Leveraging Omics Data to Expand the Value and Understanding of Alternative Splicing

By

Nathan T. Johnson

A Dissertation

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

In

Bioinformatics & Computational Biology

APPROVED:

Dmitry Korkin, Ph.D.
Advisor
Program Director

Amity L. Manning, Ph.D.
Committee Member

Zheyang Wu, Ph.D.
Committee Member

Scarlet Shell, Ph.D.
Committee Member

Ben Raphael, Ph.D.
External Committee Member

"Science, my boy, is made up of mistakes, but they are mistakes which it is useful to make, because they lead little by little to the truth."

Jules Verne, Journey to the Center of the Earth

ABSTRACT

Utilizing ‘omics’ data of diverse types such as genomics, proteomics, transcriptomics, epigenomics, and others has largely been attributed as holding great promise for solving the complexity of many health and ecological problems such as complex genetic diseases and parasitic destruction of farming crops. By using bioinformatics, it is possible to take advantage of ‘omics’ data to gain a systems level molecular perspective to achieve insight into possible solutions. One possible solution is understanding and expanding the use of alternative splicing (AS) of mRNA precursors. Typically, genes are considered the focal point as the main players in the molecular world. However, due to recent ‘omics’ analysis across the past decade, AS has been demonstrated to be the main player in causing protein diversity. This is possible as AS rearranges the key components of a gene (exon, intron, and untranslated regions) to generate diverse functionally unique proteins and regulatory RNAs. AS is highly prevalent, where on average 10 AS transcripts occur for every gene in humans. Furthermore, multiple transcripts can be expressed at the same moment leading to different protein products that can interact within their molecular environment in unique ways. The prevalence on which transcripts are alternatively spliced has been demonstrated to be based on age, tissue, cell type, and disease state.

This work brings together different ‘omics’ data to expand our understanding and promote the value of AS Specifically, there are six projects described here which make use of transcriptomics, proteomics, genomics, and epigenomics, which often overlap, on the focus in a couple of complex genetic diseases as well as analyzing a parasite, which infects soybeans. The projects range from systemically profiling machine learning methods utilizing RNA-Seq based alternative splicing expression data to promote its use, development of a method to predict whether an alternative spliced protein affects its interaction, a systematic analysis across the transcriptome for comparing binding sites and domains with alternative splicing and expression patterns, assessment of single nucleotide variation on protein binding sites in cancer, assessment of epigenomics with transcriptomics within the context of acute lymphoblastic leukemia, and looking for patterns of alternative splicing on parasites infecting soybeans.

ACKNOWLEDGEMENTS

I, first, would like to thank my Ph.D. mentor Dr. Dmitry Korkin, who has provided valuable guidance to shape how I approach science. His insight and invaluable ideas constantly refined and challenged me to be a better researcher. I will be forever indebted to him for his valuable mentorship. Additionally, I would like to thank my committee members, Drs. Amity Manning, Ben Raphael, Scarlet Shell, and Zheyang Wu for their valuable feedback and suggestions to improve my quality of communication both orally and in written format.

I, second, would like to thank my fellow lab members and collaborators, who without, this work would not have been possible to be completed: Andi Dhroso, Hongzhu Cui, Suhas Srinivasan, Oleksandr (Alex) Narykov, Xingyan Kuang, Katelyn (Katie) Hughes, Anastasia (Ana) Leshchyk, Nan Hu, and Nan (Alan) Zhao. Specifically, I would like to give a double thanks to Andi, Suhas, Hongzhu, and Alex due to their lengthy discussions, debates, and collaborative work that has allowed the various works within this dissertation to be the quality it is.

I, third, am grateful to the University of Missouri, Worcester Polytechnic Institute, the National Science Foundation, and National Institute of Health for the environment and resources that allow research to exist. Explicitly, would like to thank my external and internal collaborators, Drs. Kristen Taylor, Melissa Mitchum, Gerald Arthur, Amity Manning, Rodrigo Aldecoa, Dmitri Krioukov, Nematode Consortium, Suzanne Scarlata, Michael Green, and Haiyuan Yu. As without their problems, insight, and data, the work described here and elsewhere would not be possible. Lastly, would like to thank Takeda Pharmaceuticals for their internship, which opened my eyes to the possibilities within science in industry.

I, fourth, would like to acknowledge the countless people that can be attributed for their valuable contribution to the scientist and person that I have become prior to entering this Ph.D. program: Chestnut Laboratories and Drs. Glenner Richards, Stephen Badger, Elizabeth Bryda, and Robert White. Without their countless impact, I would not have ever considered a scientific career path or been equipped with the tools to excel.

Finally, I would like to thank my family for their constant support and understanding during this process. Their admiration and support helped give me the persistence to finish this process. Exceptional thanks go to my wife, Amanda Johnson, to whom without I would never have finished. Her constant love, support, and support (on purpose twice) gave me the motivation and perseverance.

My honors and achievements are dedicated to all these people, as without, impossible.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1: Introduction and Literature Review

1.1 21st Century Problems

Fighting disease and food production have been among the main challenges that civilizations have tried to resolve for thousands of years. Yet, our comprehension of molecular mechanisms behind genetic and infectious diseases in humans, animals, and plants has begun to crystallize only in the last 100 years (1). Our scientific understanding of molecular biology starts in the late 1800's in the form of Gregor Mendel laying the groundwork for genetics, but then re-discovered in the early 1900's (2, 3). The field quickly expands due to the discovery of the inheritability of DNA (4), the structure of DNA (5), and the central dogma of molecular biology (6) to name only a few of the many tremendous achievements.

Furthermore, these accomplishments help lay the groundwork for success in the treatment and understanding of many Mendelian diseases cataloged initiated in the 1960's by Dr. Victor McKusick (7). These Mendelian diseases defined as single loci patterns of inheritance with examples such as sickle-cell anemia and cystic fibrosis represent over 15,000 genes with known relationships between phenotype and genotype (8). However, multi-loci diseases such as cancer, diabetes, and neurological disorders do not have a clear relationship between phenotype and genotype. 38.4% of all people will be diagnosed with cancer at some point in their lifetime (9). This represents 14.1 million new cases globally and 8.2 million deaths (9). Focusing on the US, 1.7 million new cases and 825,000 deaths (9). In the US, 30.3 million (9.4%) of the U.S. has some form of diabetes (10). 10 million children (15%) are currently diagnosed with autism in the U.S. In the past 100 years, there has been incredible growth in food production despite global population growing from 1.6 to 7 billion and there was a surplus for 1.6 billion (11, 12). Innovations from synthetic fertilizers, hybridization of crops, and expansion of landmass for farming have in large part been responsible for the ability to match food demand (13, 14). Furthermore, while current farming innovations have increased productivity it has reduced food variety and lowering the quality of nutrients due to lack of ripening on the vine (15), Additionally, due to more land being used for farming, it causes a shift in pests preference for crops as a nutrient resource causing more than one billion dollars in yield loss for soybean alone (16). This observation is extended to food productions globally being reduced 20-40% due to pests alone (17). There is a need for food production to increase by 59-98% by 2050 that will likely not be resolved by known innovations (18).

1.2 Value Of Omics Data

In the past decade, the technological advancements have propelled genome sequencing with the rate that has surpassed the Moore's Law (19), generating petabytes of information and making genomics one of the first scientific areas that entered the era of Big Data. The diversity of Next Generation Sequencing (NGS) methods (20), ranging from whole-genome (21) to whole-exome (22) to RNA-sequencing (23) and reaching a single-cell precision (24) has allowed investigating the genetic material between the healthy and disease tissues of an individual and across populations (25). Substantial improvements have been made in the ability of utilizing high-throughput ''omics'' data for use in clinical environment during the same time period (26). Scientists have recently started looking at a new important target: diagnostics and treatment of complex genetic disorders (CGDs) by leveraging the omics data of different types, including genomics, proteomics, metabolomics, transcriptomics, glycomics, epigenomics, lipomics, and others (27-29). However, like most new concepts there are multiple problems surrounding attempts to utilize the omics data for treatment and diagnostics, such as a lack of a standardized protocol (30), reproducibility of the results, limited computational resources for data integration, and most importantly, the extraordinary complexity of CGDs. In the beginning of the 21st century, we are witnessing a truly pandemic growth of common diseases that are molecularly and genetically complex. GLOBOCAN 2012 estimated that 14.1 million new cancer cases including and 8.2 million deaths occurred only in 2012, with 43% of new cases and 35% of deaths coming from the economically developing countries (31). More than 1.7 million new cancer cases and almost 600,000 cancer deaths are estimated to occur in the United States alone in 2018 (9). Being the 7th leading cause of death in the U.S. in 2015, diabetes affected 30.3 million children and adults in the U.S. (9.4% of the population) (10). The number of neurodevelopmental, psychotic, and neurodegenerative disorder cases are also on the rise: for instance, the number of U.S. children aged 3–17 diagnosed with developmental disabilities, such as autism or ADHD, has reached a staggering 10 million affecting 15% of children of this age (32). To cope with the complex diseases, doctors and scientists have been relentlessly trying to improve diagnostics and therapeutic intervention through the use of experimental and computational approaches. However, for many of these diseases the tasks of early diagnostics and successful treatment are challenging and in some cases still unfeasible, impeded by our lack of knowledge of the disease at the molecular level. Understanding of the molecular mechanisms driving CGDs is in turn hindered by the multiple layers of complexity due to dozens, often hundreds, of pathogenic mutations affecting many genes, targeting multiple regulatory mechanisms and perturbing multiple pathways and systems. Complex diseases commonly manifest changes at the genetic, post-transcriptional, and epigenetic levels (33-39). Single nucleotide variations (SNVs) and indel mutations occurring in coding as well as non-coding regions of genomes are perhaps the most widely studied class of genetic changes owing to the recent progress in next generation sequencing (33-35). The transcriptional complexity of CGDs is further complicated by post-transcriptional diversity—one of the most recent discoveries is the intrinsic role of post-transcriptional variations, such as alternative splicing variations in a number of diseases (38, 40). Finally, another recent finding supported by the rapidly increasing volume of evidence is the link between the epigenetic variations and complex diseases (39). In combination, these mutations affect similar molecular targets such as TP53 (Figure 1.1.1).

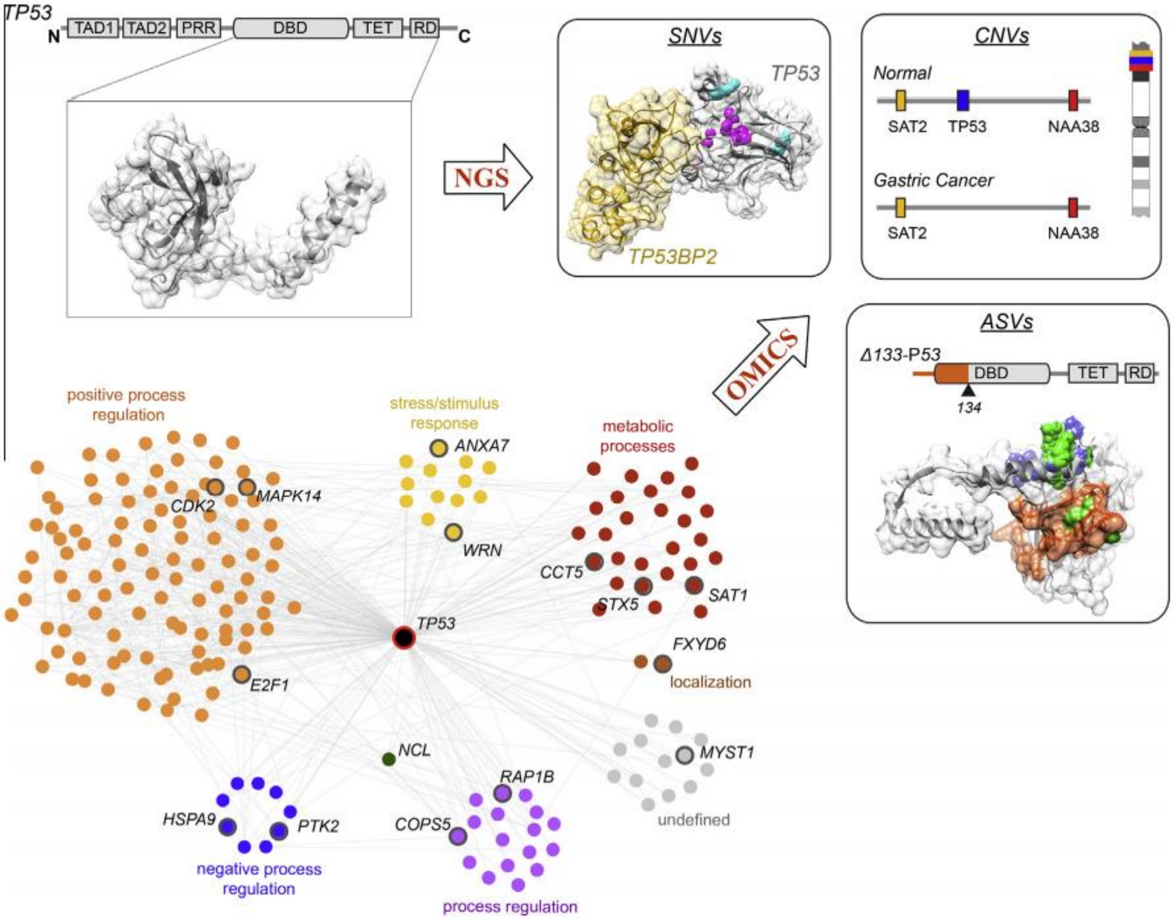


Figure 1.2.1. Genetic, structural and posttranscriptional variations of TP53 can all affect the macromolecular interactions mediated by this gene. Shown clockwise are: (i) domain architecture of P53 and the structure of its DBD-TET-RD C-terminal part; (ii) effects of nsSNVs on the PPI between TP53 (gray) and TP53BP2 (yellow) products with the neutral nsSNVs shown in cyan and disruptive in magenta, (iii) pathogenic CNV in chromosome 17 including deletion of TP53 which is associated with gastric cancer and other disorders, (iv) disruptive effect of pathogenic ASV on protein-DNA interaction but not PPI where the deleted part of DBD (orange) also includes a part of DNA binding site (green) but not TP53BP2 binding site (blue), and (v) P53-centered PPI network, with the interaction partners grouped and color-coded based on the GO annotation and well-studied proteins labeled.

1.3 Alternative Splicing and postranscriptional variation in Complex Genetic Disorders

Alternative splicing (AS), since its discovery in 1977 in viruses, is increasingly seeing an explosion of interest from the general research community (Figure 1.2.1) (41-43). This interest stems from the observation that this mechanism is predicted to be one of the widespread causes of protein diversity, however this observation is recently under scrutiny discussed later (44-47). AS occurs across all analyzed eukaryotes and in some viruses, bacteria, and archaea (48-50). Estimations of 90-95% of human genes undergo AS, and on average, there are seven to ten AS events predicted per multi-exon gene in humans (51-54). AS is a universal regulatory mechanism of gene expression that allows the initiation of more than one unique RNA species using the spliceosome from a single gene, which results in both coding and non-coding RNA (55). During posttranscriptional processing of a precursor mRNA (pre-mRNA), the spliceosome can incorporate over 200 proteins to generate different AS products through multiple mechanisms (56, 57). These mechanisms include exon skipping (the removal of specific exons), the use of alternative splice sites (which can be within introns, exons, or untranslated regions (UTRs)) and intron retention (41). These mechanisms might affect RNA stability, localization or translation (41, 58). Furthermore, rearrangement can alter the starting site for translation potentially completely rearranging the protein coding sequence (41). The choice of which AS mechanism is used is based on three main factors: 1) splice site strength, 2) cisregulatory sequences in pre-mRNAs that favor or impair exon recognition, and 3) the expression levels of trans-acting factors (RNA-binding proteins (RBPs) and splicing factors) (41). Three main observations have been reported regarding the distribution of AS transcript isoforms: (1) genes tend to express multiple isoforms simultaneously, but at different quantities, (2) there tends to be major and minor dominant isoforms of a gene that account for over 30% and 15% of total transcript expression, which shift dependent on biological context, and (3) for any two discrete isoforms from the same gene, it seems that one is always more dominant (52, 59-67). The mechanisms underlying these observations are unclear due to transcriptomic studies are rarely subjected to systemic AS analysis (Figure 1.2.1) (68).

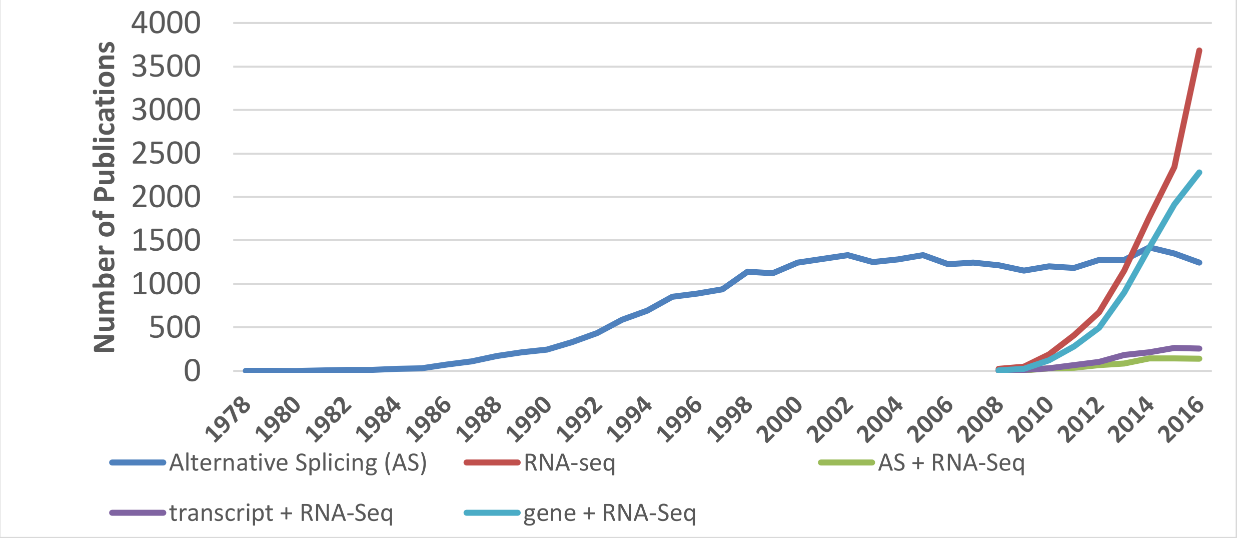


Figure 1.3.1. Number of Publications Matching Pubmed Search Terms. Searches for the key words “Alternative Splicing”, “RNA-Seq”, “RNA-Seq & Alternative Splicing”, “transcript & RNA-Seq”, and “gene & RNA-Seq”. Despite RNA-Seq being the current best tool for global assessment of alternative splicing, the research community rarely performs the analysis.

Previously mentioned, the main widespread cause for protein diversity is thought to be due to AS, but this is recently being scrutinized (69). Based on a recent large scale RNA-Seq analysis, 72% of annotated human genes undergo AS and roughly 205,000 transcripts have protein-coding potential (53, 70). Furthermore, it is predicted that in total 90-95% of all annotated human genes have the potential to undergo AS (51, 52). However, based on a high-resolution mass spectrometry of proteins on 30 histologically normal human samples resulted in proteins identified for ~84% of annotated genes, but only ~37% of them were caused by AS (71). This observation is consistent with known annotation from RefSeq (54). Furthermore, using prediction methods all known alternative spliced isoforms were tested for their protein folding ability, which concluded that approximately one third of isoforms are functional proteins (72). These discrepancies may be attributed due to the small sampling of phenotypes and/or insufficient methods (73). Another possibility is that AS transcripts are largely non-functional or play a larger unrealized role within noncoding RNA. Regardless, the functional consequences of AS are largely unexplored (41, 66). However, there are a few known patterns regarding AS. While genes tend to express isoforms simultaneously, it is not uncommon for isoforms to be specific to tissue, cell type, developmental stage or disease (41, 49, 74). Misregulation of AS underlies many diseases, including skeletal and neurodegenerative diseases, and cancer (37, 38, 58, 75, 76). Unfortunately, the precise mechanisms behind the disease-associated misregulation of AS, and the key relationships between the variations in AS and the protein function are yet to be understood. However, what is understood is the effect of post-transcriptional variation on protein-protein interactions underlie the cell's basic functioning. Therefore, it is not surprising that recent studies of disease networks have linked many genetic variations (e.g., single nucleotide variations, SNVs) and posttranscriptional variations (e.g., AS isoforms) with protein-protein interactions (PPIs) (77-79). Understanding how the AS variants can rewire the interaction network mediated by proteins associated with the disease has been defined as a critical step in studying complex genetic disorders (80-82). In our recent review, we proposed that there are four main scenarios of how AS can modify protein-protein interactions; unfolded protein, deleted protein domains, deleted functional portion of protein domain, and 3 insertion of a new protein domain potentially adding a function (Figure 1.2.2A) (83). These modifications would result in the normal protein-protein interaction network to be modified through completely or partially removing protein interactions, or adding new interactions (Figure 1.2.2B). These scenarios were later experimentally verified (84). To date, the interaction landscape determined by the post-transcriptional variants of genes is far from being fully reconstructed. Thus, fast and accurate computational methods are prone to play an important role in modeling the effects of post-transcriptional variation on PPIs. In summary, alternative splicing is a common regulatory mechanism; however its function is not fully understood. It is known to have a normal role in modifying protein function in many biological contexts (cell type, tissue, age) as well as a role in disease, but understanding this mechanism is far from complete. It is important for the research community to have a comprehensive assessment of changes to protein function as a result of alternative splicing as well as tools to gain further insight.

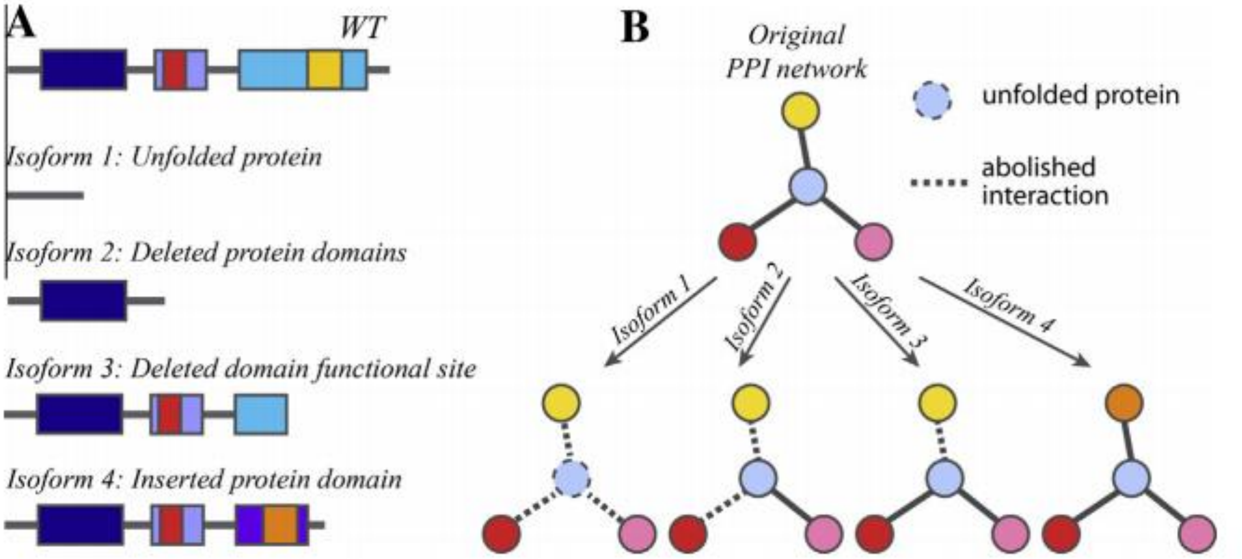

A
WT
Isoform 1: Unfolded protein
Isoform 2: Deleted protein domains
Isoform 3: Deleted domain functional site
Isoform 4: Inserted protein domain
B
Original
PPI network
unfolded protein
abolished
interaction
Isoform 1
Isoform 2
Isoform 3
Isoform 4

Fig. 1.3.2. Posttranscriptional variation and its effect on PPI network. (A) Different AS variants of a gene can have drastically different functional effects. AS can result in a small structurally disordered protein fragment. It can also remove one or several protein domains and thus abolish the functions those domains carry. Sometimes, however, only a small functional part of the domain is removed (for instance, when more than one exon corresponds to the same protein domain). Shown are AS isoforms containing protein domains with the protein binding sites. (B) Basic effects of AS variation on the PPI network. Light blue node corresponds to the four isoforms of a protein in panel A. Red/ orange binding site corresponds to the interaction between the blue node and the red/orange node.

1.4 Studying genetic variation in Complex Genetic Disorders

Due to the reduced cost of DNA sequencing and NGS's superior coverage and resolution, NGS technology is taking over traditional array-based detection methods (85). The technology provides geneticists and bioinformatics researchers with new sequence based reference datasets and necessitates revisiting the tools of the genome wide association studies (GWAS) era (34). Armored with the rapidly growing NGS data, scientists are now reaching beyond the GWAS methods, which primarily focus on genetic markers that are intended to represent causal variation indirectly, with the goal of identifying causal variants directly. This possibility is often considered the key advantage of the new sequencing approaches over genotyping methods (35), especially given the widely accepted hypothesis that many complex genetic diseases could be influenced by rare variants in many different genes (33). Many common and rare genetic variants have been associated with complex diseases. According to the National Cancer Institute (NCI)-National Human Genome Research Institute (NHGRI) (86) catalog of published GWAS, as of November 2014, there are 2060 publications describing 14,876 SNVs. Almost all common and many rare complex diseases have been addressed, including various types of cancer, cardiovascular diseases, neurological disorders, and immune system diseases. More importantly, this knowledge base has provided insights to the key molecular mechanisms underlying CGDs (87). For example, Multiple Sclerosis (MS) disease is an autoimmune demyelinating disease, whose mechanism is still not fully understood. After integrating different source of association study data, the interleukin 7 receptor (IL7R) gene stands out as a strong candidate gene with promising insights into the underling pathogenesis mechanism (88). A genetic variation in IL7R has a strong association with MS (88), and its interplay with the alternative splicing of IL7R suggests a reliable hypothesis for MS. Another interesting finding is that one genetic locus could be associated with multiple clinically distinct diseases. For example, different interleukin receptor genes that are associated with Crohn’s disease, multiple sclerosis, systemic lupus erythematosus and rheumatoid arthritis (89, suggesting that autoimmune diseases may share a common mechanism. Moreover, such ‘pleiotropy’ phenomenon has resulted in the concept of ‘diseasome’ {Goh, 2007 #5760). Last, it has become evident that in many cases, not a single gene but a group of genes, often associated with a specific pathway or biomolecular network, are targeted by the mutations (90, 91). Thus, the network and pathways information could be useful in identifying sets of genes (rather than individual genes) implicated in the disease. For instance, by applying pathway-based analysis to the whole genome association studies Askland et al. found that multiple ion channel structural and regulatory genes are likely to contribute to the susceptibility of bipolar disorder (92). Even more importantly, they propose that the heterogeneity of these gene sets across multiple studies could be the key feature of the genetic mechanism behind the susceptibility to this CGD.

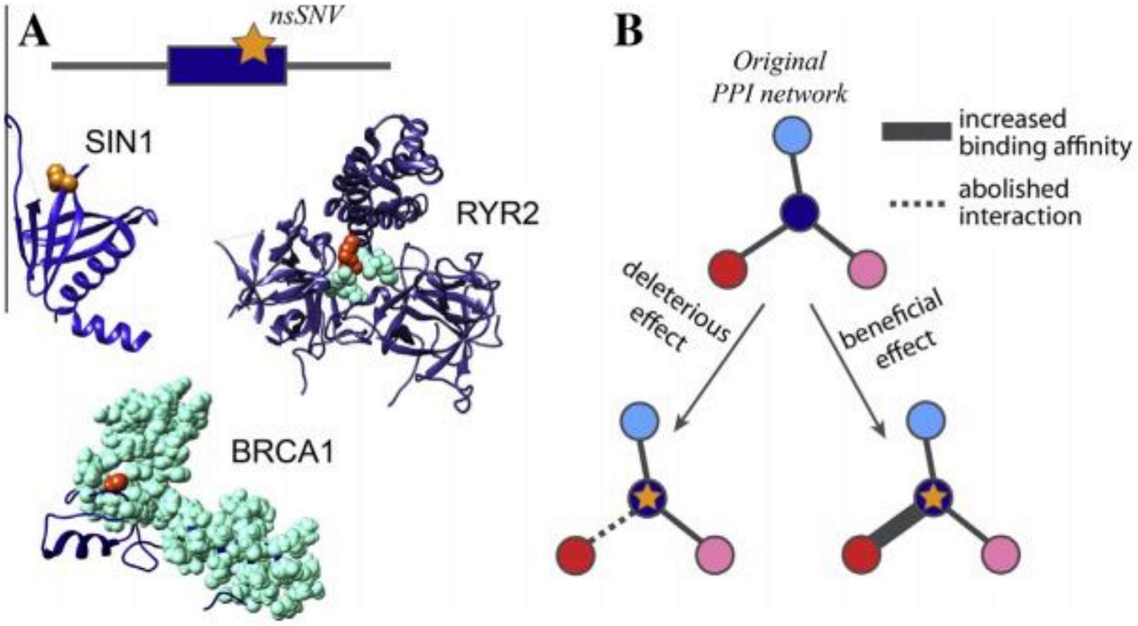


Figure 1.4.1. Genetic variation and examples of its effects on PPI network. Genetic variation such as nsSNVs and indels can affect a variety of functions. (A) Shown are examples of known mutants that are linked to complexes genetic disorders and that affect a phosphorylation site (protein SIN1), disrupt ligand binding (binding of a chloride ion by RYR2), or abolish PPI interaction by modifying the protein binding site (protein BRCA1). (B) Effects of nsSNVs can be observed at the systems level, for instance by studying a PPI network centered around the mutant proteins.

1.5 Studying epigenomic variation in Complex Genetic Disorders

A generic definition of epigenetics is regulation of gene expression from DNA that is not through variation in the genetic sequence. Regulatory mechanisms in epigenetics can be grouped into three main categories: DNA methylation, histone modifications, and nucleosome positioning. All three categories have been implicated in CGDs (39, 93). This review will focus on DNA methylation, the most widely studied type of epigenetics in CGDs. Epigenetic changes play an important role in cancer progression. Changes in the DNA methylation and histone modification can regulate transcription of the tumor suppressor genes and/or oncogenes (94). Specifically, in hypermethylation transcription of the promoter regions lead to gene silencing, while hypomethylation promotes gene expression. CpG clusters, known as CpG islands, are often targeted for DNA methylation in cancerous genes (95) in spite of the fact that in healthy tissues most CpG islands are unmethylated, even when the corresponding genes are not expressed (94). Furthermore, CpG islands, particularly on the promoter regions, are commonly associated with the gene silencing function, potentially becoming a part of the tumorigenesis process. Retinoblastoma cell control cycle gene (RB) was one of the first epigenetic lesions to be identified in carcinogenesis (94). A study conducted by Stirzaker et al. identifies three characteristics of RB identified as significant: (i) significant hypermethylation transpires throughout the CpG island, (ii) it is not limited to transcription factor binding sites, and (iii) abnormal methylation patterns are maintained in unmethylated CpG islands suggesting that the RB gene is active in the precursor cells of tumors (96). Colon cancer is another tumor type in which aberrant methylation occurs, frequently caused by germline mutations in MutL homolog 1 (MLH1) gene (97). Both DNA and histone proteins should be in an open, or unlocked, state in order for transcription to take place (94). The introduction of mutated genes, which are able to control the epigenetic state has a significant role in the promotion of cancer, e.g., myeloid leukemia and myeloid cancer (98). Epigenetic variations are implicated in a number of neurodevelopmental and neuropsychiatric disorders. In recent years, autism spectrum disorders (ASD) have received a great deal of attention

(99-103). However, the exact biomolecular components and their implications behind the ASD phenotypes remain unclear (104). It has been suggested that autism can be caused by a synergistic activity of genetic mutations as well as epigenetic dysregulation (105). A recently identified mechanism that plays a role in ASD is the loss epigenetic regulatory patterns, responsible for gene expression (106). In another work, mutations in the amyloid-beta precursor (APP) protein and the presenilin genes PSEN1 and PSEN2 are found to associate with Alzheimer's disease (AD). Significant accumulations of APP proteins are one of the earliest pathological events to take place in AD, triggering a cataclysm event, leading to neuro-degeneration (107). Age-specific epigenetic drift has been found to support the idea of potential epigenetic effects in late-onset Alzheimer's disease (LOAD). Moreover, significant epigenetic variability in genes participating in the processing of amyloid beta proteins and methylation regulation suggests predisposition to LOAD, and that aberrant epigenetic control of CpG-island may contribute to LOAD pathology (108). Malfunctioning of epigenetic mechanisms is also linked to heart diseases. A study by Angrisano et al. revealed epigenetic changes involved in the heart failure of murine under pressure overload (109). Specifically, the authors identified histone H3 modifications on the promoter regions of endoplasmic reticulum Ca2+-ATPase (SERCA-2A) and beta-myosin heavy chain (Mhc-b) genes in murine heart. DNA methylation components were also found on the myosin heavy chain 7 (Myh7) promoter regions, suggesting that epigenetic modification may indeed be involved in the heart failure. The results obtained for the animal model also pose a question if the mechanics of aging plays a direct role in alteration of epigenetics. Liver diseases are another group of CGDs where the epigenetic effect have been studied in depth, with an increasing attention to hepatic stellate cells (HSC) that play significant role in liver fibrosis. These cells can become highly proliferative, and synthesize a fibrotic matrix rich in type I collagen in the event of liver injury

(110). Mann et al. have described multiple proteins and miRNAs that are implicated in the epigenetic relay pathway that is implicated in HSC transdifferentiation (111). Understanding the epigenetic constituents may bring us closer to therapeutically controlling or reducing the fibronegesis in chronic liver diseases.

1.6 Summary

21st century society demands an increase food production by at least 58% by 2050 and increasingly higher incidence of complicated diseases such as cancer, diabetes, and neurological disorders. Following the central dogma laid out by Crick (6), molecular mechanisms through DNA, RNA, and protein are capable of complex data through omic's data. With the tremendous growth of omics data during the last years, the computational systems biology and bioinformatics approaches that leverage these data often share three common traits: (1) methods and tools are becoming increasingly data-intensive, (2) new methods are integrative in their nature; (3) the abundance of the above data provides a new critical goal for computational and informatics methods to assist in diagnostics and treatment of complex genetic disorders. Furthermore, while this dissertation highlights alternative splicing, the molecular mechanisms involved in alternative splicing cross many scientific disciplines. For example, there are cases where genetic variation, structural variation, post-transcriptional variation, and epigenetic variation are all implicated as being a factor in CGDs.

1.7 Summary of Dissertation

The main aim of this dissertation is to expand our understanding of alternative splicing and the value in evaluation in the hope of solving 21st century problems of increasing our food production and improving our understanding of complicated diseases. This work brings together different 'omics' data to expand our understanding and promote the use of A.S. Specifically, there are six projects described here which make use of transcriptomics, proteomics, genomics, and epigenomics, which often overlap, on the focus in a couple of complex genetic diseases as well as analyzing a parasite, which affects soybeans. The projects start with a basic analysis of RNA-Seq that is integrated with epigenomics to identify disease mechanisms in acute lymphoblastic leukemia. The focus shifts to looking for patterns of alternative splicing on parasites infecting soybeans that causes an estimated 1 billion in crop damage every year (16). Then the focus shifts to assessment of single nucleotide variation within eight cancer types on protein binding sites. Then the projects focus on systemically profiling machine learning methods utilizing RNA-Seq based alternative splicing expression data to promote its use, development of a method to predict whether alternative splicing occurs affects its interaction, a systematic analysis across the transcriptome for comparing binding sites and domains with alternative splicing and expression patterns.

This dissertation is organized as follows. The overall pattern is each chapter or subsection of a chapter encompasses a published paper with the citation reference listed at the end of the title. Generally speaking, the published papers have been unmodified except for dissertation formatting purposes. The exceptions to this rule are Chapter 1, 2.3, and 4.

The introduction in Chapter 1 includes a peer reviewed review paper (83) I was a part of, but due to it was published 3 years ago was updated, split up, and rearranged for the purposes of this dissertation. Both subsections for chapter 2.3 and chapter 4 had supplementary information removed due to their magnitude.

In summary, Chapter 1 includes the introduction, literature review, and organization. Chapter 2 utilizes transcriptomics in three sections. 2.1 uses epigenomics and transcriptomics to look for regulators in acute lymphoblastic leukemia (ALL). 2.2 presents a large systematic analysis demonstrating the importance of using alternative splicing from RNA-Seq to build classification models. 2.3 demonstrates the importance of A.S. in a worm (nematode) on proteins, which effect its ability to infect soybeans (effectors). Chapter 3 employs proteomics in three different subsections. 3.1 covers our developed method (AS-IN), which predicts when A.S. occurs whether it alters a known protein-protein interaction. 3.2 incorporates both genomics and proteomics by assessing whether single nucleotide variations (SNVs) in cancer effect protein binding sites. 3.3 summarizes a systemic wide analysis of A.S. effect on protein binding sites and functional domains for human. Chapter 4 summarizes some main conclusions that can be determined from this dissertation as well as possible future directions. Appendix A-B contain supplementary tables and figures for Chapters 2.1 and 3.1.

CHAPTER 2: TRANSCRIPTOMICS

2.1 Epigenetic and RNA effects in Acute Lymphoblastic Leukemia (ALL)

2.1.1 Abstract

Acute lymphoblastic leukemia (ALL) is the most common cancer diagnosed in children under the age of 15. In addition to genetic aberrations, epigenetic modifications such as DNA methylation are altered in cancer and impact gene expression. To identify epigenetic alterations in ALL, genome-wide methylation profiles were generated using the methylated CpG island recovery assay followed by next-generation sequencing. More than 25,000 differentially methylated regions (DMR) were observed in ALL patients with »90% present within intronic or intergenic regions. To determine the regulatory potential of the DMR, whole-transcriptome analysis was performed and integrated with methylation data. Aberrant promoter methylation was associated with the altered expression of genes involved in transcriptional regulation, apoptosis, and proliferation. Novel enhancer-like sequences were identified within intronic and intergenic DMR. Aberrant methylation in these regions was associated with the altered expression of neighboring genes involved in cell cycle processes, lymphocyte activation and apoptosis. These genes include potential epi-driver genes, such as SYNE1, PTPRS, PAWR, HDAC9, RGCC, MCOLN2, LYN, TRAF3, FLT1, and MELK, which may provide a selective advantage to leukemic cells. In addition, the differential expression of epigenetic modifier genes, pseudogenes, and non-coding RNAs was also observed accentuating the role of erroneous epigenetic gene regulation in ALL.

2.1.2 Introduction

Acute lymphoblastic leukemia (ALL) is a hematological malignancy associated with precursor B-cells. ALL is the most common type of cancer in children with an annual occurrence rate of 35 to 40 cases per 1 million people in the United States (112). The development and differentiation of B-cells comprises numerous stages and is a highly synchronized and controlled process governed by stage-specific gene expression (113, 114). Any deviation from normal stage-specific gene expression could lead to disease conditions including ALL. The general known mechanisms underlying the induction of ALL include chromosomal translocation, hyperdiploidy, and aberrant expression of proto-oncogenes. Advancement in deciphering additional mechanisms that may be responsible for the induction of all ALL is lacking. Therefore, the identification of key regulatory regions in the genome that may impact the development of ALL is critical to gaining a better understanding of ALL pathogenesis. DNA methylation is responsible for tissue specific gene expression and plays a significant role in hematopoiesis (115, 116) and malignant transformation (117, 118). A reduced level of CpG methylation was one of the first epigenetic alterations to be found in human cancer when compared with normal-tissue counterparts (119). Although hypermethylation of CpG islands within gene promoters has been the main focus of studies on malignant cells, the role of differential DNA methylation in other regions is gaining favor (120, 121). One such region harbors transcriptional enhancers, which reside within noncoding regions of the genome and are known to work over long distances to promote cell/tissue type specific gene expression. Active enhancers are often accompanied by DNA demethylation, (122) and alterations in enhancer methylation are seen in malignant transformation. Recently, it has been shown that differential methylation of these regions exhibit a higher correlation with gene expression than differential promoter methylation (123). As a step toward better understanding the consequence of altered DNA methylation on gene expression in pre-B ALL, MIRA-seq was used to identify altered DNA methylation throughout the genome and then correlated with transcriptome data. We show that differential comparisons of DNA methylation between normal and diseased tissue can identify potential regulatory regions of the genome and that when paired with gene expression data the functionality of the regulatory regions can be determined.

2.1.3 Results

Genome-wide DNA methylation profiles MIRA-seq was utilized to generate genome-wide DNA methylation profiles for 19 pre-B ALL patient samples from diagnostic bone marrow. Normal precursor B-cell populations (pre-BI and pre-BII) were isolated from 10 human umbilical cord blood (HCB) samples to generate methylation profiles for healthy tissue to be used as a comparator (115). On average, 188 million reads were generated for HCB samples and 174 million reads were generated for ALL patient samples (Figure 5.1.3.1A). Methylation peaks were more abundant in HCB samples (305,736) than in ALL samples (162,832) and across all chromosomes revealing an overall genome-wide reduction of methylation in ALL (Figure 5.1.3.1B). Genomic distribution analysis showed that »90% of the methylated peaks were located within intronic and intergenic regions (Figure 5.1.3.1C). The distribution of methylation peaks relative to CpG islands (CGIs) revealed that 9,814 CGIs were methylated in HCB samples and 11,015 CGIs were methylated in ALL samples but the overwhelming majority of methylated peaks were present in regions of the genome not associated with CGIs (Figure 5.1.3.1D).

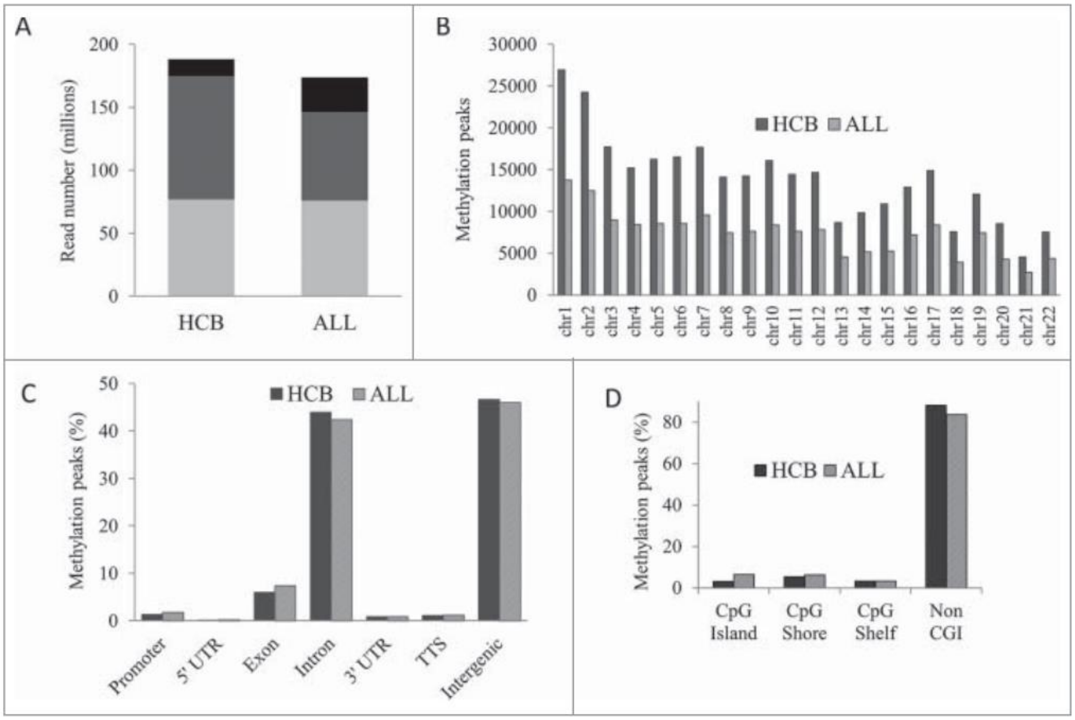


Figure 2.1.3.1. Genome-wide DNA methylation profiles in HCB and ALL. (A) Average read and alignment statistics. Reads were averaged across all individuals for HCB and ALL samples. The top of each bar represents the total number of reads for each category. Black bars: total reads; Dark gray bars: reads mapped; Light gray bars: unique reads. (B) Chromosome-wise methylation peaks. The X and Y chromosomes were excluded from analysis. (C) Genomic distribution of methylation peaks. TTS: transcription termination site. (D) Methylation peaks in CGI context.

Differentially methylated regions in ALL

To determine methylation patterns distinct to ALL, differentially methylated regions (DMRs) between ALL and HCB samples with at least a 2-fold change and an FDR of 5% were identified. A total of 15,492 regions lost methylation and 9,790 regions gained methylation in ALL compared to the normal HCB samples and the genomic distribution of loci harboring DMRs differed in the hypomethylated versus hypermethylated DMRs (Figure 5.1.3.2A and B). Hypermethylation was more prevalent in the 50 regulatory regions of genes than hypomethylation. The majority of the DMRs coincided with intergenic and intronic genomic regions. DMRs have applicability as disease specific biomarkers and may also play regulatory roles in the expression of genes that are involved in the pathogenesis of ALL. To further elucidate the importance of DMRs, we sought to identify the DMRs with regulatory potential. DMRs are associated with regulatory sequences: The promoters of protein coding genes harbor regulatory sequences required for the initiation of transcription. A total of 1,568 differentially methylated gene promoters were identified (corresponding to 1,252 hypermethylated genes and 240 hypomethylated genes) in ALL. To explore the association of DNA methylation and gene expression, MIRA-seq data and RNA-seq data were correlated. Sixty-two promoter DMRs were hypermethylated and downregulated in ALL and were significantly enriched for genes involved in the regulation of transcription and apoptosis, whereas 37 promoter DMRs were hypomethylated and upregulated and were significantly enriched for genes involved in GTPase activation, the regulation of cell proliferation, and those that play a role in protein complex assembly. Additionally, hypermethylated DMRs were identified in the promoters of 3 tumor suppressor genes, MTSS1, PAWR, and EXT1, and corresponded with a significant decrease in gene expression. In addition to protein coding gene promoters, differential methylation was observed within 1,000 bp upstream or Figure 5.1.3.1. Genome-wide DNA methylation profiles in HCB and ALL. (A) Average read and alignment statistics. Reads were averaged across all individuals for HCB and ALL samples. The top of each bar represents the total number of reads for each category. Black bars: total reads; Dark gray bars: reads mapped; Light gray bars: unique reads. (B) Chromosome-wise methylation peaks. The X and Y chromosomes were excluded from analysis. (C) Genomic distribution of methylation peaks. TTS: transcription termination site. (D) Methylation peaks in CGI context downstream of the TSS in non-coding RNAs and pseudogenes (Figure 3.1.3.2C). MicroRNAs (miRNA) are non-coding RNAs that regulate expression through imperfect base-pairing with the 30 UTR of multiple target genes. A total of 69 miRNAs were differentially methylated in ALL including miR-375, miR-196a, miR-3545, miR-9-1/2/3, miR-124-1/3, and miR-34b, which have been implicated in human malignancies.13-15 RNAseq libraries were prepared from poly(A) RNA and excluded the capture of miRNA; therefore, correlation studies between methylation and gene expression were not performed for miRNA. The regulatory potential of DMRs associated with miRNAs warrants further attention. Long intergenic non-coding RNAs (lincRNAs) are emerging as key regulators of numerous cellular processes and regulate the expression of multiple target genes. Differential methylation occurred in 65 lincRNAs. Of these, hypomethylation and upregulation was observed in AC002398.5, DIO3OS and LINC00642. Lastly, 55 pseudogenes were differentially methylated in ALL. No correlations between expression and promoter methylation was observed in the pseudogenes; however, pseudogenes, much like lincRNAs, have the potential to epigenetically regulate their parental genes and were further investigated. It is well known that transposable element activities are often silenced by DNA methylation (124), and that transcriptional activation of these elements results in transposable element mediated insertions and chromosomal rearrangements in many cancers (125). Many of the intergenic DMRs were associated with transposable elements and repeat sequences (Figure 5.1.3.2D). Non-autonomous short interspersed nuclear elements (SINE) were the most abundantly present transposable element within the differentially methylated intergenic regions followed by long terminal repeat (LTR), autonomous long interspersed nuclear elements (LINE) and satellite repeats. Centromeric a satellite repeats were often hypermethylated in ALL, which may block CENP-A and result in centromere inactivation.

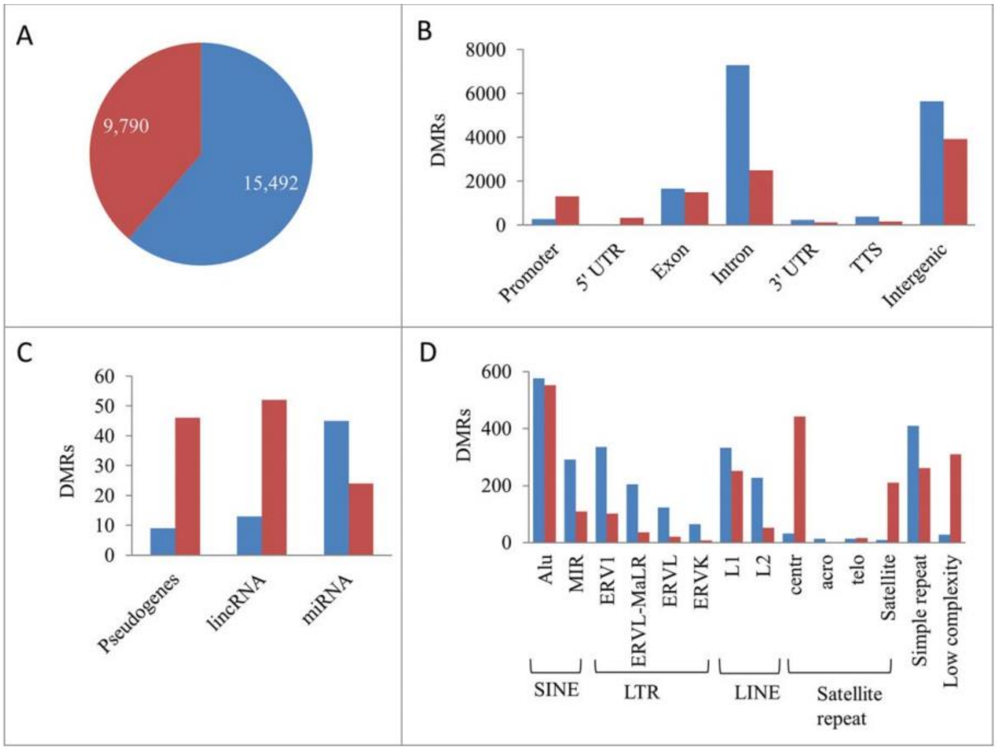


Figure 2.1.3.2. Differentially methylated regions in ALL. (A) Hypomethylated (blue) and hypermethylated (red) regions. (B) Genomic distribution of hypo- and hyper-methylated DMR. (C) DMRs associated with the 50 regulatory region of pseudogenes and non-coding RNA. (D) Intergenic DMRs associated with transposable elements and repeat sequences

DMRs are associated with predicted regulatory sequence

Differential methylation predominately occurred in intergenic and intronic regions in ALL. One third of the intronic DMRs (3,341) were located within 150 base pairs of the 50 or 30 splice sites and could potentially alter appropriate splicing in ALL. To investigate whether the intergenic and intronic DMRs coincided with the location of regulatory enhancer elements, the sites for intergenic and intronic DMRs were overlaid with ENCODE ChIP-seq data for enhancer related histone marks (H3K4me1 and H3K27ac) in the GM12878 lymphoblastoid cell line. Overall, 765 intergenic and intronic DMRs overlapped with potential enhancer like regions (eDMR). Of these, 453 were hypomethylated and 312 were hypermethylated. Enhancer methylation has been shown to have a stronger association with gene deregulation than promoter methylation in cancer (123). To investigate the association between enhancer methylation and gene expression in our data, lists were constructed of the nearest upstream and downstream gene to identify the potential target genes for each eDMR. A total of 81 genes exhibited significantly decreased expression in ALL that corresponded with hypermethylation of potential neighboring eDMRs, and 111 genes showed significantly increased expression that corresponded with eDMR hypomethylation. Functional annotation clustering revealed that downregulated genes with eDMR hypermethylation included those involved in cell cycle processes, cell division, regulation of gene expression, cytoskeleton, and a large number of zinc finger proteins, whereas upregulated genes with eDMR hypomethylation included those involved in lymphocyte activation, cell migration, apoptosis, DNA replication, and DNA metabolic processes.

Gene body DMRs are associated with gene expression

Associations between gene body methylation and gene expression were also observed. Increasing gene body methylation along with promoter methylation has been shown to have a stronger repressive effect on gene expression during normal B-cell development than promoter methylation alone (126). However, the effect of gene body methylation in the absence of promoter methylation is less clear. Both inverse and positive correlations between gene body methylation and gene expression were observed. Gene body hypermethylation and a significant decrease in expression was observed in Figure 5.1.3.2. Differentially methylated regions in ALL. (A) Hypomethylated (blue) and hypermethylated (red) regions. (B) Genomic distribution of hypo- and hyper-methylated DMR. (C) DMRs associated with the 50 regulatory region of pseudogenes and non-coding RNA. (D) Intergenic DMRs associated with transposable elements and repeat sequences. 261 genes and included protein kinases (CDK5R1, NRBP1, LYN, NUAK2, PHKB, BLK, PRKAG2, MKNK2, SMG1, TRIO, GAK, PRKD2, ULK1, RIOK3, WNK4, MAP3K9, PDGFRA, NEK8, DCLK2, TLK2, LRRK1, CDC42BPB, CAMK1D), cell morphogenesis genes

(CDK5R1, GDF7, ULK1, LAMA5, NR4A2, MAPK8IP3, SEMA3B, MYCBP2, NFATC1), lymphocyte differentiation genes (CHD7, IL7, CEBPG, HDAC9, FOXP1), chromatin modifiers (RSF1, CREBBP, BANP, ARID1B, UIMC1, CHD8, CHD7, WHSC1L1, PHF21A, TLK2, IRF4, HDAC9, RERE), and regulators of MAPK, JNK, JUN kinase activity. Conversely, gene body hypomethylation and a significant increase in expression was observed in 815 genes and included the DNA methyltransferases (DNMT3A and DNMT1), antiapoptotic genes (IL2RB, PRDX2, BCL2L1, TCF7L2, DAPK1, AKT1, ATF5, BAX, TGM2, NOS3, THBS1, and MYO18A), and telomere organization genes (TERT and TNKS1BP1). Additionally, many genes showed positive correlations between methylation and expression. For example, several members of the protein tyrosine phosphatase family that regulate many cellular processes, such as cell growth, mitotic cycle, cellular differentiation, and malignant transformation, were upregulated and hypermethylated in ALL. Alternately, genes that play roles in B-cell activation were downregulated and hypomethylated in ALL.

B-cell development genes and epigenetic modifiers are aberrantly expressed in ALL

To investigate the deregulation of gene expression in ALL, genome-wide gene expression profiling of ALL patients and healthy precursor B-cells was performed using RNA-seq. A total of 3,700 genes were significantly upregulated in ALL vs. healthy samples and 2,734 genes were significantly downregulated. Forty-three genes known to play roles in B-cell differentiation and activation were differentially expressed and may contribute to the pathogenesis of ALL (Table 2.1.3.2). The aberrant expression of epigenetic modifiers was also observed. The DNA methylation catalyzing enzymes DNMT1, DNMT3A, and DNMT3B were significantly upregulated in ALL. Conversely, 2 genes known to actively demethylate DNA (127), TET2 and TET3, were significantly downregulated in ALL. In addition, 22 genes encoding histone proteins were significantly upregulated in ALL. Finally, the chromatin activating histone lysine acetyltransferases (MGEA5, CDYL, CREBBP, EP300, and NCOA3) were downregulated and the chromatin inactivating histone deacetylases (HDAC11 and SIRT2) were upregulated in ALL.

Table 2.1.3.1 Patient Characteristics

Patient ID	Blast rate (%)	Age (months)	WBC, 10^3/µl	Sex	Immunophenotype	Cyotgenetics
A4	88	4	7.8	M	19;10	hyperdiploidy
A15	94	36	7.8	M	19;10	hyperdiploidy
A18	97	17	4.3	F	19;10	46, XX-15der(1) t (1;?), del(6)(q21),t mar
A19	88	36	3.7	M	19;10	hyperdiploidy
A20	92	120	3.6	M	19;10	46, XY
A21	91	36	6.6	M	19;10	46, XY t(3;19)(p25;p13)
A22	94	60	2.5	F	19;10	47, XX C21; 48, XX
A23	96	180	2.3	M	19;10	46, XY del(6)(q21;q27)
A24	94	108	3.7	M	19;10	45, -7 -9 +der(9) t(8;9)(q112;p11)
A25	96	48	13.7	M	19;10	46, XY
A26	91	48	4.3	M	19;10	47, XY
A28	96	36	1.5	M	19;10	none available
A29	93	24	10.2	F	19;10;20	46, XX
A30	94	24	3.7	F	19;10;20wk	46, XX
A31	94	132	18.8	M	19;10;20	45, XY -7
A32	92	36	3.4	M	19;10;20	none available
A33	88	180	4.5	M	19;10;20	46, XY
A35	97	22	25.9	M	19;10;20	46, XY
A36	91	72	2.7	F	19;10;20	46, XX

A37	93	20	2.5	M	19;10	hyperdiploidy

Differential expression of transcripts with epigenetic regulatory functions

LincRNAs epigenetically regulate gene expression by a number of diverse mechanisms including recruitment of histone methyltransferases through polycomb repressor complex 2 to modify chromatin states (128, 129), and the differential expression of lincRNA has been shown to play critical roles in many diseases (130, 131). Differential expression analysis of lincRNAs in ALL patients compared to healthy controls revealed 197 lincRNAs were differentially expressed. Among them, 104 lincRNAs were upregulated and 93 were downregulated in ALL. Pseudogene transcripts play a significant role in cancer pathogenesis and are differentially expressed in different types of cancer (132, 133). The relationship between differentially expressed pseudogene transcripts and the expression of parent gene targets was diverse in our data. In some instances, the upregulation of a pseudogene was associated with the downregulation of its parent gene. For example, the pseudogene GRK6P1 was upregulated and associated with downregulation of their parent gene GRK6. Interestingly, GRK6 phosphorylates the activated forms of G protein-coupled receptors (GPCRs) thereby instigating their deactivation. Further, the overexpression of GPCRs is known to contribute to cancer cell proliferation (134). Thus, the upregulation of GRK6P1 may lead to the constitutive activation of GPCRs and contribute to the proliferation of cancer cells. Conversely, in other instances the downregulation of a pseudogene was associated with the upregulation of its parent gene. For example, the downregulation of AC007041.2, RP11-368P15.1 and KRT18P4 was associated with the upregulation of DRG1 and NDUFB3, and KRT18 respectively. KRT18 (cytokeratin 18) is involved in multiple cellular processes including apoptosis, mitosis, cell cycle progression, and cell signaling and is hypothesized to be involved in carcinogenesis through multiple signaling pathways (135). Therefore, the pseudogene mediated upregulation of KRT18 may lead to the aberrant regulation of signaling pathways in ALL. A positive correlation was also observed in which upregulated pseudogene transcripts were associated with upregulated parent gene transcripts and downregulated pseudogene transcripts were associated with downregulated parent gene transcripts. In these cases the pseudogene transcripts may upregulate their parent gene by competing with endogenous RNAs that share miRNA response elements (136), or by competing for RNA binding proteins that degrade their parent gene and vice versa. In ALL, the upregulation of pseudogenes RP11-423H2.1, FAM86C2P, and HMGB1P41 was associated with the upregulation of their parent genes THOC3, FAM86A, and HMGB2. Previous studies have shown that HMGB2 is overexpressed in a variety of cancers and that there is a decline in the proliferation of cancer cells when siRNA is used to knockdown expression of HMGB2 (137, 138), suggesting a putative role in the pathogenesis of ALL. Likewise, some pseudogenes were downregulated and their parent genes were also downregulated. Specifically, the downregulation of PABPC1P3 was associated with the downregulation of its parent gene, PABPC1, which encodes a poly(A) binding protein involved in stabilizing the 50 cap of mRNA. The downregulation of PABPC1 has also been reported in esophageal cancer(139). It is possible that the pseudogene mediated downregulation of PABPC1 results in unstable mRNA transcripts and contributes to the pathogenesis of ALL.

Table 2.1.3.2: Differentially expressed genes in ALL involved in B-cell development and epigenetic modifications

Gene	Fold change*	Gene	Fold change*
B-cell development genes		**B-cell development genes**	
DNTT	-10.25	LYN	4.08
VPREB1	-9.08	IRF8	4.30
RAG1	-8.81	**DNA demethylase**	
RaAG2	-8.52	TET3	1.02
IGLL1	-7.37	TET2	1.17
FCER1G	-6.41	**DNA methyltransferase**	

LEF1	-5.49	DNIMT3B	-5.54
TNFSF4	-4.91	DNMT1	-1.89
HMGB2	-3.49	DNMT3A	-1.44
LCP2	-3.30	**Histone deacetylases**	
OAS3	-3.21	HDAC11	-1.75
VPREB3	-3.15	SIRT2	-0.99
IL18R1	-2.94	**Histone lysine acetyltransferases**	
BST1	-2.87	CDYL	1.15
CD59	-2.63	CREBBP	1.53
CTSC	-2.31	EP300	1.56
SOX4	-2.15	MGEA5	1.96
ADA	-1.91	NCOA3	3.39
IGJ	-1.89	**Histones**	
LRRC8D	-1.84	HIST1H2BJ	-8.21
NOTCH1	-1.62	HIST1H3H	-8.07
TCF3	-1.22	HIST1H2BO	-7.60
CEACAM1	-1.21	HIST1H3J	-7.47
PAXBP1	-1.12	HIST1H2BH	-6.90
MALT1	1.05	HIST2H2AB	-6.60
IGHM	1.17	HIST1H3D	-6.43
ETS1	1.39	HIST1H4F	-6.43
BCL2	1.40	HIST1H2AC	-5.91
HLA-DMB	1.58	HIST1H2BF	-5.81
RFX1	1.63	HIST1H4H	-5.76
BCL10	2.08	HIST2H2BF	-5.29
IL24	2.16	HIST1H2AD	-5.07
BTG1	2.18	HIST1H1E	-4.95
HLA-DQB1	2.39	HIST1H2BK	-4.82
IRF4	2.46	HIST1H4I	-4.76
FCGR2B	2.56	HIST1H2BC	-4.65
ADAM8	2.60	HIST1H4E	-4.64
CARD11	3.24	HIST2H2BE	-4.43
ADAM19	3.59	HIST1H2BN	-4.35
MS4A1	3.87	HIST1H1C	-4.13
IL4R	4.00	HIST1H2BD	-3.72

Negative log fold change = upregulated in ALL, positive fold change = downregulated in ALL *Log2 fold change.

2.1.3 Discussion

On average, more than 50 million unique mapped MIRA-seq reads were generated providing genome-wide coverage of the methylome in 19 pediatric ALL patients. Importantly, these profiles were compared to healthy precursor B-cells isolated from umbilical cord blood, the normal counterparts of malignant precursor B-cells to identify DMRs. To determine the regulatory potential of DMRs, transcriptomes were also generated and differential expression was determined between ALL patients and normal controls. Previous studies in ALL have identified inverse correlations between gene expression and DNA methylation in CGIs, CGI shores and gene promoters (140). In this study, 99% of DMRs associated with a CGI were hypermethylated in ALL; however, these only accounted for a small number of the total DMRs. In fact, more than 80% of DMRs were identified in intronic or intergenic regions and not within a CGI context. Since DMRs can be used as biomarkers and as targets for novel therapeutics, we sought to identify the most likely candidates with regulatory potential. Strikingly, 70% of the intergenic DMRs were concomitant with functional regulatory elements including transposable elements, enhancers, transcription factor binding sites, ncRNA, and pseudogenes. Inverse and positive correlations between DNA methylation in regulatory regions and gene expression were observed. In addition, inverse and positive correlations were observed between gene body methylation and expression. The cause and effect of DNA methylation within gene bodies is not fully understood; however, mechanisms leading to faulty gene expression have been postulated including the regulation of transcriptional elongation (141), cell-type specific selection of alternative promoters (142), modulating alternative RNA splicing (143), or defining alternative polyadenylation sites (144). Genes that are regulated by DNA methylation and provide a selective growth advantage to cancer cells have been referred to as epi-driver genes (145). The ability to weed out driver epimutations from passenger epi-mutations is crucial in the quest to delineate potential therapeutic targets from a multitude of passenger events. Integrated DNA methylation and gene expression analysis identified potential epi-driver genes including SYNE1 (cytokinesis), PTPRS (signaling molecule), PAWR (pro-apoptotic gene), HDAC9 (downstream target of KRAS), RGCC (cell-cycle regulator), and MCOLN2 (unknown function), which were hypermethylated in the 50 regulatory region and downregulated in ALL. These genes have also been shown to be hypermethylated and/or downregulated in other malignancies (146-148), indicating the potential for tumor suppressor activity and supporting the role of DNA methylation as a regulator of gene expression. Although the function of MCOLN2 is unclear, the B-cell lineage specific activator PAX5 regulates its expression, strongly implicating its involvement in early Bcell development (149). Taken together, the downregulation of these genes due to DNA methylation may play important roles in the development of ALL. Perhaps the most paramount finding of this study was the identification of potential regulatory enhancers (eDMR). In relation to this, potential epi-drivers regulated by DNA methylation of an eDMR were also identified. Three of the genes with hypermethylated promoter DMRs (SYNE1, PTPRS, and MCOLN2) also possessed a hypermethylated eDMR. In addition, LYN and TRAF3 were downregulated in ALL patients and associated with a hypermethylated eDMR. LYN plays an important role in the regulation of B-cell differentiation, proliferation, survival and apoptosis, and TRAF3 negatively regulates the activation of the NF-kB2 pathway in B-cells. Conversely, FLT1 and MELK were upregulated and associated with a hypomethylated eDMR. Both genes have previously been shown to be upregulated in cancer (150, 151). Furthermore, FLT1 has been shown play a role in the proliferation of tumor cells (152), and suppression of MELK expression by siRNA has been shown to inhibit the growth of cancer cells. Therefore, the aberrant expression of these genes due to DNA methylation may provide a survival advantage to malignant cells and play a role in pediatric ALL. In summary, novel differentially methylated regulatory regions and differentially expressed genes were identified that may contribute to the pathogenesis of ALL. As expected, genes associated with B-cell development and epigenetic modifier genes were differentially expressed. The de novo DNA methyltransferases

(DNMT3A, DNMT3B) responsible for the establishment of DNA methylation patterns and chromatin inactivating deacetylase genes were upregulated, whereas TET1 and TET2, which are responsible for actively demethylating DNA and chromatin activating acetyltransferase genes were downregulated in ALL. The upregulation of methylating enzymes along with the downregulation of demethylating genes supports the theory that the loss of methylation is a passive event that occurs during DNA replication over multiple uncontrolled cell divisions (153). Accordingly, the overall result of the aberrant expression of the epigenetic modifier genes observed in this study may effectively be the inactivation of key genes that contribute to ALL. In addition, pseudogenes and lincRNAs genes were also aberrantly expressed in ALL and have functional roles in epigenetic regulation through diverse mechanisms including behaving as antisense RNA, endo-siRNA, competing endogenous RNA, or competing for RNA-binding proteins to regulate their target genes. Moreover, for the first time, putative transcriptional enhancers were identified that were differentially methylated and associated with the expression of a neighboring gene. Importantly, these may be used as prospective biomarkers for ALL and/or as targets for novel therapeutic agents that can restore altered DNA methylation and gene expression back to the normal state with the ultimate goal of improving treatment therapies and patient outcomes.

2.1.3 Materials and Methods

Patient samples

De-identified patient samples were obtained under full ethical approval of the institutional review board at the University of Missouri. A total of 20 pre-B ALL patient samples (Table 5.1.3.1) and pre-BI and pre-BII cells from 10 healthy individuals were used for this study. ALL patient samples contain at least 88% blasts. Normal control pre-BI and pre-BII cells were isolated from

10 human umbilical cord blood (HCB) samples as previously described (154). Briefly, mononuclear cells were isolated by density gradient centrifugation using Ficoll-Paque PLUS (GE Healthcare Bio-Sciences AB; cat. no. 17–1440-03) followed by depletion of all non B-cells with biotin conjugated antibodies cocktail and anti-biotin monoclonal antibodies conjugated to magnetic beads using human B cell Isolation Kit (MACS Miltenyi Biotec; order no. 130-093-660). For the methylation studies, purified B-cells were fluorescently labeled with antibodies against cell surface antigen (CD19, CD34, CD45; BD Biosciences) specific to individual stages of B-cell development. Finally, the fluorescently labeled cells were sorted as pre-BI (CD19C/CD34¡/CD45low) and pre-BII (CD19C/CD34¡/ CD45med). Because no regions of differential methylation were observed in pre-BI versus pre-BII cells, transcriptomes were generated for precursor B-cells which include both pre-BI and preBII subsets. To obtain this population of cells, purified B-cells were fluorescently labeled with antibodies against CD19 and IgM and precursor B-cells (CD19C/IgM¡) were isolated by flow cytometry.

Antibodies

The following antibodies were used for flow cytometry and non B-cell isolation through column purification: BD PharMingenTM PE Mouse Anti-Human CD34 (BD Biosciences; cat. no. 560941); BD PharmingenTM APC Mouse Anti-Human CD19 (BD Biosciences; cat. no. 555415); CD45 FITC (BD Biosciences, cat. no. 347463); BD PharmingenTM PE Mouse Anti-Human IgM (BD Biosciences, cat. no. 555783); B cell Isolation kit (MACS Miltenyi Biotec; order no. 130-093-660).

MIRA-seq library preparation

Genomic DNA from ALL patient samples was isolated using DNeasy Blood and Tissue Kit (Qiagen; cat. no. 69506) according to manufacturer’s instructions. MIRA-seq libraries for normal precursor B-cells were prepared as previously described.4 For ALL patient samples, 1.0 mg of DNA from each ALL patient was sonicated with alternating 30 seconds on/off intervals for a total of 9 minutes to generate 200- to 600-bp fragments. A small portion of sonicated DNA was run on 1% agarose gel to ensure the sonication accuracy. The remaining sonicated DNA fragments were concentrated and purified using the MinElute PCR purifi- cation kit (Qiagen; cat. no. 28004). Adaptor ligation to fragmented DNA followed by MIRA using MethylCollectorTM Ultra kit (Active Motif; cat. no. 55005) was performed according to manufacturer's protocols and as previously described.4 After size selection of enriched methylated DNA on 1% agarose gel, PCR amplification of recovered methylated DNA fragments was performed for 11 cycles and then purified with the MinElute gel extraction kit (Qiagen; cat no. 28604). In order to validate the enrichment of methylated DNA, end point PCR amplification of methylated SLC25A37- and unmethylated APC1- regions was performed with the following primer pairs: 50 -CCCCC TGGACGTCTGTAAG-30 (forward) and 50 -GGCATCTGGTAGATGACACG-30 (reverse) for SLC25A37, and 50 -ACTGCCATCAACTTCCTTGC-30 (forward) and 50 -GCGGATT ACACAGCTGCTTC C-30 (reverse) for APC1. Quantity and fragment analysis was performed using Qubit and Bioanalyzer before sequencing. Four high quality MIRA-seq libraries were multiplexed in 10nM concentrations and sequenced on the Illumina HiSeq 2000 (1£100 bp reads) at the DNA Core Facility, University of Missouri-Columbia.

Identification of methylated peaks and differentially methylated regions in ALL

MIRA-seq data processing and methylated peaks for individual samples were identified using MACS2 pipeline as previously described.4 Briefly, following adaptor trimming, sequences were aligned to the human reference sequence (GRCh37 with SNP135 masked) with bowtie2 (version 2.1.0). Patient sample A32 had an insufficient numbers of reads and was excluded from subsequent analyses. Picardtools (version 1.92) were used to remove duplicate reads from the BAM files. The resulting BAM files were indexed with SAMtools "index." Methylated peaks were identified using MACS2 (version 2.0.10.20130712) (155) with default parameters. Unified peak locations across the samples were created using bedtools (version 2.17.0). Individual sample was assigned a peak when their own peak overlapped with the unified peaks. ALL and HCB peaks were included if the peak was present in at least 17 biological replicates. Differentially methylated regions (DMRs) between the ALL and control precursor B-cells isolated from HCB were identified as described previously.4 The coverage depth for each sample was analyzed and any sample with insufficient depth (saturation correlation < 0.90) was omitted from further analyses. Following normalization of data using a CpG coupling factor-based method (156), DMRs were identified. Initially regions of interest (ROIs) were determined based on read counts within 100 bp windows with a 300 bp overlaps (expected fragment size of 400 bp). Non-specific filtering was performed by discarding the ROIs with modest signal representation (<20 mean counts across all samples). Differentially methylated regions were identified from the remaining ROIs using the edgeR package called via the MEDIPS package in R/Bioconductor. The ROIs with <5% false discovery rate (FDR; Benjamini-Hochberg) and at least a fold2- change were identified as a DMRs. ROIs immediately adjacent to one another were combined into a single DMR. Hyper- and hypomethylated ROIs were merged separately so that only putatively consistent ROIs were combined. The reported log fold change for merged DMRs is the maximum log2 fold change for any of its constituent ROIs. All MIRA-seq data were deposited in NCBI Sequence Read Archive (Accession SRP058314).

Annotation and enhancer prediction

Methylated peaks and differentially methylated regions were annotated with HOMER (Hypergeometric Optimization of Motif EnRichment), version 4.3, using the default setting to identify genomic locations (157). The X and Y chromosomes were excluded from the analysis as the genders of individual normal samples were unknown. CpG island positional information from the University of California Santa Cruz (UCSC) table browser was used to determine the position of methylation peaks within a CpG island context. The genomic locations of enhancers were identified based on the enrichment of histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac) modifications in the GM12878 cell line (lymphoblastoid) available from ENCODE.

RNA-seq library preparation and data analysis

RNA samples were also obtained from the pre-B ALL patients (20 samples) utilized in the MIRA-seq assays and from normal precursor B-cells isolated from HCB (8 samples). RNA sequencing libraries were constructed with the NEBNext UltraTM Directional RNA Library Prep Kit for Illumina (New England Biolabs; cat. no. E7420) and sequenced on the Illumina HiSeq 2000 (1£100 bp reads) at the University of Missouri DNA Core Facility. The reads were preprocessed to remove poor quality reads of <20 using FastX toolkit (http:// hannonlab.cshl.edu/fastx_toolkit/). Reads were aligned to hg19 using Tophat (v2.0.13) with default settings. Differential gene expression between ALL and healthy precursor B-cells were determined using Cufflinks with default parameters (version 2.2.1)(158). The read counts along with FPKM values and their variances were calculated by cuffdiff 2 and the log fold change and p-value was calculated for each gene. Multiple testing corrections using Benjamini-Hochberg was also performed (qvalue). The same cutoffs for FDR and fold change used in the analysis of methylated ROIs were used to determine differential expression. All functional annotations were performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (159). All RNA-seq data were deposited in NCBI Sequence Read Archive (Accession SRP058414).

2.2 Novel global effector mining from the transcriptome of early life stages of the soybean cyst nematode *Heterodera glycines (160)*

2.2.1 Abstract

Soybean cyst nematode (SCN) *Heterodera glycines* is an obligate parasite that relies on the secretion of effector proteins to manipulate host cellular processes that favor the formation of a feeding site within host roots to ensure its survival. The sequence complexity and co-evolutionary forces acting upon these effectors remain unknown. Here we generated a *de novo* transcriptome assembly representing the early life stages of SCN in both a compatible and an incompatible host interaction to facilitate global effector mining efforts in the absence of an available annotated SCN genome. We then employed a dual effector prediction strategy coupling a newly developed nematode effector prediction tool, N-Preffector, with a traditional secreted protein prediction pipeline to uncover a suite of novel effector candidates. Our analysis distinguished between effectors that co-evolve with the host genotype and those conserved by the pathogen to maintain a core function in parasitism and demonstrated that alternative splicing is one mechanism used to diversify the effector pool. In addition, we confirmed the presence of viral and microbial inhabitants with molecular sequence information. This transcriptome represents the most comprehensive whole-nematode sequence currently available for SCN and can be used as a tool for annotation of expected genome assemblies.

2.2.2 Introduction

The soybean cyst nematode (SCN) *Heterodera glycines* is the most economically important pathogen of soybean, causing over one billion dollars in yield loss annually (16). This microscopic roundworm begins its life cycle as an egg in the soil, undergoing one molt before hatching as a second-stage juvenile (J2). Once the nematode has hatched, it migrates through the soil towards a host plant where it invades the root tissue and migrates towards the vasculature, selecting a single cell to establish a feeding site called a syncytium. At this point, the nematode penetrates the cell wall using its stylet and releases a set of secretions into the host cell, including effector proteins. Stylet-secreted effector proteins identified to date share many characteristics including the presence of a signal peptide, lack of a transmembrane domain, and expression in the esophageal gland cells (161). These effector proteins manipulate the host cell by modulating a variety of cellular processes to make it more suitable for the nematode, including suppression of host defense and stress responses and causing significant transcriptional re-programming in the host cell nucleus (162). Effectors harboring nuclear localization signals are recognized by host cellular machinery for targeting to the nucleus where they modulate host nuclear functions (163). Similar to effectors delivered by the stylet of piercing/sucking insects, the type III secretion system of bacteria or the haustorium of pathogenic fungi and parasitic plants, these effectors represent an interface between the nematode pathogen and host (164, 165). Once the feeding site is established, the nematode becomes sedentary and relies entirely on the host for nutrition for the remainder of its life cycle. The nematode slowly swells up as it undergoes a series of molts and differentiates into either a male or a female. Females protrude from the roots while the males regain mobility and exit the root to fertilize females, following which the males die. The females eventually die after fertilization, their bodies hardening into a protective casing for the eggs called a cyst that breaks off into the soil and begins the process anew. The early stages of the nematode infection cycle represent a key point in determining the fate of a cyst nematode. Whether or not the nematode will survive long enough to complete its life cycle depends on the ability of the nematode to survive and circumvent the hostile environment presented by the plant host.

In recent years, next generation sequencing technologies have been applied to several plant-parasitic nematode species, resulting in the assembly of complete genomes for *Meloidogyne hapla*, *M. incognita*, *Globodera ellingtonia*, *G.pallida*, *G. rostochiensis*, *Ditylenchus destructor*, and *Bursaphelenchus xylophilus (166-171)*. Despite the enormous economic importance of *H. glycines*, no finished and comprehensively annotated genome is currently available. In the absence of a sequenced genome, several other plant-parasitic nematode systems have turned to *de novo* transcriptome-level studies instead (172-175). These studies were able to identify key features of the interaction of plant host and nematode pathogen, including the discovery of new effectors.

In the SCN system, studies have primarily focused on identifying and characterizing stylet-secreted effectors produced in the esophageal gland cells, which has resulted in the identification of 72 SCN effectors (176-178). These studies based their identification of SCN effectors on the presence of a signal peptide as well as expression in the esophageal gland cells confirmed by *in situ* hybridization. Multiple functional studies have since been performed using these effectors, identifying host targets and characterizing their role in cyst nematode parasitism (179). Though the approach focused on the gland cells has been highly successful in identifying stylet-secreted effector proteins, low abundance transcripts, those harboring non-canonical secretion signals, and those encoding secreted proteins originating in other structures of the nematode, such as amphids (180), are lacking. A global analysis allows for a comprehensive assessment of effectors, enabling studies to assess effector variation within and across populations to identify highly variable effectors potentially correlated with virulence, as well as those effectors highly conserved across the population that may be key components of the SCN infection process. Effector variation has been shown to be important in other plant pathogens such as bacteria and fungi as a tactic for evading host recognition and resistance (181, 182).

To provide comprehensive biological insight and a tool for comparative analyses between different nematode species and populations of *H. glycines* in the absence of a reference genome, a *de novo* transcriptome assembly of early life stages was generated. An analysis of the transcriptome confirmed previous reports of microorganisms present within the nematode with molecular details and identified new parallels to other plant-parasitic nematode species. We then performed multiple analyses focused on effectors; both predicting novel effectors using a newly developed bioinformatics tool called N-Preffector that is not reliant on the presence of a signal peptide and investigating variation of previously identified stylet-secreted effector protein sequences. This allowed for the identification of an additional suite of novel effectors that may play a pivotal role in SCN infection and could serve as potential targets for future development of novel SCN control strategies.

2.2.3 Results

Transcriptome sequencing and assembly

To gain global insights into the transcriptomic response associated with the establishment of SCN infection, mRNA sequencing of pre-parasitic second-stage juvenile (ppJ2) and parasitic second-stage juvenile (pJ2) life stages infecting a resistant and susceptible host was conducted, yielding a total of 603.6 million paired 100 base reads. Following initial filtering steps and removal of reads mapping to the soybean genome, the final input for transcriptome assembly was 430 million reads. Trinity *de novo* transcriptome assembly resulted in a final assembly of 147,910 transcripts with a total assembly length of 46.7 Mb and estimated 23-fold transcriptome coverage. The average length of these transcripts was 658 base pairs (bp) with a N50 of 1,085 bp (Table 2.2.3.1). When translated, 78,625 resulting proteins were predicted. This transcriptome assembly was then assessed using BUSCO (Benchmarking Universal Single-Copy Orthologs, (183)). Based on the 429 single copy orthologs for eukaryotes, the SCN assembly is 68% complete, with an additional 14% of the orthologs represented in fragmented transcripts and the remaining 18% missing from the transcriptome.

Table 2.2.3.1. *de novo* transcriptome assembly statistics for the SCN early life stage assembly.

Number of transcripts	147,910
Total assembly length (Mb)	46.7
Number of trinity 'genes'	71,093
N50 (bp)	1,085
Maximum contig size (bp)	11,112 bp
Minimum contig size (bp)	201 bp
Average contig length (bp)	658 bp
Predicted proteins	78,625
BUSCO score	C: 68%, F: 14%, M: 18%

The assembly was generated from *H. glycines* pre-parasitic second-stage juvenile samples as well as parasitic second-stage juvenile samples from susceptible and resistant host interactions using the Trinity *de novo* transcriptome assembly tool. The transcriptome was assessed for completeness using the tool BUSCO (benchmarking universal single-copy orthologs) to identify complete (C), fragmented (F), and missing (M) sequences representing conserved orthologs found in all eukaryotes.

Annotation of transcripts

Transcripts from the *H. glycines* transcriptome were annotated following the Trinotate pipeline (184). Transcripts were first compared to GenBank, Swissprot, and TrEMBL databases using BLASTX, resulting in a total of 66,601 (45.03%) out of the 147,910 transcripts annotated using an e-value cutoff of 1e-5. When examining the species distribution of these significant hits, most transcripts hit to prior *H. glycines* database entries followed by animal-parasitic nematode species such as *Ascaris suum* and *Strongyloides ratti* (Figure 2.2.3.1). In total, 1315 species are represented in the BLASTX results representing a broad variety of genera. Other species of note in the annotated transcripts include *Cardinium* endosymbionts of *Encarsia pergandiella* and *Bemesia tabaci* as well as several soybean cyst nematode associated viruses (185-190). The virus sequences from the *H. glycines* PA3 population sequenced in this study are described by Ruark et al. (191) and the endosymbiont-associated transcripts were characterized in more detail as described below.

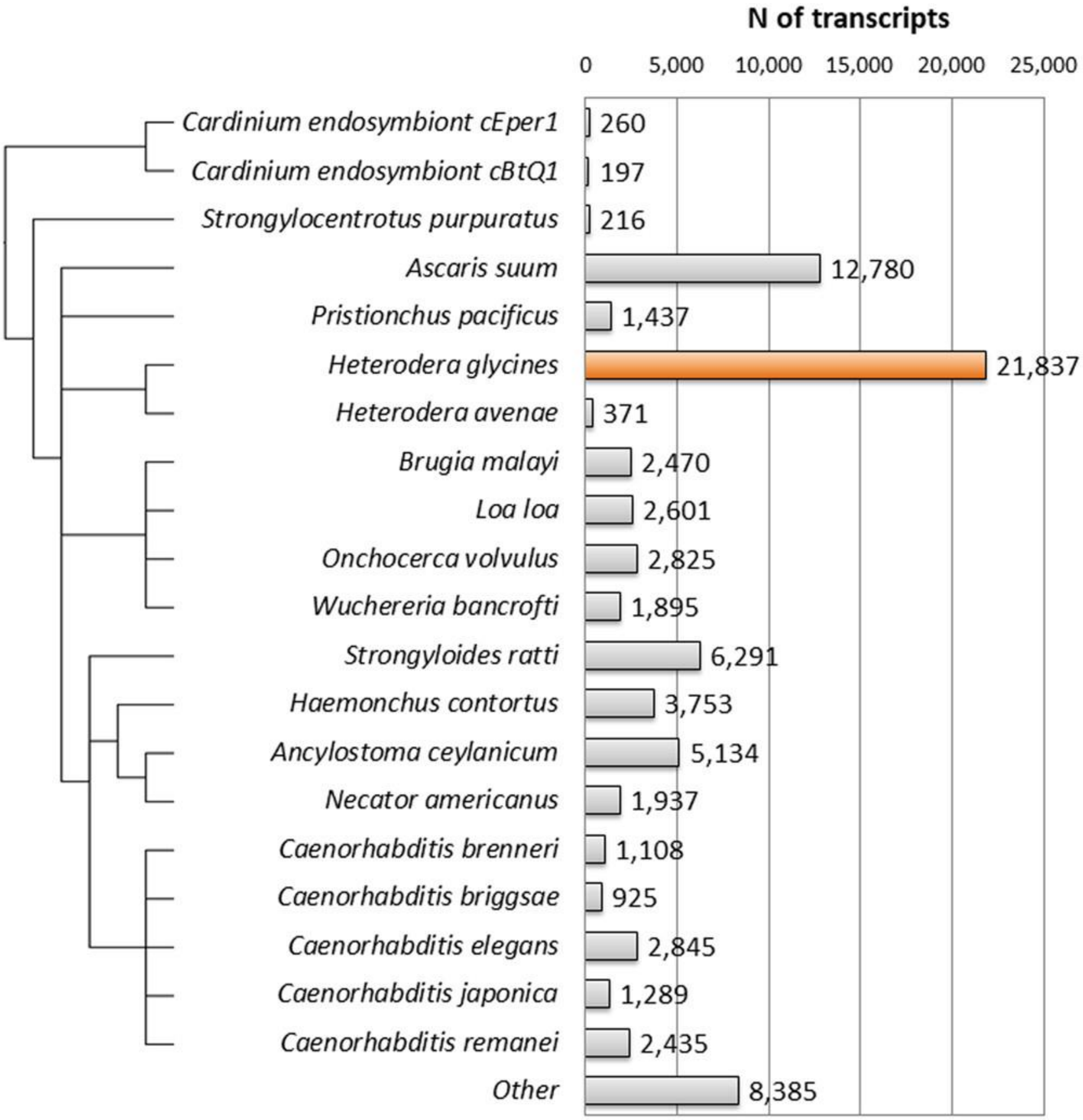


Figure 2.2.3.1. Species distribution of predicted homologues to *H. glycines*. Homologues were predicted using a BLASTX search against several protein databases at an e-value cutoff of 1e-5. The top 20 species with the most homologues are shown here. The resulting species evolutionary relationship was obtained from NCBI Taxonomy Browser [83] and visualized using IcyTree [84].

Transcripts were further compared to several nematode species with sequenced and annotated genomes representing free-living, animal-parasitic, and plant-parasitic trophic groups to identify potential overlap and genes that are uniquely shared between SCN and one other nematode species. The *H. glycines* transcriptome uniquely shares 76 potential homologs with *Bursaphelenchus xylophilus*, 313 homologs with *Meloidogyne hapla*, 200 with *M. incognita*, and 7,721 with *Globodera pallida*. In addition, the transcriptome shares 11 homologs with the free-living nematode *Pristionchus pacificus*, 28 with the free-living nematode *Caenorhabditis elegans*, and 84 homologs with the animal-parasitic nematode *A. suum* (Figure 2.2.3.2).

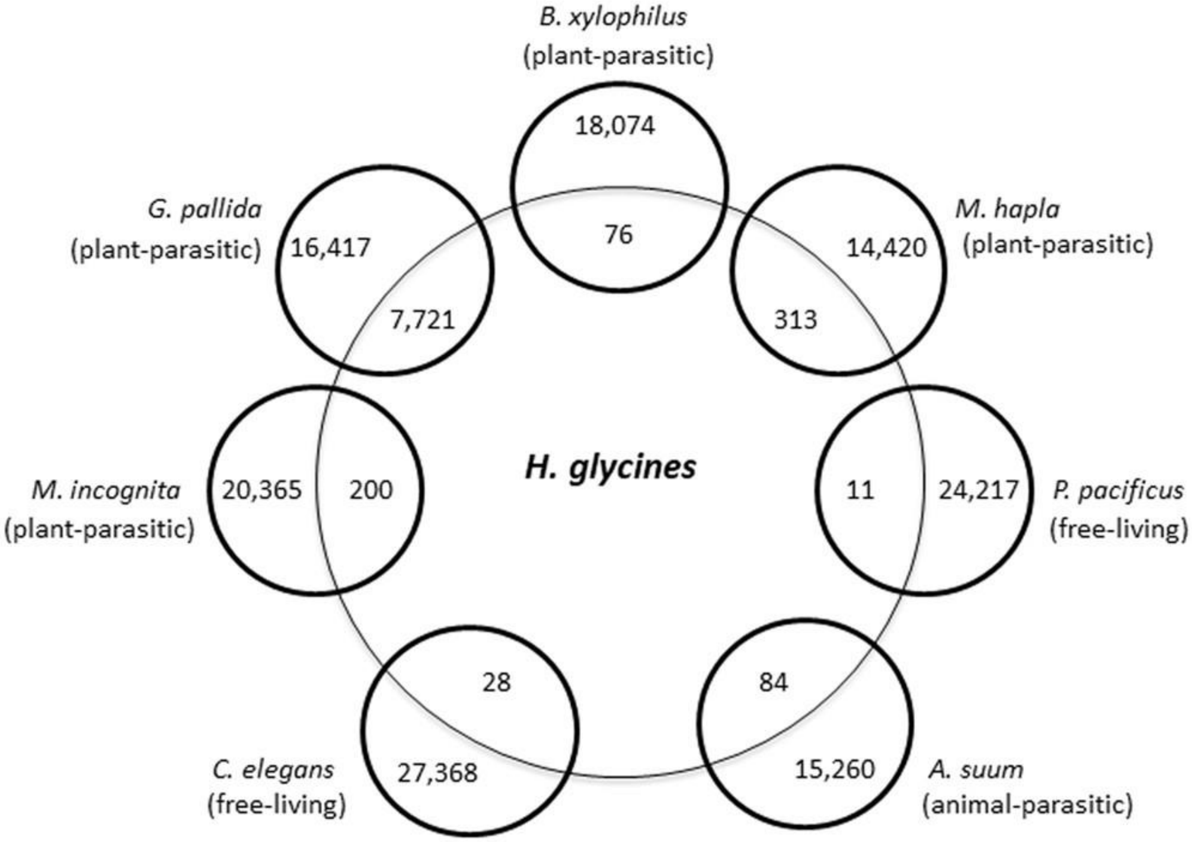


Figure 2.2.3.2. *H. glycines* orthologs in proteomes from sequenced nematodes with diverse feeding behaviors. The interior numbers represent predicted *H. glycines* proteins that only have orthologs identified in one of the seven other nematode species examined. Exterior numbers represent sequenced nematode proteins with no unique orthologs in the early parasitic *H. glycines* transcriptome.

Identification and GO annotation of endosymbiont-associated transcripts from the H. glycines transcriptome

Prior microscopic analysis of SCN indicated the presence of a bacterial endosymbiont (187, 190). Within the early parasitic SCN transcriptome we identified 468 transcripts annotated as endosymbiont-associated transcripts, all of which were confirmed by BLASTX mapping to the *Cardinium hertigii* proteome (Figure 2.2.3.3A). To further examine the potential functional significance of this inhabitant on SCN biology, GO terms were assigned to the 468 endosymbiont-associated transcripts using BLAST2GO, resulting in GO annotation of 328 of the 468 transcripts. Of those sequences with GO annotation within molecular function, the majority were involved in ATP binding, with 24% of the annotated transcripts falling into this category, followed by DNA (17%) and RNA (14%) binding (Figure 5.2.3.3B). The cellular compartment represented by the greatest number of transcripts was the ribosome (39%) (Figure 2.2.3.3C). The most significant biological processes represented among annotated transcripts were translation (14%) and transport (10%) (Figure 2.2.3.3D).

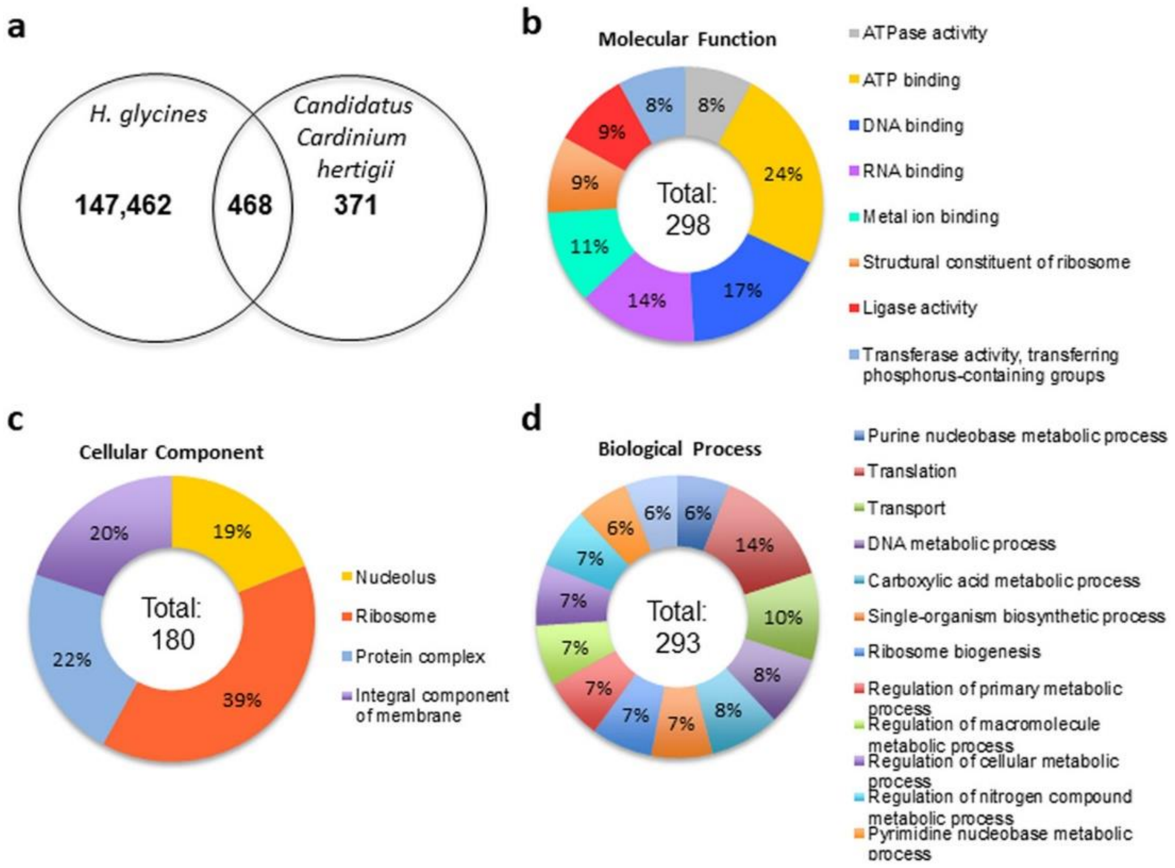


Figure 2.2.3.3. Identification and characterization of '*Candidatus* Cardinium hertigii'-associated transcripts within the *H. glycines* early life stage transcriptome. Transcripts from the *H. glycines* transcriptome were extracted and mapped against the proteome for *Candidatus* Cardinium hertigii to identify potential endosymbiont-associated transcripts, resulting in the identification of 468 of the 839 described proteins for this endosymbiont within the SCN early life stage transcriptome (A). Available gene ontology annotation was added to the endosymbiont-associated transcripts by BLAST2GO and grouped by the parent terms molecular function (B), cellular component (C), and biological process (D).

SCN stylet-secreted effector protein analysis

Effector proteins originating in the esophageal gland cells and secreted through the stylet play critical roles in the SCN infection process. Therefore, we first examined the 72 previously identified stylet-secreted *H. glycines* effectors (176-178) within the transcriptome. Of these, transcripts corresponding to each effector were identified using a BLASTN search, indicating that the transcriptome contained sufficient depth to detect expression of the known gland cell effector repertoire of SCN. An analysis of effector variation within the population was then performed. We first grouped the known effectors into stylet-secreted effector families (SSEFs) with greater than 70% sequence identity. To assess the level of variation of these known effectors within the sequenced *H. glycines* population, the predicted peptide sequences were mined for protein variants using BLASTP at a 1e-5 cutoff. Protein variants were identified for 69 of the 72 known effectors (Figure 2.2.3.4). The remaining three (17G06, 30C02, and GLAND9) were found to have single nucleotide insertions and/or deletions leading to a frame shift in the predicted peptide, resulting in a completely different peptide compared to the reference sequence, and consequently were not examined for sequence variation. A wide scope of variation was identified, with some effectors having over 70 predicted protein variants across the population (e.g., annexin 4F01), while others were limited to a single, highly conserved protein sequence (e.g., 7E05, protein unknown function).

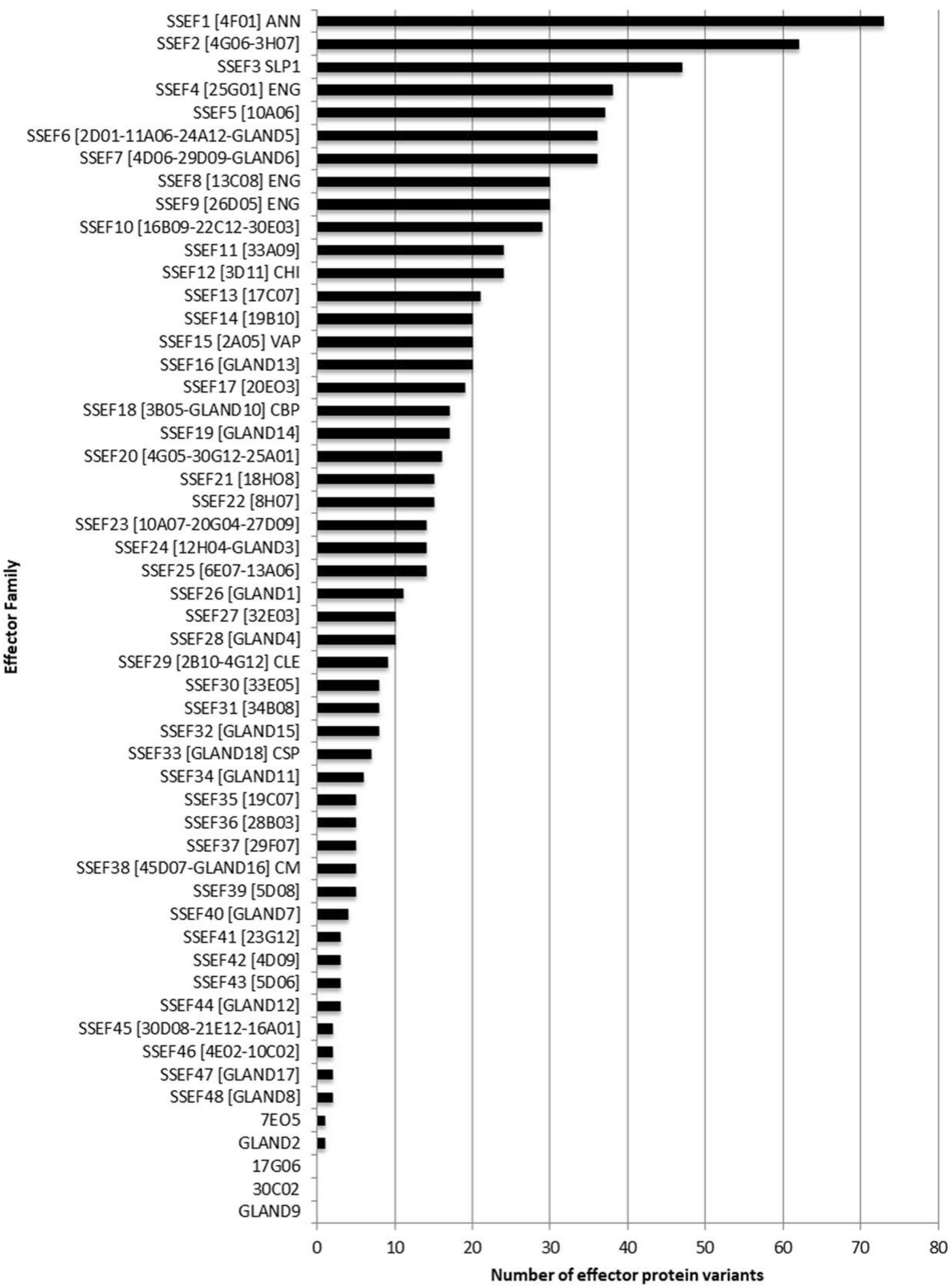


Figure 2.2.3.4. Variation of known effectors in the *H. glycines* early life stage transcriptome. Protein variants of previously published *H. glycines* effectors [17-19] were identified using a BLASTP search at a 1e-5 cutoff and counted. Known effector sequences with >70% amino acid identity were grouped into stylet-secreted effector families (SSEF) to facilitate the analysis. Available functional annotation for effector families is indicated as follows:

ANN=annexin-like; SLP1= SNARE-like protein 1; ENG=endoglucanase; CHI=chitinase; VAP=venom allergen-like protein; CBP=cellulose-binding protein; CLE= CLAVATA3/EMBRYO SURROUNDING REGION (CLE)-like; CSP=circumsporozoite protein; CM=chorismate mutase.

We then examined the expression of known SCN effectors during a compatible and an incompatible interaction to determine if the host environment influences the expression of any of these effectors. The effectors were split into two different groups (upregulated or downregulated) based on their expression pattern from the pre-parasitic second-stage juvenile (J2) stage to the parasitic J2 life stage and then compared across the two conditions. Most of the known SCN effectors followed the same pattern of expression across both comparisons, but the level of expression change was slightly reduced in the incompatible interaction. However, a subset of effectors exhibited an opposite trend of increased expression in the incompatible interaction, including members of SSEFs 1 [4F01], 9 [26D05], 17 [20E03], 22 [8H07], 45 [30D08, 21E12, 16A01], 39 [5D08], and 11 [33A09].

Effector alternative splicing analysis

To analyze alternative splicing (AS) as a potential mechanism of effector variation, we used the 72 previously identified stylet-secreted *H. glycines* effector candidates (176-178). Similar to the protein analysis of known SCN effectors, transcripts corresponding to each effector were identified using a BLASTN search in order to determine AS relationships. The major differences from the protein analysis were the use of a higher sequence similarity threshold (>85% identity) and the use of a gap penalty of 0. These two constraints were implemented to reduce false positives and improve true positives since gaps are expected to occur and should have a higher percent identity if AS occurs. In total, 395 AS transcripts were identified for the 72 previously known SCN effectors (Table 2.2.3.2), with the number of AS variants per each effector ranging from 1 to 38.

Using these 395 AS transcripts, differential expression analysis was conducted to determine statistically significant AS transcripts for comparison between the ppJ2 and pJ2 life stages as well as between two different host interactions in the pJ2 life stage, an incompatible and compatible interaction. In total, 129 AS transcripts representing 44 known SCN effectors were determined to be statistically significant with respect to host interaction groups and 276 AS transcripts representing 58 known SCN effectors were statistically significant with respect to life stages, with 127 overlapping transcripts (98.4%) between stages (Table 2.2.3.2).

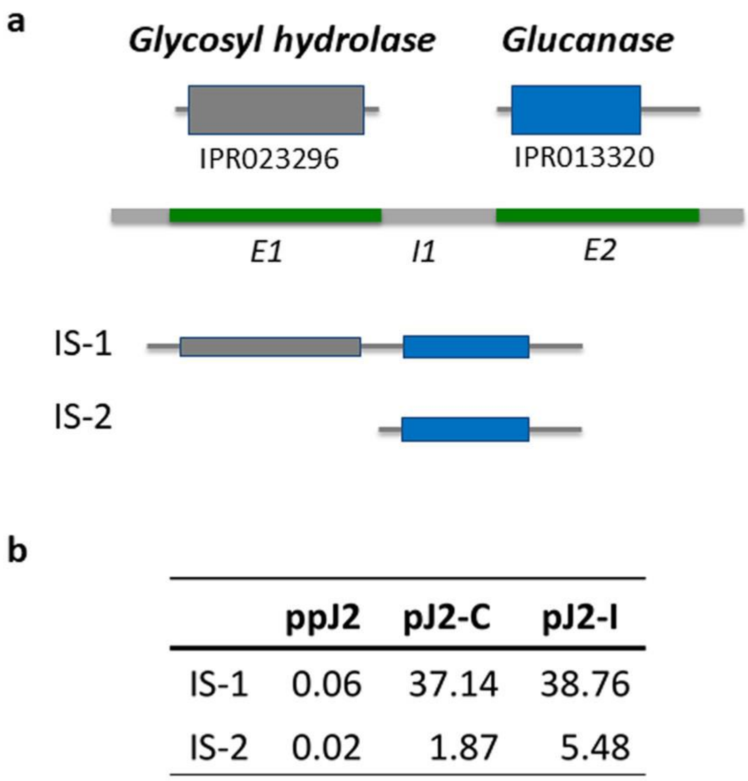
Table 2.2.3.2. Summary statistics for alternative splicing analysis of known SCN effectors.

	Known effectors	AS transcripts
Total	72	395
Significant for host interaction (compatible vs incompatible)	44	129
Significant for life stage (ppJ2 vs pJ2)	58	276

Alternative splicing analysis was performed on the previously published SCN effectors using the *de novo* transcriptome assembly. Splice variants were identified for known effectors and then analyzed for differential expression based on host interaction and nematode life stage.

To explore the effect that AS may have on protein function, functional domain analysis was conducted on the 395 AS transcripts. For this, we determined the changes in the functional domain architectures between specific AS isoforms. Since AS often alters the reading frame, all six reading frames were analyzed. Of the 72 effectors, 7 did not have any identified functional protein domains. In total, 513 protein functional domains for the remaining 65 effectors (7.9 domains per an isoform, on average) were identified using InterPro (192). For the 395 AS transcripts, 910 protein functional domains were identified (2.3 domain, on average), with 108 transcripts with no functional domains identified. When considering each effector and their AS transcripts, 37 out of 65 effectors (57%) had AS events that altered the predicted domain architecture. The 395 transcripts included 198 architectures with no change, 247 with at least one added functional domain, and 78 with one or more functional domains deleted. We note that the numbers of domain architectures do not add up to 395 because in some cases a transcript belonging to one effector was identified as the AS transcript from a different effector.

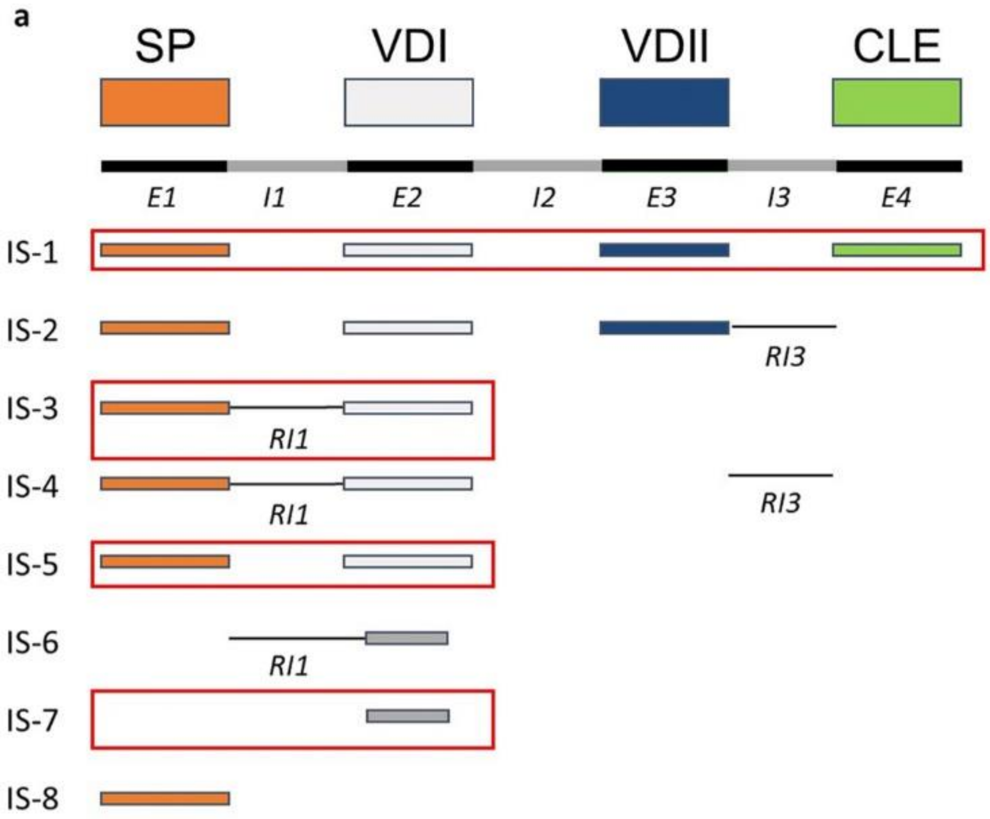
To analyze the functional changes in more detail, case studies of two effectors, GLAND13 and HgCLE (*Heterodera glycines* CLAVATA3/EMBRYO SURROUNDING REGION-like), were considered together with their AS transcripts. GLAND13 was chosen to demonstrate a simple example of a clear association between a protein function and AS variation due to the differentially spliced isoforms. On the other hand, HgCLE was chosen to demonstrate the structural and functional complexity that could be invoked through alternatively spliced isoforms. The architecture of the GLAND13 protein was predicted to have two functional domains that corresponded almost exactly to the two exons (Figure 2.2.3.5A). These two protein domains were associated with glycosyl hydrolase, a five-blade beta propeller domain, and concanavalin A-like lectin/glucanase protein domain (InterPro IDs: IPR023296 and IPR013320, respectively). Both of the functional domains are known to associate with metabolism. Our *de novo* AS analysis determined two different transcripts associated with GLAND13. The primary transcript included both protein domains, while the secondary transcript had exon 1 spliced out. It is possible for the reading frame to be altered if an AS event modifies the beginning of the gene. However, in our case the reading frame was preserved, which caused a removal of the glycosyl hydrolase domain, while leaving intact the glucanase domain. The functional implications of this removal are yet to be experimentally characterized. However, it was clear from the analysis that the primary transcript was important for life stage and was upregulated in the parasitic stage (p-value is 9.07E-11). Additionally, the secondary transcript was important for both life stage and host interaction (p-value 1.29E-5, Figure 2.2.3.5B). This transcript was upregulated in the parasitic stage, but to a greater extent in nematodes infecting a resistant host plant.



	ppJ2	pJ2-C	pJ2-I
IS-1	0.06	37.14	38.76
IS-2	0.02	1.87	5.48

Figure 2.2.3.5. Gene structure, protein functional domain architecture, and isoform protein products for GLAND13. Domain architecture and the retained protein domains in each of two isoforms, IS-1 and IS-2 (A). Expression of each isoform (transcripts per million) in pre-parasitic second-stage juveniles (ppJ2) and parasitic J2 (pJ2) life stages during a compatible (C) or incompatible (I) host interaction (B). The first isoform was significant for life stage change (p-value is 9.07E-11), while the second isoform was significant for both life stage and host interaction changes (p-value 1.29E-5).

HgCLE genomic architecture includes four exons that were consistent with four functional subunits: signal peptide, variable domain I, variable domain II, and the CLE domain (193). The N-terminal signal peptide is important for secretion of the peptide out of the nematode esophageal gland cell while variable domain I has been shown to function in targeting of the effector within the host plant cell (193). The HgCLE effector family [2B10-4G12] contains two known members with high levels of sequence conservation at the amino acid level, the only differences existing within the variable domains. The CLE domain is processed to release a small peptide that functions within the host plant as a ligand mimic (194). Based on the AS analysis, there were 8 transcripts associated with HgCLE. To improve the AS analysis, the corresponding HgCLE2 genomic DNA sequence was retrieved from NCBI GenBank (GenBank ID: FJ503005.1) and compared with these 8 transcripts (Figure 2.2.3.6A). While the genomic sequence was obtained from a nematode population that was different from the one used in this study, it was expected that there would be a significant sequence similarity between the gene sequence and the AS isoforms if there were AS events associated with intron retention. Transcript 1 corresponded to the full sequence of HgCLE2 retaining all four exons. Transcript 2 included exon 1-3, but retained intron 3 and lacked exon 4, which was associated with the CLE domain. Transcript 3 contained exon 1 and 2, but retained a modified version of intron 1. Transcript 4 was similar to transcript 3 except intron 3 was retained. Transcript 5 included just exon 1 and 2. Transcript 6 included modified versions of intron 1 and exon 2. Transcript 7 included a modified version of exon 2. Transcript 8 included exon 1. With respect to the differential expression analysis, transcript 1, 3, 5 and 7 were statistically significant (p-values ranging from 5.50E-04 to 9.638E-05) for life stage and host interaction, transcripts 2 and 4 were statistically significant (p-value 5.27E-07 and 4.96E-10) only in regards to the life stage, and transcripts 6 and 8 were not differentially expressed between any group (Figure 2.2.3.6B).



b

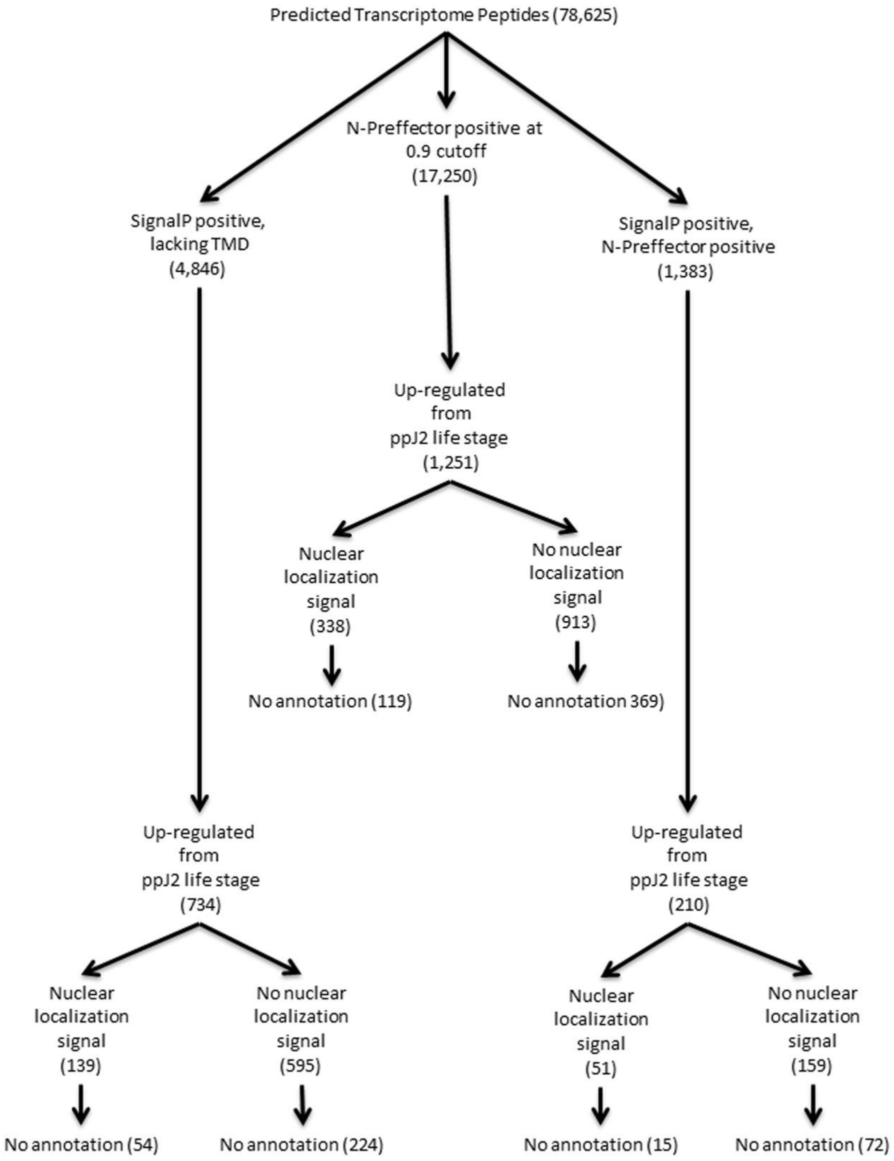
	ppJ2	pJ2-C	pJ2-I
IS-1	0.71	930.61	705.41
IS-2	0.18	82.05	74.32
IS-3	0.02	4.13	1.65
IS-4	0.00	2.08	1.47
IS-5	0.00	2.85	1.53
IS-6	0.00	0.36	0.34
IS-7	0.15	17.40	10.25
IS-8	0.13	0.24	0.29

Figure 2.2.3.6. Gene structure, protein functional domain architecture, and isoform protein products for HgCLE2. Domain architecture and the retained protein domains in each of eight isoforms, IS-1 and IS-8 (A). Shown in red are the retained introns. Each retained intron was modified as a result of AS. Dark grey boxes correspond to a modified VD1 domain due to AS. Expression of each isoform (transcripts per million) in pre-parasitic second-stage juveniles (ppJ2) and parasitic J2 (pJ2) life stages during a compatible (C) or incompatible (I) host interaction (B). Red boxes highlight transcripts that were statistically different for both life stage and host interaction groups.

Novel effector prediction

We then performed a comprehensive effector analysis on the SCN transcriptome. Effectors were predicted using two separate pipelines, and then the results were compared to determine the overlap of each pipeline (Figure 2.2.3.7). The first of these pipelines relies on the presence of a signal peptide and follows the method used in previous studies for the prediction of putative stylet-secreted effectors (177, 178). This pipeline predicted 4,846 putative effectors. To identify putative new effectors with higher confidence, we focused on genes upregulated from the pre-parasitic J2 to parasitic J2 life stage and analyzed the sequences for the presence of a nuclear localization signal (NLS). A NLS combined with an N-terminal signal peptide is a strong indicator for localization of these effectors into host cell nuclei where they can play a variety of functions including regulation of plant defense responses (179, 195). Following these filtering steps, this pipeline predicted 734 effector candidates, including 139 nuclear localization signal (NLS)-positive effector candidates up-regulated from the pre-parasitic J2 to the parasitic J2 life stage (Figure 5.2.3.7). The 72 known SCN effector proteins, known to contain signal peptides, were re-discovered at a rate of 74% using this pipeline. This pipeline is reliant upon the presence of a N-terminal signal peptide, which may not be present if the N-terminus is absent from the transcript. This is reflected in the fact that several known SCN effectors were not recovered despite their nucleotide sequences being present within the transcriptome. A second pipeline independent of the presence of a signal peptide was performed using N-Preffector, a machine learning algorithm trained on known nematode and bacterial effectors. The N-Preffector-based pipeline predicted 1,251 putative effectors, including 338 NLS positive effector candidates up-regulated from the pre-parasitic J2 to the parasitic J2 life stage (Figure 2.2.3.7). In this pipeline, 67% of the known SCN effectors were re-discovered. When the two pipelines were compared, 210 effector candidates were found to be overlapping, including 51 NLS positive candidates (Figure 2.2.3.7).

Many of these sequences have little or no annotation available. Among those sequences with available annotation are many homologs of effectors from other plant-parasitic nematodes that were not previously identified or characterized in *H. glycines*. These include effectors such as glutathione synthetase (196) and members of the SPRYSEC family (197).



Predicted Transcriptome Peptides (78,625)
N-Preffector positive at 0.9 cutoff (17,250)
SignalP positive, lacking TMD (4,846)
SignalP positive, N-Preffector positive (1,383)
Up-regulated from ppJ2 life stage (1,251)
Nuclear localization signal (338)
No nuclear localization signal (913)
No annotation (119)
No annotation 369)
Up-regulated from ppJ2 life stage (734)
Up-regulated from ppJ2 life stage (210)
Nuclear localization signal (139)
No nuclear localization signal (595)
Nuclear localization signal (51)
No nuclear localization signal (159)
No annotation (54)
No annotation (224)
No annotation (15)
No annotation (72)

Figure 2.2.3.7. Secreted effector protein prediction in the early life stage transcriptome of *H. glycines*. Predicted peptides from the transcriptome were put through two separate pipelines to identify candidate effectors. One pipeline utilized prediction of a signal peptide and lack of a predicted transmembrane domain (TMD) while the other utilized N-Preffector, a machine learning algorithm. Numbers shown here are predicted peptides remaining after each step in the pipeline.

2.2.3 Discussion

In this study, we sequenced the transcriptome of the early life stages of the plant-parasitic nematode *Heterodera glycines,* including the infective (pre-parasitic) second-stage juvenile (J2) life stage and the parasitic J2 life stage in two different host conditions, resistant and susceptible. We then carried out a *de novo* transcriptome assembly with an emphasis on assessing the level of variation of known effectors within a single population and identifying novel secreted effectors within *H. glycines* that may play important roles in establishing a parasitic interaction with its host, soybean. The resulting transcriptome from these samples consisted of nearly 150,000 transcripts encoding 78,625 predicted proteins. There are several possible explanations for the large number of transcripts identified. First, to generate the transcriptome a large population of nematodes was sequenced. The inherent genotypic heterogeneity present within the population may lead to many variants of the same gene being represented within the transcriptome. In addition, following transcriptome assembly no expression threshold was applied. This was done to capture any rare or lowly expressed transcripts within the population. Of the 147,910 transcripts contained within the *H. glycines* transcriptome, 66,601 (48%) were annotated based on BLAST homology. Many of these potential homologues existed in other nematode species, including plant- and animal-parasitic nematodes. In addition, some transcripts showed homology to a bacterial endosymbiont from the genus *Cardinium*, of *Encarsia pergandiella*, a parasitic wasp, and *Bemesia tabaci*, a whitefly. Previous work identified this endosymbiont and characterized it as *Candidatus* Paenicardinium endonii, later renamed to *Candidatus* Cardinium hertigii (187, 198). However, little is known about the function of this endosymbiont and what role it may play, if any, in plant parasitism. Related endosymbionts found in insects and arachnids have been shown to have prominent impacts on their hosts, leading to changes in host reproductive capacity and also modulating host immunity (199). To better understand the function of a putative endosymbiont in *H. glycines* all transcripts associated with this endosymbiont were identified and extracted from the transcriptome, representing a majority of the characterized sequences for this endosymbiont. Those transcripts identified were primarily associated with metabolic processes, which may contribute to both nematode and endosymbiont metabolism. Further studies into the function of this endosymbiont and any effect on parasitism removing the endosymbiont has will be vital in elucidating what role it plays inside the nematode. In addition to a bacterial endosymbiont, several putative homologs from viruses were also identified. Previously, researchers found representative viruses from the *Bornaviridae, Rhabdoviridae,* and *Bunyaviridae* families contained within *H. glycines* (186, 191). Thus, there appears to be a significant microbial community active within *H. glycines* that has until now remained largely unexplored. Further examination of these organisms could reveal vital connections that can be exploited for improving resistance against SCN.

Stylet-secreted effectors represent a key component of the plant-nematode interaction, serving a wide variety of functions required for successful invasion and establishment of the nematode feeding site. Previous studies have identified a suite of these effectors using microaspiration techniques to isolate the contents of the esophageal gland cells where these genes are expressed (177, 178). These studies then prioritized potential effectors based on those sequences possessing a signal peptide and lacking a transmembrane domain. Despite previous knowledge about the effector repertoire of SCN, very little is known about the structure of these sequences within a population, specifically how these sequences vary from one individual to another. To address this question, we undertook an effector variation analysis within the transcriptome, identifying putative variants of known effectors and examining their level of variation within the population. Within the SCN transcriptome, predicted sequence variants of known effectors ranged from over 70 (effector 4F01) down to one (effectors 7E05 and GLAND2). This effector variation may be a result of different alleles being present in the population and/or reflect variation in the copy numbers of genes encoding related effectors. The level of variation of these effectors is likely related to the function of the effector in question. For example, a highly variable effector such as 4F01 may be under constant selection pressure to avoid host recognition, resulting in a wide level of variation across the gene pool. A prior study demonstrated that 4F01 might function as a mimic of host plant annexins to promote successful plant infection (200). By contrast, effectors with limited variation across the population are likely constrained by their function. It would be interesting to see how a highly virulent population or a field population compares to the highly inbred population used here for sequencing. Certain effectors may be expanded or reduced depending on the population and host selection pressure. Effectors with a very low number of variants across populations may represent key elements of infection that could be targeted for further study in the attempt at identifying a novel source of broad spectrum SCN resistance.

One potential mechanism of gene regulation that can introduce variation into genetic sequences is alternative splicing. Previous work has identified alternative splicing in stylet-secreted effectors from SCN on an individual basis and demonstrated that expression of these variants was impacted by the life stage of the nematode (201, 202). As sequence data become available for plant-parasitic nematodes, these types of analysis can be expanded to larger scales. For example, a comprehensive analysis of alternative splicing events conducted across the effector complement of the potato cyst nematode *G. pallida* using the sequenced genome found that 38% of these genes undergo alternative splicing and that certain families of effectors show increased occurrence of splicing relative to others (203). With the early parasitic transcriptome generated in this study we were able to perform a large-scale alternative splicing analysis on the known effectors of SCN and identified significant changes in the expression of alternatively spliced transcripts for a majority of the effectors between the ppJ2 and pJ2 life stages as well as between compatible and incompatible host interactions. Changes in effector splicing across life stages as the nematode begins infection may be important for altering the protein function or activity to facilitate migration and establishment of the nematode feeding site. We then examined alternative splicing of effectors between a compatible and an incompatible host interaction, identifying a smaller subset of effectors with significant expression changes between these two conditions. These splice variants may be useful once again for altering function and activity of the effectors, potentially after being triggered by perception of host resistance by the nematode. By expressing an alternate version of the effector sequence, the nematode may avoid direct recognition of the host or recognition of the function that effector performs. Once additional populations of SCN have been sequenced it will be interesting to see whether these splice variants are involved in virulence on other sources of SCN resistance and if these can be targeted to improve overall resistance to this pathogen.

We also mined the early parasitic transcriptome to identify additional effectors expressed within *H. glycines* using the SignalP predictive tool, as well as a novel pipeline called N-Preffector. The use of N-Preffector allowed for the identification of an entirely new class of effectors not necessarily containing a signal peptide. Examples of secreted effectors lacking a signal peptide have been shown in other plant-parasitic nematode species such as *G. rostochiensis*, where they have been shown to play a role in disrupting host reactive oxygen species production (204, 205). These effectors may contain a previously unknown secretion signal or utilize a novel secretion pathway in order to be secreted. Between the two pathways utilized for effector discovery, 86% of known SCN effectors were re-discovered within the early parasitic SCN transcriptome. The remaining 14% were not re-discovered either due to truncated sequences relative to the reference sequence or a change in the predicted protein sequence between the transcriptome and reference sequence. It is interesting to note that 47% of these effectors were identified by both pipelines, but included different effectors. This illustrates the potential advantage of using both pipelines to accurately detect all possible effectors including those that one pipeline may not identify. The signal peptide-dependent method is excellent at predicting putative effectors, but misses out on transcripts that may be truncated or simply lack the signal peptide, which can be complemented using the N-Preffector pipeline. It should also be noted that in this study an expression change between life-stages was used as a parameter for effector prediction and to limit the overall number of false positives. For this reason, the possibility exists that some putative effectors with very low expression levels may have eluded discovery. One example is HgCLEB, which is expressed at low levels and therefore was not discovered in the effector pipeline, but later identified using a targeted search of the transcriptome (206).

The novel effector candidates identified by these two pipelines represent a set of genes for downstream expression and functional analysis to investigate the interaction between SCN and soybean. Many of these sequences have little or no annotation available, much like the original gland isolated effector sequences obtained for *H. glycines* (176-178). These novel effector sequences may play pivotal roles in nematode parasitism and will require more in depth functional studies to determine their function. Among those sequences with available annotation are many homologs of effectors from other plant-parasitic nematodes that were not identified or characterized in *H. glycines* previously. Included in this category are genes such as the glutathione synthetase family, the novel *G. rostochiensis* effector E9, and candidates showing homology to the SPRYSEC family of effectors from *G. rostochiensis*. Glutathione synthetases have many potential roles in the interaction between the nematode and host plant. In the interaction of the root-knot nematode *M. incognita* it was found that glutathione is needed for successful infection of the host plant *Medicago truncatula* (195). In addition, glutathione synthetase genes were found to be greatly expanded in the genome of the potato cyst nematode *G. pallida*, where these genes are theorized to be involved in protection of the nematode from antioxidant proteins as well as potentially in nematode nutrition (167). Several transcripts annotated as glutathione synthetase also contained a secretion signal, something that differentiates them from glutathione synthetases found in animal parasites that may function within the nematode. The putative effector E9 has been identified in both *G. rostochiensis* and *G. pallida* and was confirmed to be expressed in gland cells via *in situ* hybridization (203, 207). Thus far little is known about the function of the E9 effector, other than it being expressed in the gland cells of *Globodera* species. The SPRYSEC effectors on the other hand have been heavily investigated in the *Globodera*-tomato pathosystem, with demonstrated roles in the suppression of plant immune responses (197, 207). To date, SPRYSEC effectors have not been identified in the genome sequence of root-knot nematodes (167); however, entries in non-redundant sequence databases suggest they may be present in other cyst nematodes and lesion nematodes (208). Thus, these could be very interesting candidates for comparative analysis across virulent populations of SCN to determine whether or not they play the same role as in *Globodera spp*. Another effector candidate of note is a putative secreted calreticulin. A calreticulin secreted by *M. incognita* is necessary for successful infection and may play a role in suppression of plant defenses; functions that may be retained in *H. glycines* (209). Another nematode effector homolog group identified in the transcriptome involved in suppression of host defenses are the C-type lectins (CTLs) from *Rotylenchus reniformus*. These effectors were identified in the *R. reniformus* transcriptome and subsequently shown to be expressed in the hypodermis of parasitic stages of the nematode (210). It is hypothesized that these effectors are involved in protecting the nematode from environmental stress. While these homologs are all predicted to have the same function in *H. glycines* as their originating species, further functional characterization is necessary to confirm this.

Interestingly, we identified several effector candidates with sequence similarity to proteins originating in plants and other organisms. These included multiple effector candidates with homology to members of the plant RING/U-box superfamily of proteins. These proteins are typically involved in protein modification and regulation of plant pathways, including defense responses and regulation of cell death (211). Nematode mimics of these proteins may be involved in manipulation or suppression of host defense pathways in order to allow successful establishment of the feeding site. Among the identified effector candidates are also several homologs related to plant metabolism and cell wall degradation. These included arabinosidase, fructosidase, glycoside hydrolase, and expansin. These cell wall modifying proteins have been shown to aid in the loosening and degradation of polysaccharides present in the plant cell wall (212-214) and have been identified from other plant-parasitic nematodes where they play a crucial function in migration and establishment of the nematode feeding site (215, 216). Therefore, these plant mimics all represent avenues of study to be pursued in order to better understand the interplay between SCN and its plant host, soybean.

In conclusion, a *de novo* transcriptome of the pre-parasitic and parasitic second-stage juvenile life stages of *H. glycines* has been generated, annotated, and comprehensively mined for putative effector sequences. Within this transcriptome novel effector candidates were identified utilizing a new prediction tool not reliant on sequences possessing a signal peptide, N-Preffector.

In addition, the level of variation of previously identified *H. glycines* effectors was examined for the first time at the population level and identified highly conserved and highly variable effectors. Finally, this transcriptome provides a useful genetic resource that will aid in annotation of the SCN genome. Combining these data will provide insights into the biology of SCN with the hopes of discovering new ways to combat this pathogen.

2.2.4 Materials & Methods

Nematode cultivation and isolation

The SCN inbred population PA3 (HG Type 0) was propagated under greenhouse conditions on susceptible soybean Williams 82 or EXF63. Freshly hatched pre-parasitic second-stage juveniles (ppJ2) were inoculated onto 10-day old seedlings of the susceptible host or the resistant host (cv. Forrest) and the inoculated plants were placed in the greenhouse. The remaining ppJ2 nematodes were pelleted by centrifugation and flash-frozen in liquid nitrogen and stored at -80°C prior to RNA isolation. Five days post-inoculation, parasitic second-stage juveniles (pJ2) nematodes were isolated from the roots by blending the roots for 30s in a kitchen blender. Following this, the root homogenate was poured over a nested stack of sieves with pore sizes of 850µm, 250µm, and 25µm before purifying the nematodes from the sample using sucrose centrifugal flotation (217). Samples were frozen in liquid nitrogen and stored at -80°C prior to RNA isolation**.**

RNA isolation and sequencing

RNA was isolated from frozen nematode pellets using the PerfectPure Fibrous Tissue Kit (5Prime) and a modified version of the manufacturer’s extraction protocol. Tissue was homogenized in 30 second intervals in the provided lysis solution containing 0.5 µM TCEP using a bead beater and 1.0 mm zirconia beads, followed by a 30 second incubation on ice. This was repeated three times. The sample was centrifuged briefly at room temperature before transferring the supernatant to a fresh tube. Following lysis and homogenization, 10 µl of the provided Proteinase K was added and the sample was allowed to incubate on ice for 10 minutes, after which the manufacturer's protocol for RNA purification was followed. RNA quality was determined using a Fragment Analyzer (Advanced Analytical) and quantified using a Qubit Fluorometer prior to library preparation. RNA-seq libraries (ppJ2, pJ2 infecting susceptible host, pJ2 infecting resistant host) were constructed using the TruSeq mRNA Stranded Library Prep Kit (Illumina) and sequenced on the Illumina HiSeq 2500 platform in a paired-end manner (2x100 for ppJ2 and pJ2-Compatible samples and 2x50 for pJ2-Incompatible sample). Library preparation and high-throughput sequencing services were performed at the University of Missouri DNA Core Facility. Three biological replicates of each sample were sequenced.

De novo transcriptome assembly

Prior to assembly, raw reads from these libraries were filtered using Trimmomatic (218) to remove low quality reads. The remaining reads were paired and orphan reads discarded. High quality paired-end reads were used as input for transcriptome assembly. *De novo* transcriptome assembly was completed using the de Bruijn graph-based tool Trinity (219). As part of the assembly process, an *in silico* read normalization step was performed. Assembly quality was then assessed by mapping raw reads back to transcripts using Bowtie2 (220) at default parameters.

Transcriptome annotation and quantification

The transcriptome was annotated following the established Trinotate pipeline (219). Homology searches were performed against the protein sequences contained in Genbank (221) and UniProt (222) databases using BLASTX at an evalue cutoff of 1e-5 (223). Transcripts were translated into protein using TransDecoder, a component of Trinity (219). HMMER and Pfam databases were used to predict protein domains contained within each transcript (224, 225). Presence of a signal peptide was determined using SignalP version 4.0 and TMHMM version 2.0 was utilized to identify predicted transmembrane domains (226, 227). The resulting annotation information was then combined and pooled into a SQLite database. In addition, sequenced nematode genomes were leveraged to identify potential homologs within the transcriptome. For this, predicted protein datasets from the genomes of *Bursaphalenchus xylophilus, Meloidogyne hapla, Meloidogyne incognita, Globodera pallida, Pristionchus pacificus, Ascaris suum,* and *Caenorhabditis elegans* were downloaded from WormBase (http://ws204.wormbase.org/) and used (228). BLASTP hits from the *H. glycines* transcriptome with e-values less than 1e-5 were considered potential homologs. Lists of potential homologs from each of the seven species examined were then compared and contrasted to determine uniquely shared homologs between the sequenced nematode and *H. glycines*.

For quantification and differential expression analysis, reads from the libraries used for assembly were mapped and quantified using RSEM (229) to determine transcript abundance. RSEM was utilized as it has been shown to correlate well with RT-qPCR measurements and produce expression values with high accuracy (230). Following quantification, differential expression analysis was conducted using edgeR (231), identifying all genes with a minimum 4-fold expression difference and under a p-value cutoff of 0.001 between any of the samples.

Identification of endosymbiont sequences within the H. glycines transcriptome

The entire *de novo* early parasitic transcriptome for *H. glycines* was mined for transcripts related to the endosymbiont "*Candidatus* Cardinium hertigii". All transcripts annotated with the species designation 'Cardinium endosymbiont' were extracted from the transcriptome and combined into a file. A database was then constructed from the complete proteome of the closest available sequenced bacterial isolate, *Cardinium hertigii* cEper1 isolated from *Encarsia pergandiella* (188). Then all putative *Cardinium*-associated sequences were mapped against the proteome database using BLASTX at an e-value cutoff of 1e-5 to confirm their identity as putative endosymbiont-associated transcripts. The resulting transcripts were then used for gene ontology analysis.

Gene ontology analysis of endosymbiont-associated sequences from the H. glycines transcriptome

Gene ontology (GO) analysis was performed to identify the putative function of endosymbiont-associated sequences within SCN. To do this endosymbiont-associated sequences from the SCN transcriptome were used in the research tool BLAST2GO (232). This tool uses a similarity searches to assign GO annotation to sequence data lacking well-characterized GO annotation. In BLAST2GO BLASTX was performed at an e-value cutoff of 1e-5 and the top available BLAST hit used to pull available GO annotation. Once available GO annotation was assigned to the 468 endosymbiont-associated transcripts the results were examined for their potential role in SCN biology.

Variation of known SCN effectors

The protein sequences for the 72 known SCN effector sequences (176-178) were aligned using MUSCLE (233) and then a maximum likelihood tree was constructed based on sequence homology in MEGA7 (234). MUSCLE (multiple sequence comparison by log-expectation) is a high accuracy tool for protein alignment. Effectors with bootstrap values greater than 50 were grouped into stylet-secreted effector families (SSEFs). Predicted transcript peptide sequences from the SCN transcriptome were then mapped to known SCN effector protein sequences using BLASTP at an e-value cutoff of 1e-5 and quantified for each known effector. Variants of known SCN effectors in a SSEF were pooled for quantification.

Effector alternative splicing analysis

De novo alternative splicing analysis represents a challenging task since a complete *H. glycines* genome is not available to assess exon and intron relationship (232). However, it is possible to associate known genes of interest and build associated relationships to infer alternative splicing events by comparing known regions of overlap and extract exons associated with specific alternative splicing isoforms. This alternative splicing analysis relies on the transcripts that are assembled with the Trinity pipeline (184). The alternative splicing quantification is then carried out with the *kallisto* tool (235) using the preprocessed reads and the pseudo alignment on the assembled transcripts, which allows the analysis to be computationally more efficient, without losing its quality. Using these quantified transcripts, *sleuth* tool was employed to determine statistically significant differentially expressed transcripts (236). From the list of 72 known SCN effector genes, the inferred alternative splicing relationship is built based on significant overlap between effector sequences and transcripts, defined by sequence identity of greater than 85%. The overlap and sequence identity are determined using the BLASTN tool, with the gap penalty parameter set to 0 (237). This high sequence identity threshold is used because a true alternative splicing event is expected to have a significant exon overlap between the effector sequences and transcripts. The reason that a higher identity threshold is not used is because the SCN population used as a source for the effector genes is different from the SCN population used as a source for transcriptomic data obtained in this study. Combining the high sequence identity threshold and zero gap penalty in the BLASTN search, thus, allows for alternative splicing events of known exons from the effector genes to be identified, while not allowing the discovery of new relationships. New relationships that will be missed due to the data and methodology limitation are primarily the intron retention events and require the assembled genome as a reference. Using the identified associated alternative spliced transcripts, protein functional analysis is done by predicting the domain architectures and characterizing protein domains using InterPro (192). Since it is expected for the reading frame to change, all 6 reading frames (forward and reverse) are assessed for the domain architectures and protein functions. In summary, this approach allows one to functionally characterize the differential expression changes for alternative spliced transcripts. These functionally characterized differentially expressed transcripts were compared between different nematode life stages and host interactions.

Effector prediction

The effector prediction pipeline started with all predicted peptides from the SCN transcriptome. First, sequences represented in the gland cell transcriptome were subjected to two different prediction tools: SignalP (226) and N-Preffector, developed in this study. For the SignalP-based prediction, peptides were run through SignalP 4.0 and TMHMM (227) to predict signal peptides and transmembrane helices, respectively. Predicted peptides containing a signal peptide and lacking a transmembrane domain were then filtered based on their expression between the ppJ2 and pJ2 life stages of the nematode, with those peptides showing a minimum 4-fold up-regulation into the pJ2 life stage retained. Finally, nuclear localization signals were predicted using NLStradamus (238). For N-Preffector based prediction, predicted peptides were run through a machine-learning algorithm trained on 72 known *H. glycines* effector sequences and 150 known non-effector sequences from *H. glycines* in addition to the original sequences (gram negative bacteria) in which the Preffector model was trained (239). For each protein sequence, N-Preffector calculates a vector of length-invariant features; the feature vector is then used as an input for the classification model. Feature categories that were considered are: residue composition, sequence/structure information, and physico-chemical properties of proteins. To select highly correlated features with the class and not correlated with each other, Preffector utilizes the correlation-based feature selection (CFS) method (240). Our goal was to minimize the number of proteins erroneously misclassified as effectors, *i.e.*, false positives, while trying to maximize the number of predicted real effectors, using the same exact protocol utilized in Preffector. N-Preffector achieves this through a more stringent classification criterion. Given an SVM model *M* and a training data of size *n*, for each training example x_k, let $f_k \in [-1, -1]$ be its decision value predicted by the SVM model, and $y_k \in \{+1, -1\}$ be its true annotation of being an effector or non-effector. Given the SVM model *M*, the prediction probability for a training example x_k is defined as

$$p_k^{(i)} = \frac{1}{\left(1+\exp\left(A^{(i)} f_k^{(i)} + B^{(i)}\right)\right)}. \qquad (1)$$

The coefficients $A^{(i)}$ and $B^{(i)}$ are estimated during the SVM training process by minimizing the log-likelihood function. Those peptides predicted by N-Preffector at or above a 0.9 confidence score cutoff were then filtered based on expression, retaining peptides with a minimum 4-fold up-regulation from the ppJ2 to the pJ2 life stages. Nuclear localization signals were then predicted using NLStradamus for the remaining peptides (238).

Data availability

Raw sequence reads are available under the Short Read Archive (SRA) accession no. SRP122521. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GFZZ00000000. The version described in this paper is the first version, GFZZ01000000.

2.3 Biological classification with RNA-Seq data: Can alternative spliced transcript expression enhance machine learning classifier? (241)

2.3.1 Abstract

The extent to which the genes are expressed in the cell can be simplistically defined as a function of one or more factors of the environment, lifestyle, and genetics. RNA sequencing (RNA-Seq) is becoming a prevalent approach to quantify gene expression and is expected to gain better insights to a number of biological and biomedical questions, compared to the DNA microarrays. Most importantly, RNA-Seq allows to quantify expression at the gene and alternative splicing transcript levels. However, leveraging the RNA-Seq data requires development of new data mining and analytics methods. Supervised machine learning methods are commonly used approaches for biological data analysis and have recently gained attention for their applications to the RNA-Seq data.

In this work, we assess the utility of supervised learning methods trained on RNA-Seq data for a diverse range of biological classification tasks. We hypothesize that the transcript-level expression data is more informative for biological classification tasks than the gene-level expression data. Our large-scale assessment is done through utilizing multiple datasets, organisms, lab groups, and RNA-Seq analysis pipelines. Overall, we performed and assessed 61 biological classification problems that leverage three independent RNA-Seq datasets and include over 2,000 samples that come from multiple organisms, lab groups, and RNA-Seq analyses. These 61 problems include predictions of the tissue type, sex, or age of the sample, healthy or cancerous phenotypes and, the pathological tumor stage for the samples from the cancerous tissue. For each classification problem, the performance of three normalization techniques and six machine learning classifiers was explored. We find that for every single classification problem, the

transcript-based classifiers outperform or are comparable with gene expression-based methods. The top-performing supervised learning techniques reached a near perfect classification accuracy, demonstrating the utility of supervised learning for RNA-Seq based data analysis.

2.3.2 Introduction

Ever since the intrinsic role of RNA was proposed by Crick in his Central Dogma (6), there has been a desire to accurately annotate and quantify the amount of RNA material in the cell. A decade ago, with the introduction of RNA sequencing (RNA-Seq) (242), it became possible to quantify the RNA levels on the whole genome scale using a probe-free approach, gaining insights into cellular and disease processes and illuminating the details of many critical molecular events such as alternative splicing, gene fusion, single nucleotide variation, and differential gene expression (232). The basic assessment of RNA-Seq is focused on utilizing the data for differential gene expression between the groups of biological importance (158). However, there are additional patterns that can be elucidated from the same raw sequencing data by extracting the expression levels of the alternatively spliced transcripts (243).

Alternative splicing (AS) of pre-mRNA provides an important means of genetic control (37, 244). It is abundant across all eukaryotes and even occurs in some bacteria and archaea (48-50). AS is defined by the rearrangement of exons, introns, and/or untranslated regions that yields multiple transcripts (74). Furthermore, 86-95% of multi-exon human genes is estimated to undergo alternative splicing (66). Genes tend to express many transcripts simultaneously, 70% of which encode important functional or structural changes for the protein (66). RNA-Seq data encompasses expression at both gene and transcript levels: the gene-level expression amounts to the combined expression of all transcripts associated with a particular gene. It has been previously demonstrated that the gene-level expression is an excellent indicator of the tissue of origin as well as certain cancer types (245-249). However, transcript-level expression has been shown to provide a more precise measurement of gene product dosage, resulting in the superior performance in predicting the cancer patient prognosis or survival time, and providing further insights into the functional transformations driving cancer (243, 250-252). Differential AS depends on many factors, including the epigenetic state, genome sequence, RNA sequence specificity, activators and inhibitors from both, proteins and RNAs, as well as post-translational modification (37, 253-255). These diverse mechanisms control AS to obtain developmental, cell-type, and tissue-specific expression. Furthermore, the patterns driven by AS and specific to cancer and other diseases have been recently identified (256, 257).

Machine learning tools developed over the last several decades have significantly advanced the analysis of the vast amount of next generation sequencing and microarray expression data by discovering the biologically relevant patterns (258-260). Previous studies have utilized unsupervised and supervised machine learning techniques on the microarray gene expression data with variable success rates (261, 262). Along with the individual approaches (263), large-scale comparative studies have been carried out (264, 265). Some studies evaluated both basic and advanced clustering techniques, such as hierarchical clustering, k-means, CLICK, dynamical clustering, and self-organizing maps, to identify the groups of genes that share similar functions or genes that are expressed during the same time point of a mitotic cell cycle (249, 266, 267). Other studies compared the ability to perform disease/healthy sample classification tasks by state-of-the-art supervised methods, such as Support Vector Machines (SVM), Artificial Neural Nets (ANN), Bayesian Networks, Decision Trees, and Random Forest classifiers (265).

When it comes to the biological classification, the RNA-Seq data present an attractive alternative to microarrays, since it is possible to quantify all RNA present in the sample without the need of the *a priori* knowledge. With RNA-Seq rapidly replacing microarrays, it is necessary to assess the potential of the supervised machine learning methodology applied to the RNA-Seq data across multiple datasets and biological questions (268). Recently, there have been limited studies that have assessed RNA-Seq data with supervised and unsupervised machine learning techniques (269). However, these studies utilized RNA-Seq data by leveraging only gene-level expression data rather than more detailed transcript-level, or transcript-level, data available for the alternative splicing transcripts (37). Most recently, a study analyzed the utility of RNA-Seq transcript-level data for the disease/non-disease phenotype classification of the samples, showing the advantage of the transcript expression data for the disease phenotype prediction task (270). However, the question of whether or not the utility of transcript-level expression presents a general trend across all main biological and biomedical classification tasks remains open.

This work aims to systematically assess how well state-of-the-art supervised machine learning methods perform in various biological classification tasks when utilizing either gene-level or transcript-level expression data obtained from the RNA-Seq experiments. The assessment is done from three different perspectives: (i) by analyzing RNA-Seq data from two organisms (rat and human), (ii) by using the increasingly difficult datasets, and (iii) by considering different technical scenarios. The datasets were analyzed using six supervised machine learning techniques, three normalization methods, and two RNA-Seq analysis pipelines. Altogether, the performance on 61 major classification problems include 2,196 individual classification tasks were compared. We define a *major classification problem* as a combination of the biological class and the dataset used. We then define the *individual classification task* as a combination of all machine learning methods, normalization techniques, as well as the major classification problem. The use of multiple datasets allows us to determine if the success of a classification task is due to the discovery of distinct biological patterns by a machine learning algorithm, or if it is due to biologically unrelated patterns such as caused by differences in library preparation and/or the lab source. Finally, we assess whether using the information on alternatively spliced transcripts presented in the form of transcript expression data can provide the higher classification accuracy, compared to the gene expression data.

2.3.3 Results

The goal of this work is to examine the capabilities of supervised machine learning methods in performing biological classification based on RNA-Seq data. Specifically, we analyzed whether the performance is influenced by (1) the power of the machine learning classifier, and/or (2) more detailed information extracted from the RNA-Seq data. In the first case, we assessed several supervised classifiers, ranging from the very basic methods to the state-of-the-art supervised classifiers, across three different normalization techniques. In the second case, we compared the same classifiers using either gene-level or transcript-level expression data. Together, the study setup utilized three RNA-Seq datasets, six supervised machine learning techniques, and three normalization protocols (Figure 2.3.3.1). Furthermore, each of the 61 classification problems was set to use the numerical features generated either from the gene-level or transcript-level expression data. To the best of our knowledge, this is the largest comparative analysis of biological classification tasks based on RNA-Seq data, performed to date.

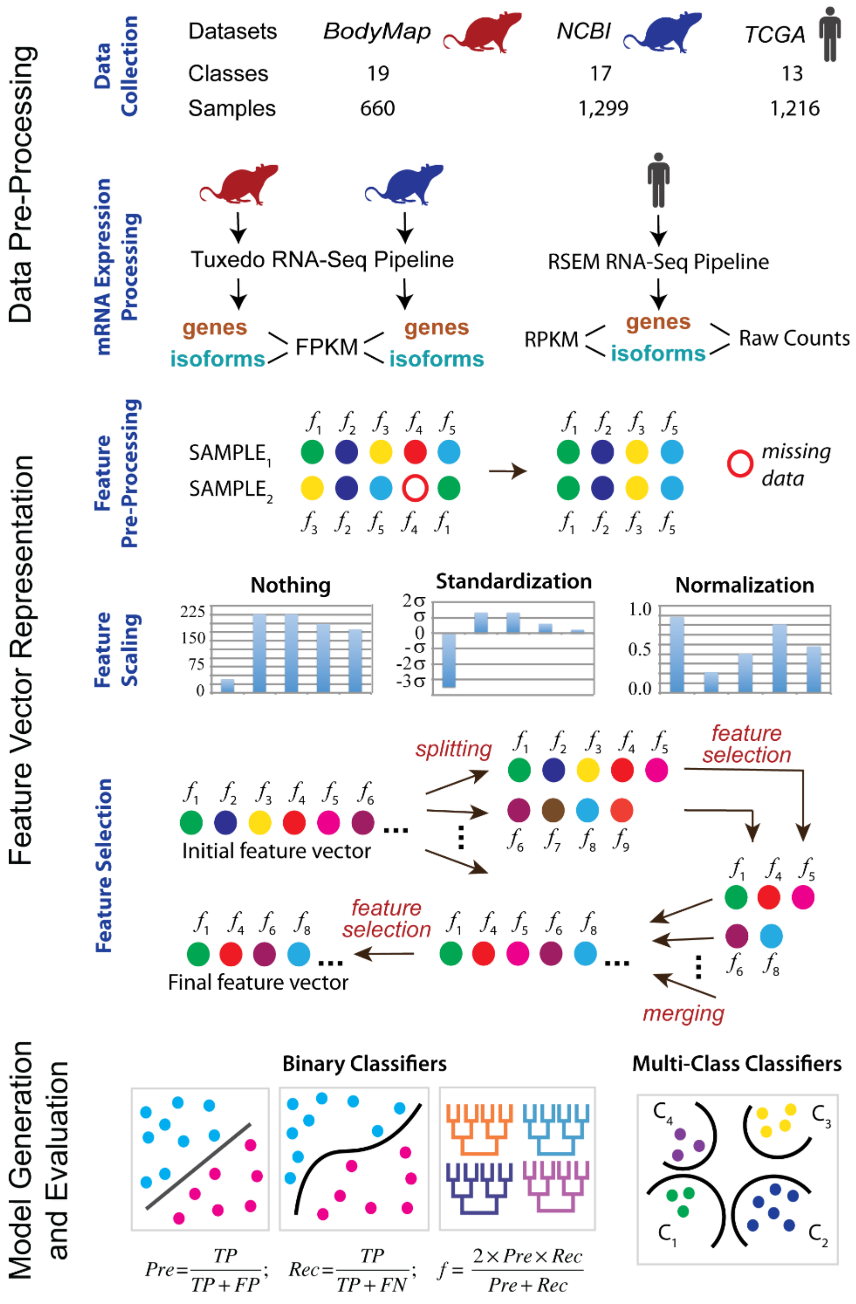


Figure 2.3.3.1. Overall computational pipeline. The samples from each of the three datasets are collected. The classification tasks are then defined. The expression data are processed for each sample at the gene and isoforms levels using two RNA processing pipelines and three different count measures. Next, feature pre-processing, scaling, and selection are done for each classification task. Finally, the binary as well as multiclass supervised classifiers are trained and tested.

Classification Tasks Analyzed

Two categories of classification tasks were considered: normal phenotype and disease phenotype. In the first category, we determined whether it was possible to distinguish between age groups, sex, or tissue types in normal rats based on transcriptome analysis. The second category focused on classification tasks associated with breast cancer, with the main goal to differentiate between the pathological tumor stages. Both categories were analyzed using RNA-Seq data at the gene and transcript levels. Two types of classification were considered for each category of tasks: binary classification and multiclass, or multinomial, classification. These classification types center around two conceptually different classification problems. The binary classification distinguishes a sample as either belonging to the class or not. The multiclass classification distinguishes which class a specific sample belongs to. For example, for a binary tissue classification task, a sample can be classified as extracted from the brain tissue or not. In the context of a multiclass classification, the same sample is classified as extracted from exactly one of several tissue types.

Dataset Statistics

Three datasets were used to carry out the classification tasks: two datasets for the normal phenotype classification tasks and one dataset for the disease phenotype classification tasks (Figure 2.3.3.1). The first dataset was obtained from the Rat Body Map and is referred to as RBM dataset. It consisted of 660 normal rat samples whose transcriptomes were sequenced from the same rat strain and served as a reference dataset for the community **(271)**. The transcriptomes were obtained at 40 M reads per sample on average. The data were evenly split between the male and female rats, four age groups, and eleven tissue types (Appendix A1.1). The four age groups included 2, 6, 21, and 101 weeks. The eleven tissue types included adrenal gland, brain, heart, thymus, lung, liver, kidney, uterus, testis, muscle, and spleen. All samples used the same library preparation protocol, sequencer, and were prepared by the same laboratory. As a result, the dataset was expected to have the least impact from the data inconsistency that arises from the non-biological sources, such as utilizing different sequencing protocols, instruments, and other parameters.

The second dataset, also used in the normal phenotypes classification tasks and referred to as NCBI dataset, included over 1,100 samples (Appendix A1.2) with the sequencing depth ranging between 6 and 116 M reads. This dataset was prepared by analyzing the collection of rat transcriptomes that were sequenced on Illumina Hi-Seq 2000 platform and were publicly available from the NCBI SRA database **(272)**. The dataset was obtained from the sequencing experiments of 29 research projects (Appendix A.1.1). It contained highly variable transcriptomes due to the differences in library preparation, project goals, and rat strains. The classification tasks for the NCBI dataset were the same as for the RBM dataset, but with one modification. The age classification was modified from the four age groups into either embryo or adult age groups and is described later in this section.

The last dataset included raw RNA-Seq data from 1,216 human breast cancer patients from the Cancer Genome Atlas (referred to as TCGA dataset) and was used in the disease phenotype classification tasks **(273)**. At the preprocessing stage, two RNA-Seq data normalization techniques were implemented and compared. Classification was performed to distinguish between the pathological cancer stages, as defined by the American Joint Committee on Cancer (AJCC) **(274)**. The AJCC breast cancer staging is based on size of tumors present within the breast, presence or absence of detection of metastases that are not within the breast, and the presence, size, and type of metastases within the lymph nodes. The patients were distributed with high variability especially when considering sub-cancer stages (Appendix 1.1.3).

Feature Selection and Analysis

The numerical features for this study represented either gene or transcript expression levels. As a result, the number of features ranged from 10,711 to 73,592, depending on the dataset and representation (Suppl. Table S2). Utilizing all features for a classification task greatly increases the computational complexity. Moreover, not all expressed genes or transcripts may be important for a given classification task; using the uninformative features during the training process could potentially decrease the accuracy of the classifier. To reduce the dimensionality of the feature space, a feature selection method **(275)** was applied in a classification-specific and dataset-specific manner, resulting in a significant reduction of features ranging from 107 to 735 folds (Figure 2.1.3.2A, Appendix A1.4-7, Appendix A2.3).

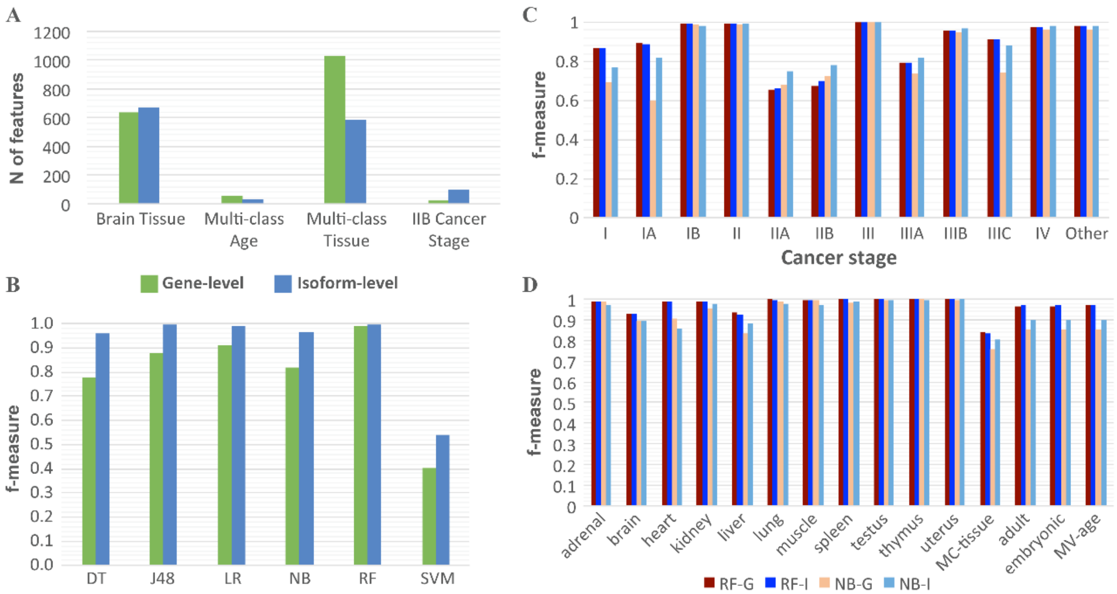


Figure 2.3.3.2. Overview of feature selection and the performance of classifiers using gene and isoform level expression data. A. Comparison of the number of features between gene and isoform after feature selection. Each classification task has the same number of features selected for each classifier at the gen-level and isoform-level. The four selected classes represent the four types of patterns seen between gene-level (green) and isoform-level classifiers. The Brain Tissue class is the most common pattern of feature selection. In general, more features are selected for isoform-level classifiers versus gene-level. B. Example of the variability of gene and isoform performance determined by f-measure across the six methods (DT = Decision Table, J48 = J48 Decision Tree, LR = Linear Regression, NB = Naïve Bayes, RF = Random Forest, SVM = Support Vector Machine). This example is from the RBM dataset for the Multi Age class without normalization. While there is a high degree of variability in performance, isoform-level classifiers consistently perform either comparably or better than gene-level classifiers. C. and D. Summary of the performance variability across classes for gene and isoform f-measure for the most frequent top and bottom performance methods (RF-G = Random Forest Gene, RF-I = Random Forest Isoform, NB-G = Naïve Bayes Gene, NB-I = Naïve Bayes Isoform). The data used in C. is TCGA dataset and in D. is NCBI dataset. MC = stands for multiclass.

Regardless of the classification task or dataset, the normalization of the RNA-Seq data did not make a significant difference on the choice of the selected features: variation in the numbers of selected features was less than 1% (Appendix A1.8-11). An interesting observation, consistent across different tissue classification tasks, was that the number of features selected for the multiclass classification tasks was significantly greater than for a binary classification task. This observation should not be surprising, because the binary classification task is generally simpler than the multiclass classification task (Suppl. Table S3). However, in our case even if all features of the binary classification tasks related to a single multiclass classification task were combined, it would still not account for all features selected by the feature selection method for the multiclass task.

In many cases, the overall number of features selected for a binary classification task was the same or nearly the same, irrespective of whether the features were gene- or transcript-based (Appendix 1.4-7). Does it mean that the features from the gene- or transcript-based approaches correspond to the same genes? Not always: the transcripts used for the selected features in a transcript-based, or transcript-based, classifier did not always originate from the genes that were selected for the corresponding gene-based classifier. Indeed, because 70% of transcripts were expected to encode different functional gene products **(66)**, we expected cases where the gene expression features were not as specific as the corresponding transcript features. In general, there was a large portion of 73,592 transcripts from 20,524 genes that corresponded to the same gene set (70-100%). However, there were several classification tasks, including multiclass tissue classification using NCBI dataset, where there was a lower percentage of such overlap (30%). Furthermore, there were several classification tasks, including multiclass age classification using RBM, multiclass tissue classification using GEO, and stage IIB classification using TCGA dataset, where the numbers of features that used either gene or transcript level of expression were significantly different, which was usually the case when a multiclass classification task was considered (Figure 2.1.3.2A, Appendix A1.4-S7, Appendix A2.2-3). Another interesting observation was obtained when comparing the RBM and NCBI Rat datasets: the number of selected features was much smaller for the RBM dataset rather than for the NCBI Rat dataset (on average 231 versus 588), thus indicating the need for additional features to compensate for the increased data variability found within the NCBI Rat dataset.

Overall Performance of Classifiers trained on Gene-based vs. Transcript-based data

Next, we hypothesized that because of the observed specificity of alternative splicing across tissues, ages, sexes, and between disease/normal phenotypes, training classifiers with the RNA-Seq data at the transcript, or transcript, level for the biological classification tasks could increase the classification accuracy (275, 276). Consistent with this hypothesis, the supervised learning classifiers that leverage the transcript-based data performed comparably or better than the classifiers trained on the gene data for all classification tasks (Figure 2.1.3.2B, Figure 2.1.3.3). This observation also held true irrespective of the datasets used, normalization protocols, classification tasks, or supervised classifiers. Furthermore, the difference between the gene- and transcript-based classifiers were consistently less than the standard deviation across all 10-folds, supporting this hypothesis (Appendix A1.12-13). The most frequently top performing methods were the random forest and logistic regression classifiers, whereas the worst performing method was typically naïve Bayes classifier (Figure 2.1.3.2B-D). However, the former approaches were not the most accurate ones for every single classification task, since in some cases naïve Bayes classifier was capable of outperforming all other methods tested (Stage IIA & IIB, Figure 2.1.3.2C). In general, the random forest classifier applied to the data without any normalization achieved 83-100% accuracy (Figure 2.1.3.2B-D).

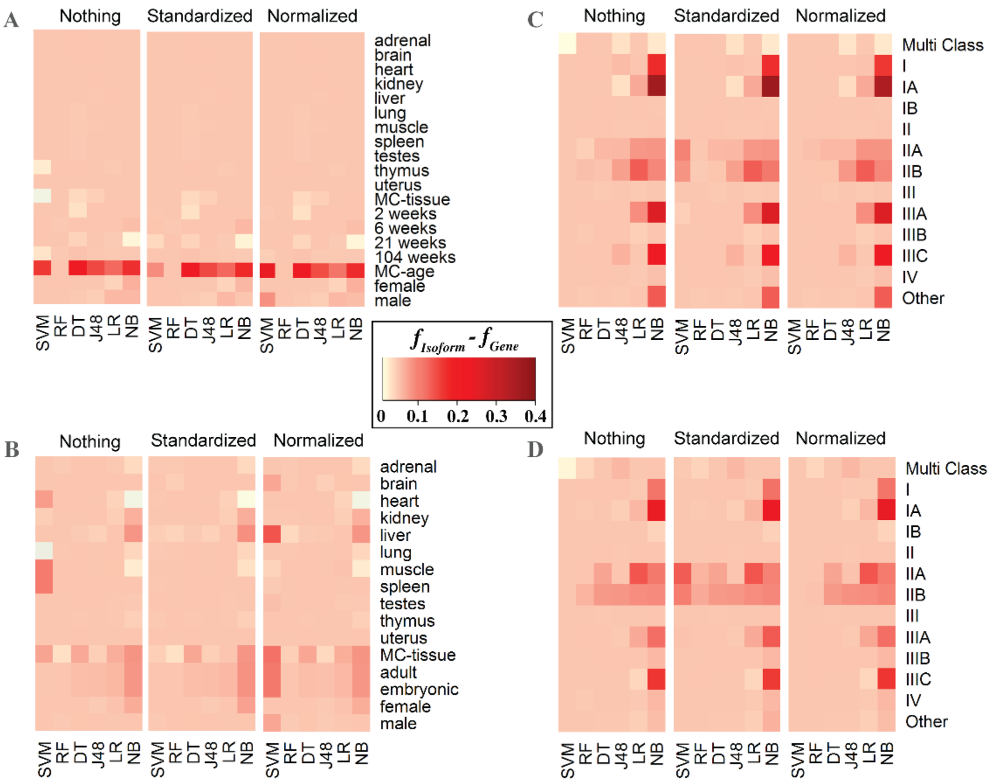


Figure 2.3.3.3. Heat map representation of the difference between Isoform and Gene *f*-measure across machine learning methods, classes, datasets, and normalization techniques. For the majority of classification tasks using isoform-level rather than gene-level expression data resulted in small to substantial increase of the performance accuracy, represented by *f*-measure values here. The bottom x-axis represents the machine learning techniques (DT = Decision Table, J48 = J48 Decision Tree, LR = Linear Regression, NB = Naïve Bayes, RF = Random Forest, SVM = Support Vector Machine). The y-axis represents the classes considered. MC stands for multiclass. The top x-axis represents normalization techniques including Nothing (no normalization), Standardized, and Normalized. Datasets for each panel are A. RBM, B. NCBI, C. TCGA – $\log_2$ normalized counts, and D. TCGA – raw counts.

It was also observed that for 63% of classification tasks, the gene- and transcript-based methods performed with similar accuracy (within 0.2 difference in f-measure value). For 37% of the classification tasks, the transcript-based methods performed better than the gene-based (more than 0.2 gain in f-measure value). The difference between the transcript-based and gene-based classification accuracies was particularly profound when comparing the classification results of naïve Bayes, which was one of the less accurate methods analyzed, while being among the fastest classifiers. However, we did not observe such a drastic difference, and sometimes no difference at all, when considering one of the most accurate classifiers, random forest, across all classification tasks. For instance, when comparing gene- and transcript-based classifiers for stage IA cancer using the raw count expression values and not performing any normalization protocols, the accuracy and f-measure values for naïve Bayes classifier ranged between 49.5%-76.4% and between 0.60-0.82, respectively, while for random forest the ranges were nearly identical (Figure 2.3.3.2C-D).

Another potential source of variability in the classifier performance was the difference in the protocols used by different studies. To determine whether the difficulty of classification task increased when using datasets from multiple laboratories rather than from a single one, the classification accuracies between the two rat datasets were compared for each binary or multiclass classification task. Not surprisingly, we found that there was greater difference in the performance accuracies when relying on the data from one laboratory compared to the data from multiple laboratories (Figure 2.1.3.4A-B). With exception of a single worst performing classifier, SVM, the classifiers performed better on the RBM dataset, which came from a single study, then on the NCBI dataset, which was obtained by merging multiple independent studies. Moreover, this difference held for both the gene and transcript-based models. Next, we evaluated if the prediction accuracy depended on the transcript counting approach. To do so, TCGA expression values were calculated based on (i) raw counts and (ii) $\log_2$ normalized counts with respect to the gene length and sequencing depth. The results showed that there was a strong preference, in terms of accuracy, in raw counts for the gene-based classifiers, but to a lesser extent for the transcript-based models. However, the opposite was observed where transcript-based models were more accurate when using $\log_2$ normalized counts (Figure 2.3.3.4C-D). There was less variability (less than 0.3 in the maximum difference of f-measure values across all methods for each classification task) when considering transcript-based versus gene-based models.

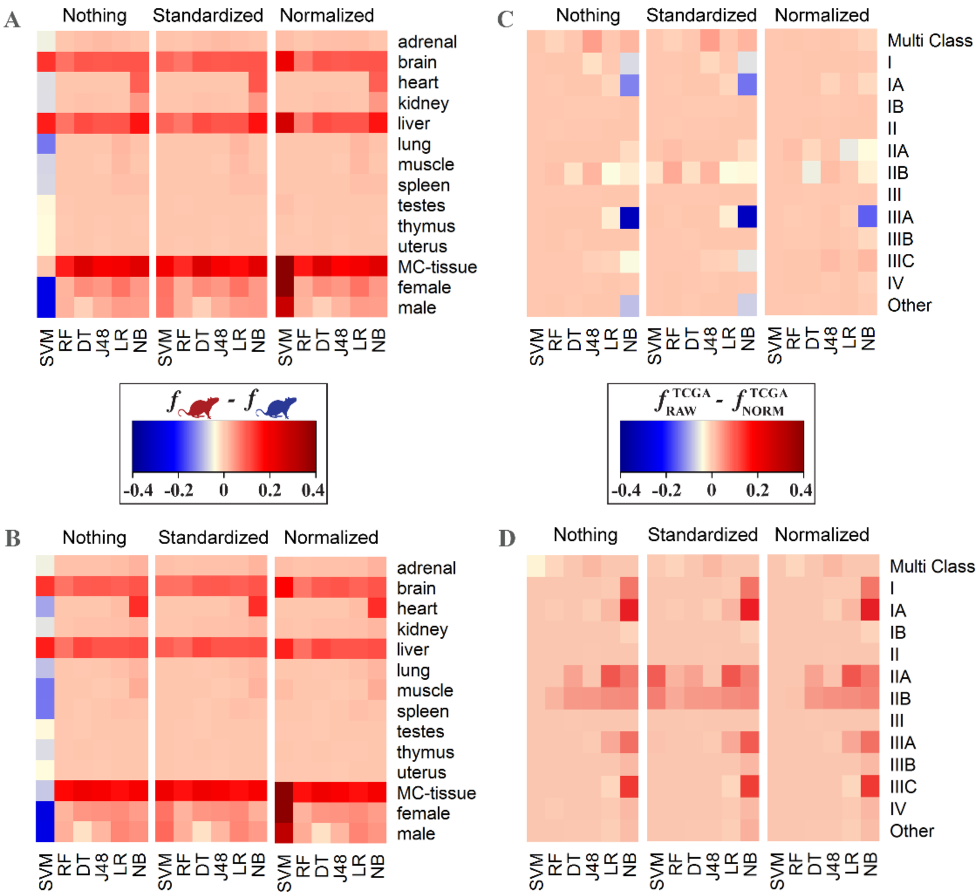


Figure 2.3.3.4. Heat map representation showing the influence of different factors on the accuracy performance. Panels A. and B. represent the difference in performance accuracies, calculated with f-measure, between RBM (single-lab) and NCBI (multi-lab) datasets for gene-based A. and isoform-based B. classifications, respectively. Panels C. and D. represent the difference in f-measure between the classifiers trained on the TCGA expression values, quantified as either raw counts or $\log_2$ normalized counts with respect to gene length and sequencing depth. Shown are f-measure differences for gene-based (C.) and isoform-based (D.) classifications, respectively. The bottom x-axis represents the machine learning techniques (DT = Decision Table, J48 = J48 Decision Tree, LR = Linear Regression, NB = Naïve Bayes, RF = Random Forest, SVM = Support Vector Machine). The y-axis represents the classes considered. MC stands for multiclass. The top x-axis represents normalization techniques including Nothing (no normalization), Standardized, and Normalized.

Finally, we considered different normalization techniques across the gene- and transcript-based classifiers. The general trend observed was little to no difference in performance accuracies using either different normalization protocols or no normalization at all. The only exception was the performance of the SVM classifier employed by both, the transcript-based and gene-based, approaches: differences in the accuracy values between the various normalization techniques for some classification tasks were as high as 40.3 and 30.7%, respectively (Figure 2.3.3.5).

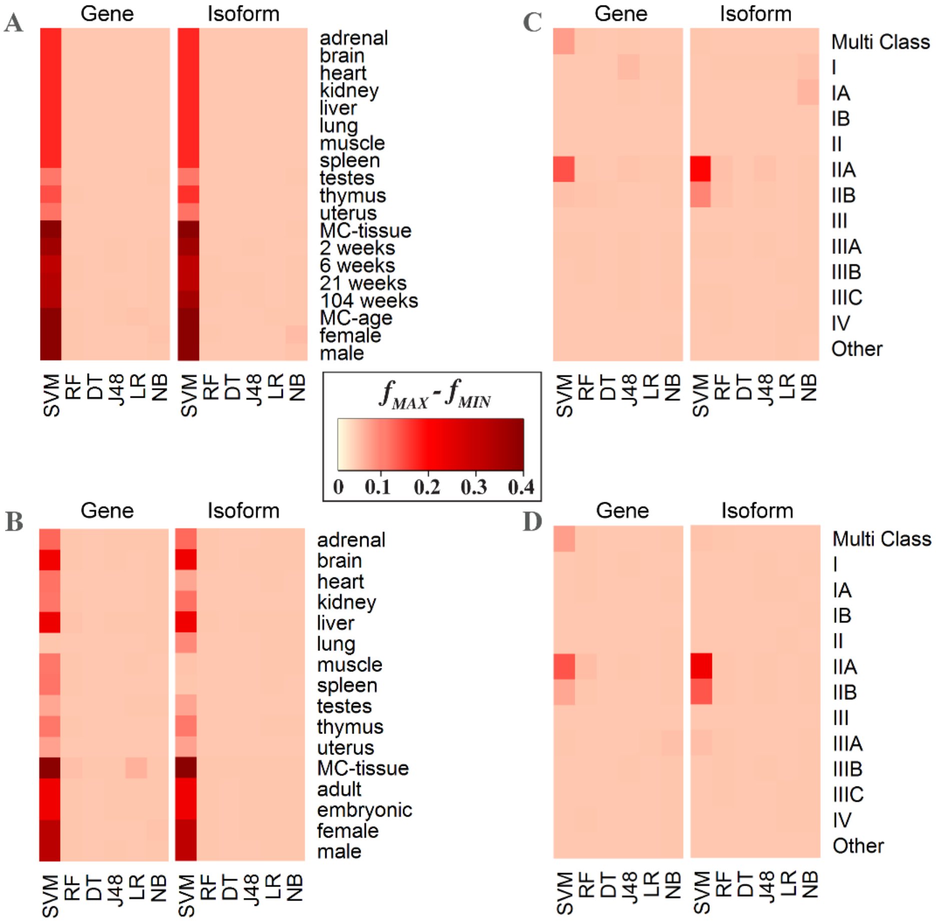


Figure 2.3.3.5. Heat map representation of the difference between maximum *f*-measure and minimum *f*-measure across normalization techniques. To demonstrate the variability attributed to the machine learning normalization technique, the intensity of the color represents the difference between the maximum and minimum *f*-measures achieved for a specific classification task and specific classifier across all three normalization protocols. The upper x-axis reflects if the difference is from gene or isoform expression values. SVM is the only method that has

significant changes due to normalization. The lower x-axis represents machine learning techniques (DT = Decision Table, J48 = J48 Decision Tree, LR = Linear Regression, NB = Naïve Bayes, RF = Random Forest, SVM = Support Vector Machine). The y-axis represents the classes considered. MC stands for multiclass. Datasets for each panel are A. RBM, B. NCBI, C. TCGA – $\log_2$ normalized counts, and D. TCGA – raw counts.

Normal Phenotype Classification Tasks: Age, Sex, and Tissue Classification of Rat Samples

The Rat Body Map (RBM) represents a dataset with the least amount of noise due to non-biological variation becomes it comes from a single laboratory, which uses the same sample and library preparation protocols and a fixed sequencing depth (Figure 2.3.3.3A). From this dataset we identified eleven tissue types, four age groups, and both sexes. We then defined 17 "one-against-all" binary classification problems. Additionally, we merged the tissue and age groups and applied a multiclass classifier.

For the tissue classification, including multiclass tissue classification, the models achieved 100% accuracy and 1.0 f-measure based on the assessment protocol and irrespective of the machine learning method. However, when considering normalization technique, SVM had the accuracy ranged between 75.3% to 99.8% and 0.39 to 1.00 f-measure. The age group classification represented a more challenging task, with the classification accuracy ranging between 40.2% to 100% and f-measure ranging from 0.40 to 1.00. For the 2-week and 104-week age groups, the classifiers again achieved nearly 100% accuracy and 1.0 f-measure across all machine learning techniques. The 6-week and 21-week age groups were predicted with over 97% accuracy using random forest, j48, and logistic regression classifiers, while naïve Bayes could only achieve 81.1% and SVM with 40.2%. Similar pattern was observed in sex classification, where logistic regression and random forest achieved more than 97.3% in accuracy, but naïve Bayes could reach only 86.1%.

The NCBI dataset was expected to result in a greater variation of the feature values, compared with the RBM dataset, since it included the data from multiple research laboratories that sequenced different rat samples and even strains using different library preparation protocols (Figure 2.3.3.3B). The same types of classification tasks were considered, including tissue, age, and sex. Since this dataset represents all publicly available data in rat obtained using the same sequencer model, it included more tissue types than the RBM dataset. For consistent comparison, only those tissue types that were previously included in the RBM dataset were chosen for the NCBI dataset for the binary classification. However, for the multiclass tissue classification problem, the labels were determined based on the entire range of organs and tissues that the samples originated from, thus including more tissue types than in the RBM dataset. In contrast, the age group classification for the NCBI dataset was more limited than that one for the RBM dataset, since some samples in the former did not include the detailed age information. Therefore, the age types for the NCBI datasets were reduced to either adult or embryonic types.

The RNA-Seq data normalization did not have an effect on the classification results for the NCBI dataset: the performance difference when using the normalized and unnormalized datasets was only observed for the SVM classifier, the method that performed the worst out of the six supervised learning methods. The binary tissue-based classification performed well overall, reaching over 99.7% in accuracy and 0.99 in f-measure for the top-performing random forest classifier. Interestingly, the worst performing classifier, SVM, achieved the accuracy of only 21.2% and 0.07 f-measure for the gene-based tissue multi class. The analysis of method performances for multiclass classification tasks revealed that classification of several tissue types was particularly challenging for some of the less accurate methods. The binary tissue type classification tasks reporting the lowest accuracies included brain and liver tissue classification, with 79.1%-94.3% in accuracies and 0.70-0.94 in f-measure values, depending on the supervised learning method used. For the harder problem of multiclass tissue classification, the performance of the classifiers was highly variable, with the accuracy ranging from 0.7% to 84.3% and f-measure from 0.07 to 0.84, and with the observation that the random forest classifier was, again, the best performing method. Differentiating between embryonic and adult samples as well as between the sexes were easier tasks compared to the tissue origin. The age classification accuracy ranged from 83.2% to 98.3% and f-measure from 0.75 to 0.97 across all six supervised learning. The sex classification task had classification accuracy ranging between 71.1% and 97.3% and f-measure between 0.59 and 0.97. Interestingly, the consistently poor performance of the SVM classifier was not dependent on the normalization technique.

Disease Phenotype Classification Tasks: Breast Cancer vs, Healthy and Stage Classification of Human Samples

Based on the promising results for the normal phenotype classification tasks, we further increased the difficulty of classification task by predicting different pathological stages of breast cancer using gene-based and transcript-based data. To evaluate if this classification task could benefit from additional information, we assessed the method performances based on the RNA-Seq data with $\log_2$ normalization in addition to the three types of normalization used in the two previous classification problem. The classification performance was heavily dependent on the supervised learning method with accuracies ranging from 20.2% to 99.8% and f-measure ranging from 0.21 to 1.00, and with naïve Bayes and SVM classifier being the worst performing classifiers (Figure 2.3.3.3C-D). Furthermore, when considering all classes and $\log_2$ normalization, the accuracies decreased by as much as 60%, and the only method that benefited from the normalization was the poorly performing SVM classifier.

For each stage of breast cancer, we were able to achieve at least 78.3% in accuracy and 0.77 in f-measure. However, there is a significant variability within all parameters tested (Figure 2.3.3.2C, Figure 2.3.3.3C-D). Similar to the analysis for the RBM and NCBI datasets, random forest had the highest performance across all stages of breast cancer based on 71.3% to 99.8% accuracy and 0.64 to 1.00 f-measure. The most difficult stages to classify were stages IIA and IIB, with the average difference in accuracy between 21.3% accuracy and 0.29 f-measure. Unlike the RBM and NCBI datasets, there were classes, such as Stage IIB, where naïve Bayes and SVM outperformed random forest by 5% in accuracy and 0.11 in f-measure. The easiest stages to classify were stages II and III with 99.8% accuracy and 1.00 f-measure.

In contrast to the RBM and NCBI datasets, the worst performing models for each class were highly variable, depending on the parameters chosen. For example, for stage IIA, the logistic regression classifier was the best performing model at 78.2% accuracy and 0.77 f-measure. However, the worst performing model was J48 at 60.1% accuracy and 0.60 f-measure. Similarly, for stage I the worst performing classifier was naïve Bayes with 53.7% accuracy and 0.63 f-measure, while the best performing classifier was random forest with 91.4% accuracy and 0.84 f-measure. On the other hand, for stage III binary classification the performance was 99.5-99.8% accuracy and 0.996-0.998 f-measure across all classifiers and parameter sets. These results demonstrated that no single method and parameter set was able to always outperform all others.

2.3.4 Discussion

This work achieves two aims. The first aim is to broadly assess how well the supervised machine learning methods perform in various biological classifications by utilizing exclusively the RNA-Seq data. This aim is supported by our rationale that the key biological patterns should be recoverable from the transcriptomics data. Our second aim is to investigate whether relying on the transcript-level expression, which provides details on the alternatively spliced transcripts, can increase the accuracy of biological classification compared to the gene-level expression. Since the data patterns detected by the machine learning techniques during their training stage are highly dependent on the type of biological classification problem, we wanted for our assessment to cover multiple aspects. Specifically, we evaluated the performance of six widely used supervised classifiers across different RNA-Seq datasets, organisms, and normalization protocols, totaling in 61 classification problems and 2,196 individual classification tasks. The different RNA-Seq datasets were selected based on the increasing difficulty of classification tasks due to the background noise. The RBM dataset represented the "easiest" dataset as the level of background oise was expected to be low due to using a single data source and well-defined biological classification problems: tissue-, age-, and gender-based. The assumption of a single data source implies a well-defined animal model, which the genetically identical specimina, and the same RNA-Seq library protocol. The NCBI dataset increases the background noise by including multiple RNA-Seq protocols and different genetic backgrounds, but keeping the classification the same, to allow for comparison with the RBM dataset. The TCGA dataset further increases the background noise due to increasing genetic and environmental variability by switching from a model organism to human, from the normal to disease-specific phenotype, and by relying on a potentially biased definition of the biological classes (breast cancer pathological stages are not defined from the molecular perspective, but by a pathologist). Each task separately utilized the gene-level and transcript-level expression datasets. The main purpose behind our study was to demonstrate the importance of enriching RNA-Seq data with the differentially expressed transcripts for the biological classification tasks, suggesting that limiting the RNA-Seq analysis to the differentially expressed genes would, in turn, limit the capabilities of machine learning algorithms. As a result, several important conclusions were made.

First, we found that the accuracy of machine learning classifiers depended on how much data variation associated with the type of sequencers, library preparation, or sample preparation was introduced. Our rat datasets were specifically selected to compare the differences in data variation and in classification accuracies. The first dataset (RBM) was chosen because it included samples representing multiple age groups, tissue types, as well as sex (271), while these data were generated by only one research group and using the same sequencer. Thus, possible variation due to the type of sequencers or preparation protocols was expected to be minimal. Furthermore, we downloaded and processed the raw RNA-Seq reads using our in-house protocol and thus excluding possible variation due to different RNA-Seq analysis techniques. Our second dataset (NCBI) incorporated all publically available RNA-Seq data for rat using the same sequencer model, thus minimizing possible sequencer-based bias, a well-documented source of variation (277). The NCBI dataset included 29 studies from multiple laboratories and represented the same classes as in the RBM except for the age groups. As expected, higher variation negatively affected the accuracy across predominantly all machine learning methods, normalization protocols and classification tasks. On the other hand, even for the NCBI dataset, the accuracies for all top-performing binary classifiers were never below 90% either for gene-level or for transcript-level expression data, suggesting minimal influence of the batch effect on the supervised classifiers.

Second, our study suggested that the standard data normalization techniques were not needed for RNA-Seq data except when using the poor-performing SVM classifiers. Random forest and logistic regression classifiers performed consistently well with each of the normalization technique but also without them, regardless of the classification task. However, there are several normalization techniques specific to RNA-Seq data, including RPKM (reads per kilobase per million reads), FPKM (fragments per kilobase per million reads), and TPM (transcripts per kilobase per million reads) (232). Assessing whether these normalization techniques have an effect on classification accuracy should be considered for future studies.

Third, we found that the overall performance of the most accurate machine learning classifiers was very strong, with a few exceptions. In fact, for several classification tasks including all tissue classes, 2 week, and 104 week from RBM dataset, stage I from TCGA dataset, and the top-performing classifiers achieved a perfect 1.0 f-score, while for the majority of other tasks, the accuracy and f-measure were no less than 0.9 and often achieved by more than one classifier. From the biological perspective, it was surprising to see how well the classifiers performed on the normal phenotype datasets, in spite of significant variations in the sample and library preparation by different labs as well as the difference in rat strains. Intuitively, the expression values should have high variability due to these differences. The few exceptions in excellent performance were the multiclass age group classification for the normal phenotype datasets and classifications of clinical stages I, IIA, IIB and IIIA for the disease phenotype dataset, with stages IIA and IIB performing significantly worse. The clinical definition of IIA and IIB are based on the size of the tumor as well as evidence of cancer movement, and the reduced performance on each of these stages suggests that while there is a phenotype difference there may not be a strong molecular expression difference, which would cause a higher error rate by a classifier. The results also suggested that, from the diagnostic perspective, a more accurate AJCC classification methodology to distinguish those two phenotypes might be required to improve the stage prediction accuracy. The most consistent in the overall performance across all tasks were the random forest classifiers, which had been previously shown to perform exceptionally well for a number of bioinformatics tasks (278) and can be suggested as a reliable first choice for a biological classification task. Overall, our findings provided strong evidence that the supervised learning approach is readily available for the majority of the biological classification tasks.

Finally, we found that the classifiers that leveraged the transcript-level expression never performed worse and often outperformed the classifiers that used the gene-level expression data. This observation was consistent across datasets, normalization techniques, RNA-Seq pipelines, and classification tasks. For the normal phenotype tasks, the most profound difference was when considering the most challenging classification task—the multiclass classification of age groups. For the disease phenotype tasks, the most significant difference in performances of the classifiers that used gene-based and transcript-based expression data was again for the most challenging classes, the clinical stages IIA and II B of breast cancer. The better performance for the classifiers on the transcript-level data seems to be the expected result because the methods are trained on the enriched data, from the biological point of view. However, we note that the transcript-level data provides a significantly higher number of initial features, which could result in adding more noise to or potential overfitting of a classifier. Hence the importance of the feature selection and thorough model evaluation, which in this work suggests that the transcript-level information is a better choice when developing a biological classifier. Given that the transcript extraction methods continue to improve (279, 280), we expect further improvement in the accuracy of transcript-level based classifiers.

In summary, this study demonstrates that a supervised learning method leveraging transcript-level RNA-Seq data is a reliable approach for many biological classification tasks. We conclude that an appropriate general purpose pipeline for building a RNA-Seq based classifier should use 1) transcript-based expression data, 2) feature selection preprocessing, 3) Random Forest classification method, and 4) do not use normalization. The proposed pipeline is computationally fast and can be fully automated for the projects that involve massive volumes of sequencing data and/or high number of samples. However, it is important to note that 1) there are some cases where Random Forest can be outperformed and 2) the protocols and methods used for data gathering may have an effect on the classifier. With the rapid advancements of RNA sequencing technologies as well as with continuous improvement of the transcript prediction methods, the accuracy of the machine learning approaches will only increase. We also expect for these methods to tackle more challenging tasks such as cell type classification, disease phenotype classification of common and rare complex diseases, and clinical stage classification across all major cancer types. Finally, we expect for advanced machine learning approaches, such as semi-supervised learning (281), deep learning (282), and learning under privileged information (283) to step in.

2.3.5 Materials & Methods

The methodology used in this study compares three RNA-Seq datasets, six supervised machine learning methods, three normalization techniques, two RNA-Seq analysis pipelines, and 61 classification problems in order to assess if the features derived from the expression data at the alternative splicing level (*i.e*., transcript-based) can result in a higher classification accuracy than the features derived from the gene-based expression levels. Our approach attempts to systematically evaluate the classifiers that relied on these features from multiple perspectives, with a goal to provide a comprehensive analysis. We use the increasingly difficult biological classification tasks to assess the performances of classifiers in the presence of noise due to the difference in the biological sources, sequencers, and preparation protocols. The analysis is based on three RNA-Seq datasets, two from rat and one from human. The six supervised machine learning methods tested in this work include support vector machines (SVM), random forest (RF), decision table, J48 decision tree, logistic regression, and naïve Bayes. The three normalization protocols used include (1) pipeline-specific RNA-Seq count with no post-normalization, (2) pipeline-specific RNA-Seq count with normalization from 0 to 1, and (3) pipeline-specific RNA-Seq count protocol with standardization with respect to standard deviation. The two RNA-Seq analysis pipelines in this work, each employing different RNA-Seq count methods were the standard Tuxedo suite and RSEM. The 61 classification problems include binary and multiclass classifications of tissue types, age groups, sex, as well as clinical stages of breast cancer.

Data Sources

Three datasets are used to demonstrate the usability of the transcript-level expression data for the supervised classification. The first two datasets are from rat samples of normal phenotype; the raw RNA-Seq data for both datasets is processed using our in-house protocol. The last dataset consists of already processed RNA-Seq data from human breast cancer samples (284). The first, RBM, dataset is obtained from the Rat Body Map and includes 660 samples from 12 different rats from the F344 rat strain (271) covering 4 different age groups, 11 tissues, and both male and female rats. Publicly available raw mRNA RNA-Seq data from the Rat Body Map (http://pgx.fudan.edu.cn/ratbodymap/) is downloaded and processed for the gene and transcript levels of expression. The second dataset, NCBI, includes all publically available raw RNA-Seq data from rat samples that are sequenced using Illumina Hi-Seq 2000 and available from the NCBI GEO DataSets collection (http://www.ncbi.nlm.nih.gov/gds, Suppl**.** Table S3). In total 1,308 samples are used, which represents 29 different projects. In contrast to the processing of the data for the first dataset, these 29 projects used a variety of library preparation protocols and adapters to process their samples. The third, TCGA, dataset is obtained from The Cancer Genome Atlas data repository (284) and includes 1,216 breast cancer patients diagnosed with different pathological cancer stages (as defined by the American Joint Committee on Cancer, AJCC (274)). The class distributions for all datasets are shown in Appendix A1.9-11.

RNA-Seq Pipeline

RNA-Seq analysis encompasses three main stages: preprocessing, alignment, and quantification. There are a number of methods to perform each of these three basic steps, while the debate on the most appropriate methodology continues (232). In this work, we expect for the variation due to data processing to be minimal since the same processing pipeline is used for each dataset. Two different RNA-Seq pipelines are implemented and applied to each dataset for both gene and transcript levels of expression. These two pipelines leverage different algorithms and different metrics (285). For the RBM and NCBI dataset, all raw RNA-Seq data are downloaded from the SRA repository (https://www.ncbi.nlm.nih.gov/sra) using unique project IDs (Suppl**.**

Table S3). SRA file formats are then converted into fastq format. These files are used as input for the preprocessing stage. The preprocessing is done using Fastx Tools with the settings that removed reads shorter than 20 bp. All nucleotides with quality scores of less than 20 are converted into N's (286). The alignment is done against the rat genome version rn5 (287) using Tophat v2 and its default settings (288) . Quantification for both gene- and transcript-based expression levels is performed using Cufflinks v2 (289) and Ensembl transcript annotation v75 (290). The Cufflinks is set to use the transcript annotation for quantification with other settings being default. For the TCGA dataset, MapSlice (291) is used for alignment and RSEM (229) for quantification. The final output includes expression levels for each sample at both gene and transcript levels. We note that the gene-based expression values are the summation of all transcripts determined to be associated with the corresponding gene.

Supervised Learning Classifiers

The quantified expression values obtained from Cufflinks are then used to train and assess six supervised classifiers for each task. Two types of classification tasks are considered: one-against-all and multiclass. Our classification approach leverages feature-based supervised learning methods. Each post-processed RNA-Seq sample is represented as a feature vector, where each feature represents the transcript- or gene-level expression value for a specific gene or AS transcript corresponding to this gene. Expression samples may vary in length, thus to generate feature vectors of the same length, we compute the intersection of all samples in terms of the feature set that represents each sample. We next rank the importance of each feature and select subsets of the features that best describe their respective classes using the Best First (BF) feature selection method (275). The BF method is driven by the property that the subsets of important features are highly correlated with a specific class and are not correlated with each other. The method is described as a greedy hill climbing algorithm augmented with a backtracking step, where the importance of features is estimated through one-by-one feature removal. All machine learning methods are implemented using the Weka package version 3.7.13 (292).

Due to a large number of features for genes and even greater number of features for the transcripts (~20,000-73,000) using the base classification and even BF method was not computationally feasible, thus the modification to the original methodology is implemented allowing to reduce the processing time. The modifications includes introducing multiple splits of the features followed by two rounds of BS feature selection. Specifically, we split the data into 1,000 subsets and perform feature selection on each subset independently. After feature selection is performed on all splits, the selected features are merged, and another round of feature selection is performed. Our solution reduces the time needed to compute from several weeks to hours and still able to successfully select a reduced feature set that allows for accurate classification.

Machine Learning Technique Rationale

A broad selection of supervised learning approaches were implemented to test whether performance could be improved, depending on the method tested. The machine learning methods have different assumptions on how the data are structured; the methods also vary in their treatment of the class outliers and convergence w.r.t number of training examples.

The first two classifiers, naïve Bayes and logistic regression, are often regarded as the baseline methods due to their simplicity and robustness. *Naïve Bayes* classifier is a probabilistic method that has been used in many applications including bioinformatics and text mining (293-297). It is a simple model that leverages the Bayes rule and describes a class of Bayesian networks with assumed conditional independence between the numerical features. The use of this “naïve” assumption makes the method computationally efficient during both the training and classification stages. Furthermore, while the probability estimation by naïve Bayes is reported to be not very accurate (298), a threshold-based classification performance is typically very robust. In our implementation, the numeric estimator precision values are chosen based on analysis of the training data and is set to 0.1 The batchSize parameter that specifies the preferred number of instances to process during training if batch prediction is being performed is set to 100. *Logistic regression* is another type of a simple machine learning classifier that has been compared with naïve Bayes in terms of accuracy and performance (299). Different versions of logistic regression models are often used in bioinformatics applications (300-303). In this work, we implemented a boosting linear logistic regression method without regularization and with the optimal number of boosting iterations based on cross validation.

The next three classifiers, decision tables, J48, and random forest, are the decision tree based algorithms. A *decision table* is a rule-based classifier commonly used for the attribute visualization and less commonly for classification. The rules are represented in a tabular format using only an optimal subset of features that are included into the table during training. The decision table is a less popular approach for bioinformatics and genomics classification tasks, however it has showed a superior performance in some bioinformatics applications (304), and therefore is included into the pool of methods. The decision table model is implemented as a simple majority classifier using the Best-First method for searching. *J48* is an open source implementation of perhaps the most well-known decision tree algorithm, C4.5 (305), which is, in turn, is an extension of Iterative

Dichotomiser 3 (ID3) algorithm (306). C4.5 uses the information-theoretic principles to build decision trees from the training data. Specifically, it leverages the information gain and gain ratio for a heuristic splitting criterion with a greedy search that maximizes the criterion. Furthermore, the algorithm includes a tree pruning step to reduce the size of the tree and avoid the overfitting. In this work, the implementation of J48 was done with the default confidence threshold of 0.25 and minimum number of instances per leaf set to 2. *Random forest* is an ensemble learning approach, where many decision trees are generated during the training stage, with each tree based on a different subset of features and trained on a different part of the same training set (307). During the classification of unseen examples, the predictions of the individually trained trees are then agglomerated using the majority vote. This bootstrapping procedure is found to efficiently reduce the high variance that an individual decision tree is likely to suffer from. The random forest methods have been widely used in bioinformatics and genomics applications due to their versatility and high accuracy (307). In this work, due to a large but highly variable number of features the number of attributes, *K*, randomly selected for each tree is dependent on the classification task and is defined as $K = \lfloor \log_2 n + 1 \rfloor$, where *n* is the total number of features. The number of sampled trees per each classifier is set to 100.

The last method, Support Vector Machines (SVM) represents yet another family of the supervised classifiers, the kernel methods (308). It is among the most well-established and popular machine learning approaches in bioinformatics and genomics (261, 309-311). SVM classifiers range from a simple linear, or maximum margin, classifier where one needs to find a decision boundary separating two classes and represented as a hyperplane, in case of a multi-dimensional feature space, to a more complex classifier represented by a non-linear decision boundary through introducing a non-linear kernel function. For our SVM model training, Radial Basis Function (RBF) was used, a commonly used kernel. The two parameters, *Gamma* and *C*, were set to 0.01 and 1, respectively.

Training, Testing, and Assessment of classifiers

To evaluate each of the classifiers, a basic supervised learning assessment protocol is implemented. Specifically, the training/testing stages are assessed as a 10-fold stratified cross validation to eliminate the sampling bias. This protocol is implemented using Weka (292). The reported result of assessment is based on the average *f*-measure for the 10-folds for testing dataset. *f*-measure incorporates recall (*Rec*, also called sensitivity) and precision (*Pre*) into one reported metric:

$$Ac = \frac{TP + TN}{TP + FP + TN + \ FN}; Pr = \frac{TP}{TP\ x\ FP}; Re = \frac{TP}{TP + FN}; Sp = \frac{TN}{FP + TN};$$

where *TP* is the number of true positives (correctly classified as class members for a specified class), *TN* is the number of true negatives (correctly classified as not class members), *FP* is the number of false positives (incorrectly classified as class members), and *FN* is the number of false negatives (incorrectly classified as not class members). While each of the above four measures are commonly used to evaluate the overall performance of a method, we primarily focus on the most balanced metric, *f*-measure, due to a high number of classification tasks to be reported.

Availability

The supervised machine learning methods were implemented using the Weka platform (http://www.cs.waikato.ac.nz/ml/weka/). Data used are publically available from Rat Body Map (http://pgx.fudan.edu.cn/ratbodymap/), Geo Datasets (http://www.ncbi.nlm.nih.gov/gds), and the Cancer Genome Atlas (https://portal.gdc.cancer.gov/).

CHAPTER 3: Proteomics

3.1 Determining rewiring effects of alternatively spliced isoforms on protein-protein interactions using a computational approach (241)

3.1.1 Abstract

The critical role of alternative splicing (AS) in cell functioning has recently become apparent, whether in studying tissue- or cell-specific regulation, or understanding molecular mechanisms governing a complex disorder. Studying the rewiring, or edgetic, effects of alternatively spliced isoforms on protein interactome can provide system-wide insights into these questions. Unfortunately, high-throughput experiments for such studies are expensive and time-consuming, hence the need to develop an *in-silico* approach. Here, we formulated the problem of characterization the edgetic effects of AS on protein-protein interactions (PPIs) as a binary classification problem and introduced a first computational approach to solve it. We first developed a supervised feature-based classifier that benefited from the traditional features describing a PPI, the problem-specific features that characterized the difference between the reference and alternative isoforms, and a novel domain interaction potential that allowed pinpointing the domains employed during a specific PPI. We then expanded this approach by including a large set of unlabeled interactomics data and developing a semi-supervised learning method. Our method called AS-IN (Alternatively Splicing INteraction prediction) Tool was compared with the state-of-the-art PPI prediction tools and showed a superior performance, achieving 0.92 in precision and recall. We demonstrated the utility of AS-IN Tool by applying it to the transcriptomic data obtained from the brain and liver tissues of a healthy mouse and western diet fed mouse that developed type two diabetes. We showed that the edgetic effects of

differentially expressed transcripts associated with the disease condition are system-wide and unlikely to be detected by looking only at the gene-specific expression levels.

3.1.2 Introduction

Protein-protein interactions (PPIs) underlie many key mechanisms of cellular functioning (312). With thousands of PPIs simultaneously occurring in every cell of an organism, an average protein is expected to interact with two or more other proteins forming large molecular assemblies, transporting proteins, facilitating a chemical reaction, protecting the organism from pathogens, and carrying out other basic functions (313-315). Throughout the past two decades, there have been efforts in characterizing the experimentally confirmed PPIs by describing the structure of molecular complexes and interaction interfaces formed through the PPI (316, 317), determining a protein function that is controlled by the interaction (318), and understanding the evolutionary principles shared between the homologous interactions (319, 320). More recently, several studies have been published that focus on studying the interaction-rewiring, edgetic, effects of genetic variations cause by genetic diseases (321, 322). The edgetic effects on the whole protein interactome of other types of variation, such as copy-number variation, epigenetic variation, and transcriptional variation, or alternative splicing, are far less studied (84, 313).

Alternative pre-mRNA splicing due to either natural or disease-causing variation in transcriptome is a process by which the same gene can result in different gene products through selective inclusions and exclusions of the gene’s exons and introns (323). While many alternative splicing events naturally occur in different tissues, cells, and under different cellular conditions, a growing number of alternatively spliced genes have been associated with genetic disorders, including cancer, neurodevelopmental and heart diseases, and others (313, 324, 325). Alternative splicing has been shown to alter the protein function (74). The range of functional variation between the alternatively spliced isoforms may vary drastically: from a complete loss of original function, due to misfolding and removal by the cell degradation mechanism of the corresponding alternatively spliced isoform, to a subtle difference in the protein functioning, or perhaps the gain of a new function, due to acquiring by the isoform of a new exon that encodes a new functional protein domain. Recently, a high-throughput interactomics study has demonstrated a wide-spread interaction rewiring by the alternatively spliced gene products (84). In some cases, new interactions were shown to be formed. In spite of being very accurate, these large-scale experiments are time-consuming and expensive. Thus, there is a need for a cheaper and faster, *in-silico*, approach. However to date, no computational approaches that predict the edgetic effects of alternatively spliced variants have been introduced.

Here, we propose and compare two machine learning approaches that predict if an alternatively spliced isoform will disrupt the original interaction originally formed by a reference isoform. Machine learning has been previously used in bioinformatics applications that focus on characterization of functional effects caused by the genetic and posttranscriptional variation (84, 321). The applications often define this problem as a classification task and leverage supervised learning approaches, including deep learning, where the training set includes labeled variants for which the function is known and is experimentally validated. The supervised learning approach is designed to benefit from the labeled training set in order to provide an accurate prediction, however the labeling (*i.e.*, functional annotation) may not be feasible for large datasets required by many supervised methods. As an alternative option, a semi-supervised learning method can be introduced, where in addition to the labeled training set, the method can benefit from the knowledge of a large unlabeled dataset, *i.e.*, consisting of alternatively spliced isoforms with unknown functional effects. The semi-supervised learning methods have been popular in the areas of data mining and pattern recognition (326), and have recently been applied to the biological and biomedical data (321, 327).

Both of our new methods, supervised learning and semi-supervised learning, leverage features that focus on determining and characterizing the key differences between the reference isoform that is involved in the original PPI with bait, protein and the alternative isoform whose rewiring property we need to determine. The assessment of the methods has shown that both methods perform remarkably well, correctly characterizing 9 out of 10 alternatively spliced variants. We then demonstrate the utility of this approach by applying it to the tissue-specific transcriptomics datasets obtained from the healthy and western diet-fed and obese mice with the goal to discovering the disease-specific variants with the interaction-rewiring functional impact. In summary, the proposed novel approaches for characterization of edgetic effects of alternatively spliced genes provide a cheap and fast, but nevertheless accurate, alternative to the interactomics experiments and can be used to streamline the high-throughput experimental design by focusing on the most promising candidate isoforms.

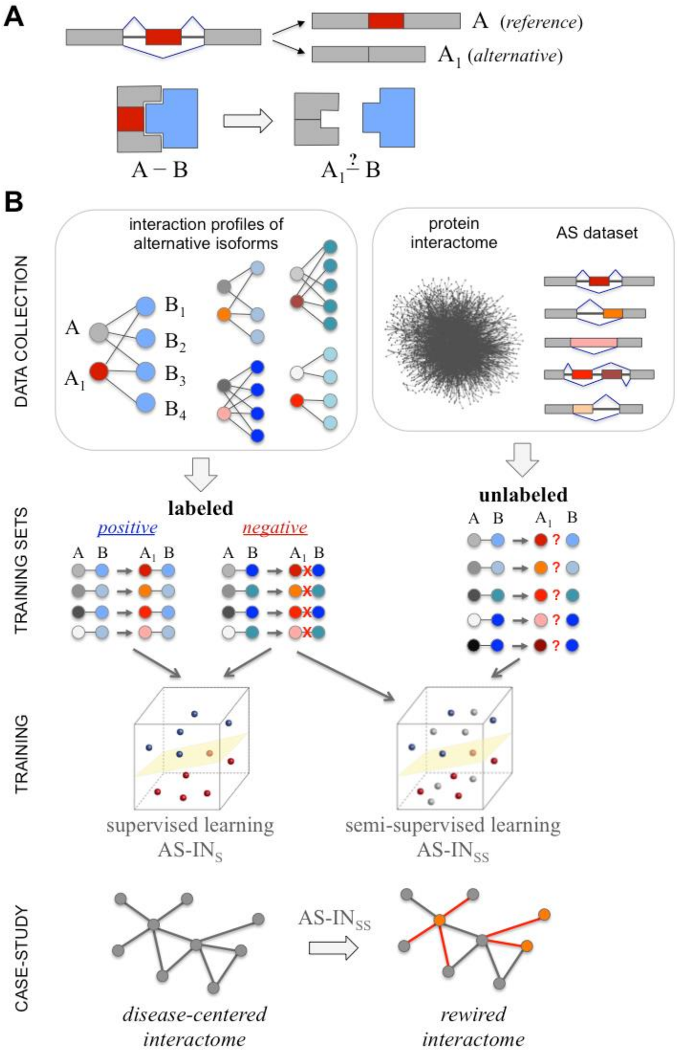


Figure 3.1.2.1. A: Characterization of edgetic effects of AS on PPI formulated as a binary classification problem. **B:** Outline of the overall computational approach.

3.1.3 Results

Datasets and feature statistics

The first dataset (D1) used to generate the training set for the supervised learning classifiers includes 2,501 interactions from 638 genes with 881 alternative spliced isoform. The number of isoform products each gene has ranges from 2 to 110, with an average of 14 isoforms per gene.

The second dataset (D2) composed of known human PPIs (84, 328-332) included 5,460 unique known interactions mediated by the total of 1,230 unique proteins (*i.e*., reference isoforms), 1,082 of which had at least one alternative isoform (in addition to the reference isoform). In total, 4,885 unique alternative isoforms were identified, and 42,654 new, unlabeled, triplets ($\mathbf{A}_1$, $\mathbf{A}_2$, **B**) were formed, where $\mathbf{A}_1$ interacts with **B**, but it is not known whether $\mathbf{A}_2$ interacts with **B**. For this second dataset, the number of isoforms for each gene ranged from 1 to 92 with an average of 32 isoforms per gene. The number of interactions per gene range from 1 to 680, with an average of 31.96 interactions per protein.

Of the three groups of features generated for each data point, perhaps the sparsest were the features corresponding to the occurrence frequency of the SCOP domains. This phenomenon was due to the fact that not all proteins were capable of having at least one SCOP domain predicted using SUPERFAMILY. On the other hand, not all SCOP families were represented across the set of proteins from D1 or D2 equally well. Of 356 proteins in D1, 260 had 1 to 8 SCOP domains predicted by SUPERFAMILY, with a mean of 1.4. Similarly, for 4,028 proteins in D2, 2,917 had 1 to 25 SCOP domains annotated by SUPERFAMILY, with a mean of 2.

Another interesting question was whether any of the delta features (third group, see Methods for more details) could be used to provide an accurate separating boundary. For instance, if an alternatively spliced isoform altered more than *k* residues of the reference isoform, then the alternative isoform would be predicted to eliminate the original interaction. There was a wide range of changes for each feature type, with the values seemingly independent of the fact if the alternative isoform disrupted the original interaction or not (Figure 3.1.3.2A, B, C). The changes in the SCOP domains architecture in the alternative isoform, compared with the reference isoform can be grouped into three categories: no change, deleted domain, or modified domain. For D1, there were 874 (90%) reference isoforms with no change, 99 (10%) isoforms with at least one SCOP domain deleted, and 374 (38 %) with at least one SCOP domain modified. For D2, there were 11,456 (33%) reference isoforms with no change, 23,120 (67%) with at least one domain deleted, or 16,390 (47%) with at least one domain modified.

Method Evaluation

First, using D1, we evaluated the prediction accuracy of three supervised machine learning classifiers: SVM with linear and radial basis function kernels and random forest (Figure 3.1.3.2D, Appendix C2.2 in Suppl. Data). The results of 10-fold cross validation showed that random forest clearly outperformed the two SVM models, reaching the accuracy of 0.86, f-measure of 0.91, MCC of 0.65 and AUC of 0.81. Next, to evaluate the importance of protein domain feature information, we assessed the same methods, but with two different feature vector definitions, one that includes the protein domain features, another one that excludes them. Without protein domains, the performance slightly dropped, with the accuracy values ranged from 0.82 to 0.84, precision from 0.85 to 0.88, recall from 0.91 to 0.94, F1-score from 0.88 to 0.89, MCC from 0.49 to 0.58, and AUC from 0.72 to 0.78. Similarly, to evaluate the importance of using the delta feature information, we assessed the same supervised classifiers with or without these features. Without delta features the performance dropped, with the accuracy values ranging between 0.73 and 0.74, precision ranging between 0.73 and 0.75, and with MCC dropping to a record low range between 0 and 0.09, with the recall being the only metric that improved, ranging from 0.96 to 1.0.

Our second machine learning approach is a semi-supervised learning classifier, which incorporates a large number of unknown label data to train the model. As a result, during the cross-validation the method provided the most accurate performance of all other methods. The assessment values were: accuracy 0.88 (improvement of 0.02 over the top supervised learning classifier), precision 0.92 (improvement of 0.02), recall 0.92 (same as the top supervise classifier), f-score 0.92 (improvement of 0.01), MCC 0.7 (improvement of 0.05), and AUC 0.84 (improvement of 0.03).

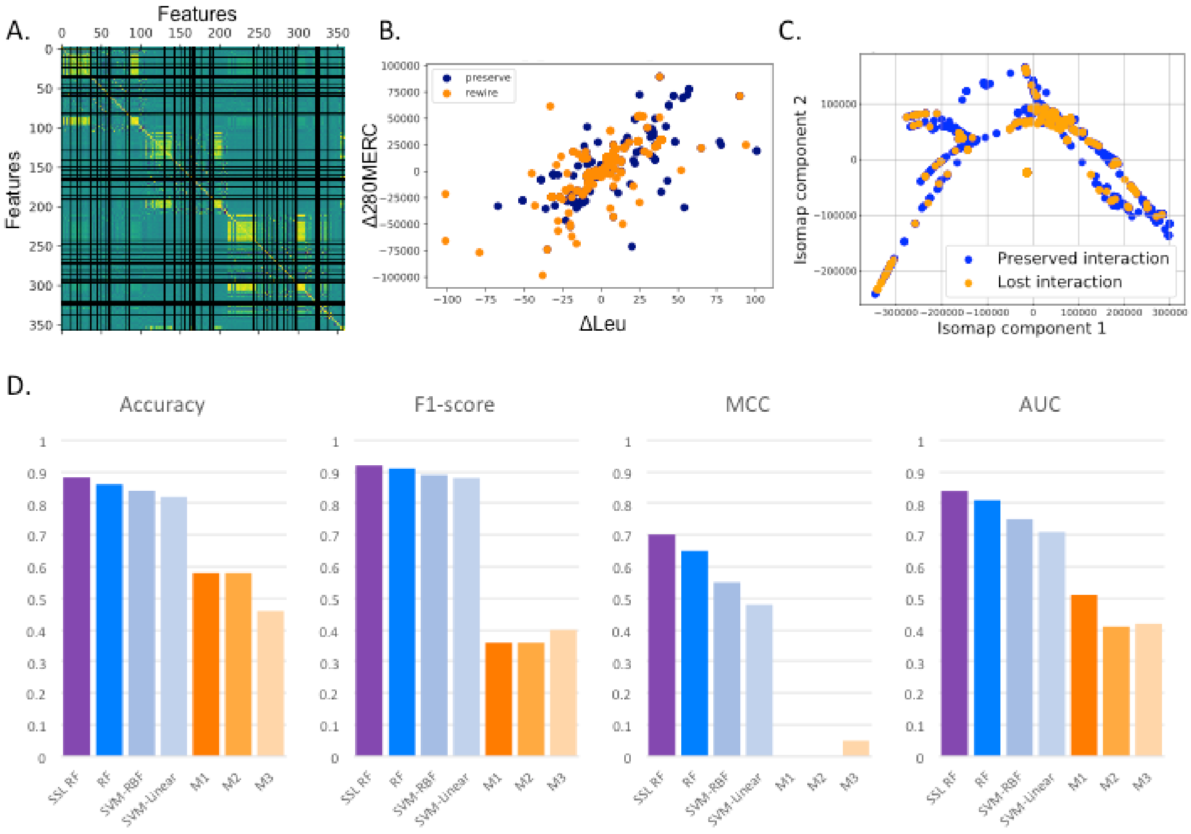


Figure 3.1.3.2. Feature analysis and comparison of our machine learning models with general PPI prediction methods across 4 different metrics (accuracy, F1-score, MCC and AUC). (A) A correlation plot between features used for training machine learning models showing three distinct blocks which are associated with biochemical features of reference isoform, biochemical features of interacting protein and delta biochemical features. Each of those blocks is separate and does not show high correlations with other blocks. (B) A scatterplot based on delta frequency of leucine and another delta of 280MERC coefficient is a typical example of how the feature values are distributed between the representatives of two classes, suggesting that the pairwise comparisons cannot separate two classes well. (C) Isomap visualization of all features through a low-dimensional embedding. Even through powerful manifold learning, we are unable to obtain separable classes in 2D space, which suggests that the problem is challenging. (D)

Performance of our supervised (blue) and semi-supervised (purple) methods vs. three current ab-initio PPI prediction methods (orange) across four metrics.

To the best of our knowledge, this is the first work where a problem of determining the rewiring effect of an alternatively spliced isoform is addressed using a computational approach. However, the same question can be potentially addressed by (1) assuming that the alternative isoform is a new protein, and (2) predicting whether the isoform interacts with the corresponding interaction partner using an *ab initio* PPI prediction method, *i.e.*, without prior knowledge about the interaction of the reference isoform and the same interaction partner. Our evaluation of the three state-of-the-art *ab initio* PPI prediction methods has shown that neither of the methods can be reliable used for our problem: the accuracy ranged between 0.46 and 0.58, recall values ranged between 0.29 and 0.5, precision was between 0.5 and 0.52, f-score was between 0.36 and 0.4, while MCC was between 0 and 0.05 (Figure 3.1.3.2D).

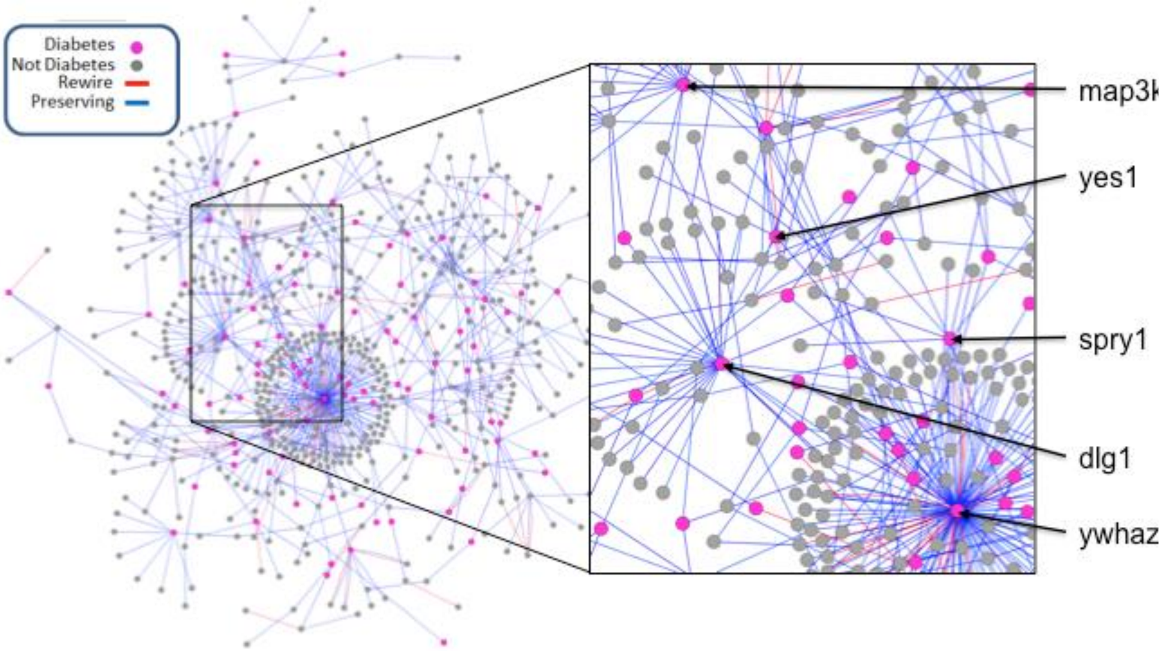


Figure 3.1.3.3: Case study of diabetes-centered mouse interactome. Network focused on alternatively spliced isoforms expressed in the liver and brain tissues, which were found drastically different (at least 5 fold of log2 expression values) between the control and T2D mice induced through Western Diet. The effect of the alternative isoforms was predicted as either disrupting the original PPI (red) or preserving it (blue). To provide context within diabetes, genes that are associated with T2D are colored magenta, while their interaction partners are colored gray. A few well-studied genes linked to T2D are highlighted: *map3k7*, *yes1*, *spry1*, *dlg1*, and *ywhaz*.

Case Study

To demonstrate the utility of AS-IS Tool and extent to which the AS can 'rewire' a disease-centered PPI network, we used our method to predict the edgetic effects due to the disease-specific AS occurring in the brain and liver tissues and obtained from the RNA-Sequencing (RNA-Seq) data extracted from the tissue samples of the healthy mouse and Western Diet (WD) fed mouse that developed T2D. Our deep RNA-sequencing data resulted in 1,899 AS isoforms from 1,608 genes for brain and 5,951 AS isoforms from 3,942 genes for liver with drastically different expression levels (>5 fold) between diabetic and normal mice samples. In total, 6,745 unique isoforms that were drastically differentially expressed were collected for both tissue types.

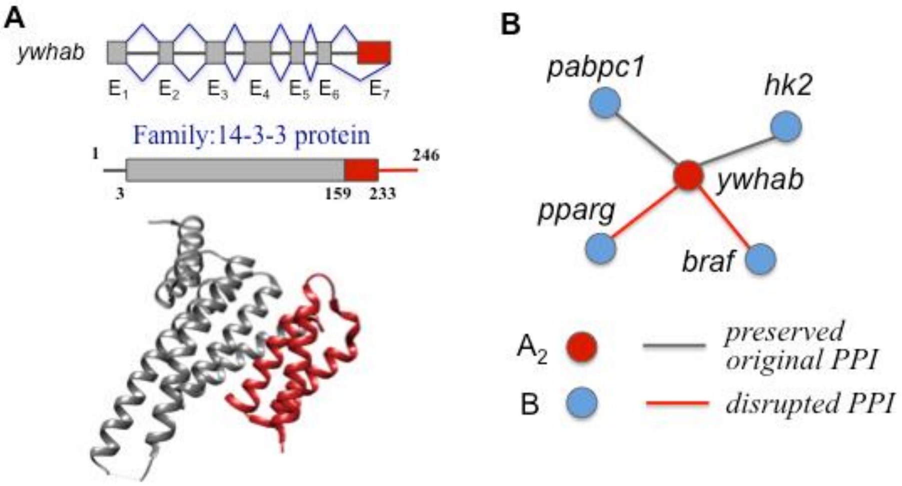


Figure 3.1.3.4: Case study of a gene associated with T2D, whose alternatively spliced isoforms were predicted by AS-IN Tool to rewire some of the currently known PPIs. A. The gene architecture, protein domain architecture, and structure based characterization of the alternatively spliced isoform of *ywhab* gene. The red part of the protein corresponds to the seventh exon and is spliced out in the alternative isoform, A_2. B. As a result, two interactions were predicted to be disrupted by the alternatively spliced isoform A_2 that had been determined to be significantly overexpressed in the tissue samples of WD-fed mouse with T2D disease phenotype, compared with the control.

We then used the experimentally confirmed interactomics data extracted from the STRING database to define 46,862 PPIs mediated by 7,730 mouse proteins corresponding to 7,630. The obtained mouse interactome was considered as the "reference" interactome. Combining this information with the obtained RNA-Seq data allowed us to provide the reference interactome for 135 out of 6,745 unique proteins that were involved in 489 PPIs (Figure 3.1.3.3). The 135 proteins corresponded to the reference isoforms for which 135 alternative isoforms were extracted, and AS-IS Tool was applied to see the edgetic effect of AS. Furthermore, we extracted 1,399 genes from T2D database, three of which were found in our dataset of 135 proteins associated with T2D (Figure 3.1.3.4); these three proteins contributed to 17 PPIs. In summary, AS-IN Tool predicted 128 (26%) interactions, including 10 (59%) T2D-associated interactions to disrupt the corresponding reference interactome (Figure 3.1.3.3).

3.1.4 Discussion

This work describes the first computational approach, AS-IN Tool, which attempts to characterize the edgetic effects of alternatively spliced isoforms on a protein-protein interaction. We formulate this problem by taking advantage of a known PPI, and then characterizing the difference between the reference and alternative isoforms. We develop two feature-based classification methods that leverage the supervised and semi-supervised learning paradigms, taking advantage of traditional features characterizing a PPI and learning the difference caused by alternative splicing. When comparing our top models with the start-of-the-art sequence-based PPI prediction tools, the accuracy of both supervised and semi-supervised methods dominated all three current methods. Furthermore, with the accuracy, precision and recall surpassing 90%, AS-IN Tool becomes a great alternative to the experimental approaches and the only accurate computational approach for this task.

While we understand that the results of predicting edgetic effects of AS isoforms on mouse interactome for our case-study are mere predictions that need experimental validations, we hope that our method can streamline the expensive and time-consuming high-throughput interactomics approach by first identifying a pool of candidate genes for the primer libraries and then pinpointing the isoforms of the outmost interest. AS-IN Tool is available for use as python software package located at https://github.com/korkinlab/asintool.

3.1.3 Methods & Assessment

Overall design and problem formulation

Our approach, **A**lternative **S**plicing **IN**teraction prediction Tool (AS-IN Tool), is designed to address a problem of characterization the rewiring, or edgetic, effects of alternative splicing, which can be formulated as the following binary classification problem (Figure 3.1.2A): Given a known, *reference*, isoform $\mathbf{A}_1$ that is involved in a protein-protein interaction $\mathbf{A}_1\mathbf{-B}$, with another protein **B**, will an *alternatively spliced* isoform of $\mathbf{A}_1$, $\mathbf{A}_2$, *preserve* the interaction with **B** or *disrupt, i.e.* eliminate, it? Triplets ($\mathbf{A}_1$, $\mathbf{A}_2$, **B**) where $\mathbf{A}_2$ preserve the interaction with **B,** given the knowledge that $\mathbf{A}_1$ and **B** interact, are labeled as members of the *negative* class. Alternatively, triplets ($\mathbf{A}_1$, $\mathbf{A}_2$, **B**) where the alternatively spliced isoform $\mathbf{A}_2$ will disrupt the interaction with **B** are labeled as members of the *positive* class. Each of the two developed methods presented in this work is a feature-based approach (Figure 3.1.2B). Specifically, the features encode the information concerning the known interaction $\mathbf{A_1}$–**B**, and information about the changes between $\mathbf{A_2}$ and $\mathbf{A_1}$ that may result in the disrupted interaction.

Supervised classifiers

Support Vector Machines (SVM) belongs to a family of widely used supervised classifiers (333). It is also among the most well-established and popular machine learning approaches in bioinformatics (260, 311). SVM classifiers range from a simple linear, maximum margin, or classifier, where one needs to find a decision boundary separating two classes and represented as a hyperplane in a multi-dimensional feature space, to a more complex classifier represented by a non-linear decision boundary through introducing a non-linear kernel function. Here, two kernel functions were explored: linear and radial basis function (RBF) implemented in *libsvm* library (334). For the SVM models, the parameter optimization was performed using grid search. Optimal values *gamma*=0.005 and C = 9 were obtained after the search in range from *gamma*=0.001 to *gamma*=1 with a step 0.002, and from C=1 to C=100 with a step 1.

Next, since a majority of our features are not correlated, one can expect for another supervised learning classifier, random forest (RF), to be well-suited for the dataset. Random forest (307) is an ensemble classifier, which combines multiple supervised learning classifiers to get a prediction. Random forest uses the ideas of bagging and random split decisions to predict a class of untrained vectors. In bagging, a random selection of the examples in the training set is used to build each decision. A random forest algorithm consists of three basic steps:

1. Draw bootstrap samples.
2. Build decision tree for each sample with the following modifications: select best predictor for node not from all available features but from their random subset.

3. Predict class based on the majority vote of resulting trees.

In this work, the random forest models were trained using scikit-learn package (335), with the default parameters, including Gini criterion.

Semi-supervised classifier based on iterative self-learning random forest

One of the main bottlenecks of supervised learning is the cost of labeling data. The idea behind a semi-supervised learning approach is to utilize a large amount of unlabeled data to improve results of the supervised algorithm. There is a number of existing approaches to the combining of labeled and unlabeled information that try to exploit the underlying structure of the unlabeled data. In most cases, the learning algorithm attempts to find clusters in order to modify the decision boundaries. Here, we implement a simple semi-supervised learning approach, called iterative self-learning random forest, that has previously shown to outperform more advanced semi-supervised learning methods on the protein interaction data represented by heterogeneous features (321). The algorithm starts with a labeled training dataset and a pool of unlabeled feature vectors (Appendix C1.1). At each step, the algorithm trains a supervised learning classifier on the labeled training set. Then, it evaluates the model using a grouped 10-fold cross-validation over the training set. Next, the algorithm is applied to the remaining unlabeled dataset, predicting their labels, selecting several examples, and adding them to the training set. After multiple iterations, the model with the best evaluation score is selected.

Feature design, evaluation, and selection

The question we are answering in this work, if the alternatively spliced isoform $\mathbf{A}_2$ would retain an interaction originally established between the reference isoform $\mathbf{A}_1$ and its interaction partner**,** is somewhat similar to the PPI prediction task. However, here we want to leverage alternative splicing information and the knowledge about the previous interaction as much as possible. This naturally imposes a structure on the features we generate. So far we are using 3 groups of features: (1) biochemical features of the reference isoform and its interacting partner, (2) domain interaction statistical potentials, and (3) so-called delta features. The first group of features are the most straightforward ones and are inspired by the PPI prediction methods (336). These features provide a general outline of different properties of the known interaction. Biochemical features include molecular weight, number of residues, average residue weight, charge, isoelectric point, A280 molecular extinction coefficient for both reduced and cysteine bridges, and several others characteristics Appendix C2.1, Supplementary Data).

The second group represents novel features derived from our DOMMINO database of macromolecular interactions (336). The rationale behind using this group of features is the following: given that an average protein includes multiple protein domains (312), it is important to know which domains are directly involved in a particular PPI. The interdomain interaction is one of the major driving forces behind a protein-protein interaction, with the protein domains often having preferences of interacting with other protein domains. Thus, the frequency of domain-domain interactions differs across different families of related domains. The quantification of those odds is defined as the statistical potential. There are two types of statistical potentials introduced in this work: (1) calculated for a domain from a specific domain family, and (2) calculated for a pair of domains. Statistical potential P_i for a single domain D_i is calculated based on the total number of interactions N_{D_i} extracted from our DOMMINO database for the specific SCOP family (337) this domain belongs to. The SCOP families for each protein sequence are defined using SUPERFAMILY tool (338). Statistical potential P_{ij} for a pair of domain D_i and D_j is calculated based on the total number of occurrences N_{ij} of the interactions between all domains from the same two SCOP families as ***D**_i* and ***D**_j*. Those numbers were transformed into probability using Maxwell-Boltzmann statistic:

$$P_i = \frac{e^{-\frac{N_{D_i}}{N_{mean}}}}{\sum_j e^{-\frac{N_{D_j}}{N_{mean}}}}, \qquad P_{ij} = \frac{e^{-\frac{N_{D_{ij}}}{N_{mean}}}}{\sum_{k,l} e^{-\frac{N_{D_{k,l}}}{N_{mean}}}}$$

where $\boldsymbol{N_{mean}}$ is the average number of interactions for one domain and $\boldsymbol{M}$ is the average number of interactions for a pair of domains present in database.

The third group, the “delta” features, includes selected characteristics of alternative splicing events. Specifically, the features are designed to capture the differences between the original reference isoform and alternatively spliced variant, which may result in a loss of interaction. There are four subgroups of these features. The first subgroup includes features describing the difference in the biochemical characteristics between the reference isoform $\mathbf{A}_1$ and alternatively spliced isoform $\mathbf{A}_2$. The second subgroup includes the difference between the statistical potentials of $\mathbf{A}_1$ and $\mathbf{A}_2$. The third subgroup is a set of simple sequence features that can be computed with a basic sequence alignment, but nevertheless may provide important knowledge. For instance, an exon skipping event that results in a large portion of protein missing, is usually more detrimental to interaction than several small exon skipping events. Similarly, the modifications in N- or C-termini are less likely to result in the interaction rewiring than an equal-sized modification occurring in the protein body. The last subgroup is reliant on SCOP family domain information defied by SUPERFAMILY tool (338), which allows determining if the alternative splicing affects specific protein domains.

To improve performance of the classifiers, three feature selection methods were explored including LASSO, recursive feature elimination (RFE), and principal component analysis (PCA) (339). LASSO is a regression model with l_1 regularization. Because of the l_1 penalty, a solution for the regression naturally contains zero coefficients for many features, thus discarding them from the model. RFE is a widely used feature selection algorithm that consecutively removes one feature from the model and evaluates the results using cross-validation. The optimal number of features is also determined by cross-validation. The last feature selection method, PCA, is a technique that performs the orthogonal transformation on the feature set to obtain linearly uncorrelated components. The number of selected principal components was determined by the 98% explained variance cutoff threshold. Feature selection methods produced varying results for SVM and failed to improve performance of the random forest classifier, which, in turn, showed the most accurate performance among all supervised methods in our study. This result was expected, since the total number of features is significantly smaller than the number of samples, so the random forest model does not overfit, and the influence of less informative features' is limited due to the random subspace selection.

Lastly, we analyze the importance of the individual features. For the calculation of feature importance, we use the mean decrease of impurity in the random forest model, our top-performing supervised classifier. This is a tree-specific metric, and is directly related to the Gini impurity, calculated at each tree node (307). The same feature is present in multiple trees in a random forest model, thus the average decrease in impurity integrates the feedback from all trees that contain this feature.

Method assessment

The performance of the supervised and semi-supervised learning methods is assessed using two evaluation protocols: the cross-validation and comparison with the state-of-the-art *ab-initio* PPI prediction methods. The purpose of cross-validation is to obtain reliable evaluation of the fitted model. It helps to avoid overfitting, a phenomenon which occurs when the model is trained to be oversensitive to some specific signals present in a sample from a training set, but not common for the general population. The main idea is to divide the dataset into *k* subsets. Then, multiple iterations of retraining and re-evaluating model are carried out. For every iteration, the dataset is divided into a test set (represented by one of *k* subsets) and a training set (the rest of the data). Many variations of the cross-validation protocol exist based on the value of *k*, with leave-one-out cross-validation (*k*=1) and 10-fold cross-validation (*k*=10) being the most common. 10-fold cross-validation is deemed to be one of the most stable protocols, so we are using it in this work.

Regular cross-validation performs well if we can consider each of the data points to be truly independent. Unfortunately, it is not a case for our dataset, where multiple isoforms are the products of the same gene. If one subset of related isoforms is present in the training set and another subset is present in the testing set, then our model is provided with unfair advantage during the evaluation. Since we are expecting the model to generalize well, and thus, to work on novel isoforms, with no prior information about them, we want our evaluation to be as close to this scenario as possible. Therefore, the original 10-fold cross validation is modified into a grouped cross-validation. Specifically, we group all isoforms that are products of the same gene, and each group is then allocated exclusively either into the training set or into the test set. This grouped cross-validation protocol is more stringent and thus is expected to reduce the reported accuracy of the method.

In our second evaluation, we compare the performance of our methods with the state-of-the-art *ab initio* PPI prediction tools, including TRI_Tool (M1) (340), LR_PPI (341) with negative set 1 (M2), and LR_PPI with negative set 2 (M3). One can apply each of these tools to predict if a PPI between $\mathbf{A}_2$ and **B** exists, independently of the knowledge of whether or not $\mathbf{A}_1$ and **B** interact.

The performance of each method is measured using standard measures, including accuracy (*Acc*), recall (also called sensitivity, *Rec*), precision (*Pre*), *f*-measure (*F1-score*), Matthews correlation coefficient (*MCC*), and area under the curve (*AUC*). Area under the curve can be computed with the help of Gini coefficient (G_1):

$$AUC = \frac{1+G_1}{2} G_1 = 1 - \sum (X_k - X_{k-1})\,(Y_k + Y_{k-1}),$$

where X_i is a true positive rate (TPR), and Y_i is a false positive rate (FPR) for the threshold *i*. A pair (X_i,Y_i) defines a point on the receiver operating characteristic (ROC) curve.

Datasets

For training and evaluation of the supervised machine learning classifiers, we use an experimental human high-throughput interactomics dataset developed for the purpose of analyzing AS effects (84). This dataset is initially randomly split for 10 fold cross validation protocol. However, the folds are then modified to ensure all isoforms related to a gene are either in the test or training split, as described in the group cross validation protocol above.

The second dataset of unknown effects by alternative isoforms is used as a source of the unlabeled data in the training of the semi-supervised classifier. To obtain the unlabeled dataset, we first consider another high-throughput human interactome (84, 328-331). We then remove

RNA-protein interactions as well as interactions from the oligomeric complexes, leaving only PPIs between two individual proteins. Then, to compile a list of AS isoforms for all proteins that are involved in the pre-processed list of PPIs, we downloaded the protein, gene, and isoform mapping from Ensembl (GRCh38 version 91) (342). All protein-coding isoforms related to a reference protein that is involved in a PPI are then included into our list of AS isoforms.

Case Study: An application of AS-IN to the diabetes-centered mouse interactome

To test the utility of our approach, AS-IN is applied to study how alternatively spliced isoforms in a mouse model of type 2 diabetes (T2D) can rewire a disease-centered interactome. The dataset used for our case study is obtained from an environmentally derived T2D mouse model (343), where we extracted and analyzed RNA-Seq from brain and liver tissues between the diabetic and normal control mice. Previous studies have demonstrated that ingesting a western diet (WD), high in fat and refined carbohydrates, leads to activation of the Akt and mTOR pathways (344). The activation of these signaling processes from the food intake, in turn, has been shown to result in inhibiting insulin metabolic signaling and leading to T2D (345). Specifically, after 3 months of feeding the WD to C57BL/6J mice, T2D is developed. To explore the AS effect on T2D in this pilot study, we selected two mice: one fed WD and one without. From these mice, brain and liver were dissected and preserved using standard techniques.

Using Qiagen's RNeasy Mini Kit, total RNA is isolated from the dissected brain and liver samples. Library preparation for RNA-Seq is done using TruSeq RNA v2 to isolate mRNA and prepare for sequencing. After validating RNA quality using RNA integrity number (RIN) on an Agilent 2500 BioAnalyzer, samples are deep sequenced on an Illumina HiSeq 2000 using 2 lanes for each sample to achieve close to 100 million 75 paired-end reads per sample. The RNA sequencing analysis pipeline includes Trimmomatic with default settings to remove the low quality reads (218), Tophat v2 to align on GRCm38.p5 (346), and Cufflinks v2 to reassemble and quantify expression levels (347). Due to only 1 sample per group (WD or wild type, brain or liver), we cannot rely on standard statistics to determine relevant isoforms. Thus, we use a strict cutoff of 5 $\log_2$-fold changes between WD and wild type mice, for each of the two tissue types, to identify the relevant isoforms.

The initial set of the relevant isoforms is further reduced based on the known gene association to T2D. To do that, we collect the data from Type 2 Diabetes Knowledge Portal (http://www.type2diabetesgenetics.org/), which houses the data from multiple genome-wide association studies (GWAS) to identify genetic associations from single nucleotide variations (SNVs) with diabetes type 2 (348). We downloaded the data from 9 GWAS studies (349-351) and selected the genes that are near to or carry SNVs with a p-value of $5*10^{-5}$ as associated with T2D. Finally, as a source of the reference PPIs, we construct the mouse interactome from STRING database, selecting all mouse PPI that have at least one experimental reference (352).

3.2 Systematic annotation of mutation underlies importance of extracellular interactions in cancer through de novo prediction of protein binding sites

3.2.1 Abstract

The standard molecular-phenotypic definition of cancer relates genomic instability with increased proliferation, capable of evading growth suppressors, resisting apoptosis, avoiding immune destruction, and metastasis. Recurrent single nucleotide variations (SNVs) as a result of genomic instability may target essential protein functions such as phosphorylation, acetylation, or ubiquitination site. One essential protein function, occurs at physical contact interfaces characterized as protein binding sites. This work's focus is to identify whether cancer SNVs target a protein's binding site. During the past decade, numerous protein binding site prediction methods have been published using sequence which can be applied to any known protein sequence; however, they are less accurate then structure-based methods, which are limited by the narrow number of protein structural data. Therefore, we developed a new *de novo* protein binding site prediction method (**Co**mparative **B**inding **R**egion **A**nnotator (COBRA)), which expands the number of proteins that can be assessed without losing accuracy. Our new *de novo* protein prediction method was assessed across eight different methods. Using COBRA we performed a large-scale annotation of SNVs across eight different cancer types to evaluate for their protein binding sites (bSNVs). bSNVs were assessed for their enrichment of known cancer drivers, functional enrichment, patient survivability, clinical significance, and protein-protein interaction (PPI) network. Finally, we analyzed whether the observed patterns were similar to SNVs that affect phosphorylation (pSNVs).

3.2.2 Introduction

Genomic instability is considered a common characteristic of cancer (353). The resulting genomic instability allows cancer cells to have increased proliferation, capable of evading growth suppressors, resisting apoptosis, avoiding immune destruction, and dissemination (354). Understanding the relationship between the phenotypic observations of cancer and genomic instability is an area of active research (355). One observation is the increased mutation rate of single nucleotide variations (SNVs) within cancerous cells (356). Recent studies have demonstrated SNVs may target essential protein functions such as phosphorylation, acetylation, or ubiquitination sites (357, 358). As a result, systemic annotation of SNVs have been utilized to identify genes and signaling pathways driving cancer progression (359).

One key observation from these analyses is disruption and dysregulation of protein interactions could be instrumental to allow cancer progression (360). Protein interactions occur at interfaces characterized as protein binding sites, which are the group (s) of amino acid residues that are in physical contact (241). During the past decade, numerous protein binding site prediction methods have been published. The existing methods make use of two types of information: sequence- and structure-based methods (241). The sequence-based methods can be applied to essentially any known protein sequence; however, they are less accurate then structure-based methods, which are limited by the narrow number of protein structural data (241). Therefore, the goal for a new binding site prediction approach would be to bridge the coverage gap while preserving and hopefully improving the prediction accuracy.

This work's focus is to identify whether cancer SNVs target a protein's binding site, which would allow for modification of protein interactions. The first goal of this work was to design a new *de novo* protein binding site prediction method. This method uses a **Co**mparative **B**inding **R**egion **A**nnotator (COBRA), which does not depend on interaction information between a protein and its interaction partner. This allows the method to increase the number of protein binding sites that can be annotated without losing prediction accuracy. This method was evaluated then compared with the state-of-the-art sequence- and structure-based approaches. The second goal of this work was to determine whether protein binding sites are a recipient of cancer SNVs. To achieve this goal, a large-scale annotation of SNVs across eight different cancer types from the cancer genome atlas (TCGA) were assessed for their protein binding sites (bSNVs). bSNVs were assessed for their enrichment of known cancer drivers, functional enrichment, patient survivability, clinical significance, and protein-protein interaction (PPI) network. Finally, we analyzed whether the observed patterns were similar to SNVs that affect phosphorylation (pSNVs) (241).

3.2.3 Results

The goal of this work is to assess whether there is any prevalence of SNVs in cancer to protein binding sites (bSNVs). In order to, conduct this analysis, we developed a new *de novo* protein binding site prediction method (COBRA, Figure 3.2.3.1) that increases the coverage of proteins that can be assessed while maintaining the accuracy achievable by structure-based methods. To validate the quality of predictions, we assessed and compared it with the sequence- and structure-based methods introduced here as well as current state-of-the-art approaches. *De novo* methods that predict protein binding sites rely only on the sequence or structure information about the target protein and do not depend on the interaction information of this protein and its interaction partner. The first group of *de novo* methods includes sequence-based methods, which rely exclusively on the features that can be generated from the query protein's sequence. The second group includes structure-based approaches that rely on both the sequence-based and structure-based features and require an experimentally obtained structure of the query protein. When an experimentally solved structure of the query protein is not available, it is possible to apply a structure-based method to a comparative model, however the prediction accuracy is likely to be worse due to structural errors in the model.

Using COBRA, we assessed bSNVs prevalence across eight different cancer types primarily from the cancer genome atlas (TCGA): breast, colorectal, liver, lung, ovarian, and pancreatic cancer, as well as glioblastoma (two datasets) and leukemia. SNVs were classified as either on the binding site (bSNV) or not (non-bSNV). Genes with bSNVs or non-bSNVs were assessed for their enrichment of known cancer drivers, functional enrichment, patient survivability, clinical significance, and protein-protein interaction (PPI) network. Finally, we analyzed whether the observed patterns were similar to SNVs that affect phosphorylation (pSNVs) (241). New approach bridges the accuracy gap between the sequence-based and structure-based prediction methods

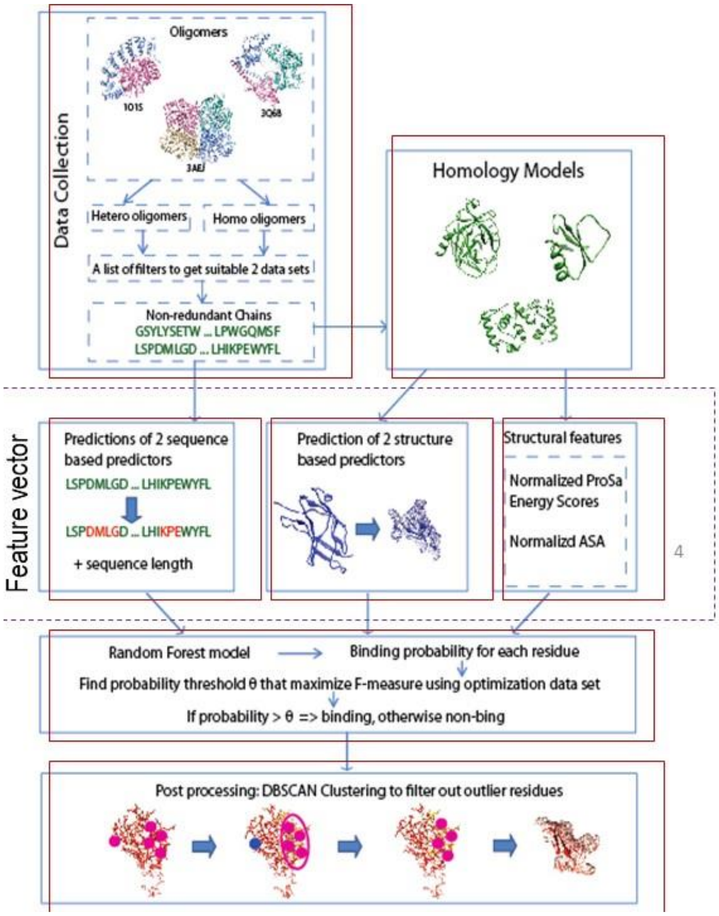

Data Collection
Oligomers
1O1S
3Q68
3AEJ
Hetero oligomers
Homo oligomers
A list of filters to get suitable 2 data sets
Non-redundant Chains
GSYLYSETW ... LPWGQMSF
LSPDMLGD ... LHIKPEWYFL
Homology Models
Feature vector
Predictions of 2 sequence based predictors
LSPDMLGD ... LHIKPEWYFL
LSPDMLGD ... LHIKPEWYFL
+ sequence length
Prediction of 2 structure based predictors
Structural features
Normalized ProSa Energy Scores
Normalizd ASA
4
Random Forest model
Binding probability for each residue
Find probability threshold θ that maximize F-measure using optimization data set
If probability > θ => binding, otherwise non-bing
Post processing: DBSCAN Clustering to filter out outlier residues

Figure 3.2.3.1. Summary of COBRA Methodology. The data collected a comprehensive non-redundant set of protein structures and comparative models of varied quality for each protein, which include at least one experimentally defined protein binding site. A feature vector was generated for each model through combining both sequence, structure, and homology-based models to summarize their properties for computational learning. Using these feature vectors, models were analyzed using both training and testing of several supervised classifiers. We post-process the prediction results by using a density-based clustering to screen the outliers.

Our comprehensive evaluation of all eight methods to classify the binding site residues of both interaction types, heteromeric and homomeric, provided several insights that were consistent across all heteromeric and homomeric classifiers. First, our structure-based approach, StrucBRA, applied to the proteins with known native structures outperformed every other tested method, while the sequence-based approach, SeqBRA, applied to protein sequences demonstrated the worst performance among all methods, as expected. The performance differences between SeqBRA and StrucBRA for the homomeric binding sites (specificity values of 0.52 *vs*. 0.79 and accuracy values of 0.55 *vs*. 0.75, correspondingly) was more profound than for the heteromeric binding sites (specificity values of 0.65 *vs*. 0.77 and accuracy values of 0.64 *vs*. 0.73, correspondingly). Both methods had considerably lower precision values (0.28 *vs*. 0.39 for heteromeric and 0.25 *vs*. 0.42 for homomeric binding sites) by allowing higher number of false positives than false negatives.

The performance of the whole family of homology-based COBRA methods fitted naturally between SeqBRA and StrucBRA methods. Not surprisingly, the performance of COBRA methods became better with the higher quality of the comparative model, as defined by the sequence identity between the target sequence and template structure. Specifically, the performance of both StrucBRA and COBRA on the comparative models as a function of the target-template sequence identity gradually improved when making prediction on the models that were obtained on increasingly similar template structures, although the improvement was not so profound for models based on the templates with sequence identities of 50% and higher. Interestingly, some of the performance measures, such as f-measure and precision for SeqBRA (performed on sequences of the comparative models) and original StrucBRA (performed on the native structures of the target proteins whose comparative models were obtained) indicated slight dependence on the target-template sequence identity. On the other hand, the accuracy and recall measures showed no such dependence. Even more importantly, when applied to the testing sets of comparative models, StrucBRA performed worse than every single COBRA method on the same sets, which indicates the importance of developing our approach for binding site prediction on comparative models.

Finally, a COBRA method that was optimized with f-score followed by clustering of the binding site residues performed better than any other COBRA method, approaching the performance of StrucBRA on the native structures. For prediction of heteromeric binding site residues, this top performing COBRA method was sometimes even better than the performance of structure-based StrucBRA on the native protein structures of the same proteins whose comparative models were tested by COBRA; the performance was evaluated by f-measure and MCC.

In summary, several important observations had been made. First, a sequence-based method for binding site prediction, SeqBRA, while applicable to any protein sequence, demonstrated the worst performance. Second, the structure-based method, StrucBRA, demonstrated the best performance on the experimental structures while applicable to the fewest proteins. Third, our comparative modeling-based method, COBRA, was applicable to any protein whose structure can be modeled using comparative modeling; this number is expected to be substantially greater than the number of existing experimental structures (241). Furthermore, when applied to a comparative model, COBRA was comparable with StrucBRA applied to the corresponding experimental structure and sometimes even outperformed the structure-based method. Fourth, the performance of StrucBRA on the same set of comparative models was worse than the performance of COBRA, indicating that our new method is a more accurate alternative to the existing structural approaches when applied to models as opposed to the native structures. Last, we showed the utility of clustering post-processing of the identified binding site residues.

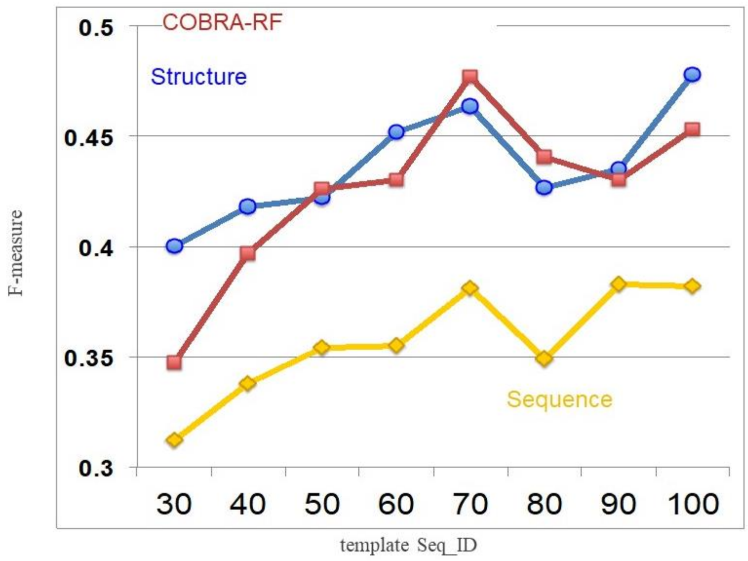


Figure 3.2.3.2. Assessment Between Sequence, Structure, and COBRA-RF. Structure based binding site prediction methods represent the highest quality of models, but suffer from only being able to be applied to the limited proteins with resolved structures. However, sequence based methods can be applied to any annotated protein, but suffer from lower prediction quality. COBRA-RF bridges the gap by allowing increased coverage of the number of proteins that can be assessed, without losing quality of predictions.

Characterization of cancer-associated somatic mutations in protein binding sites

The second goal of this study was to determine whether single nucleotide variations (SNVs) that were localized on the predicted protein binding regions (bSNVs) had any potential role in cancer. We utilized The Cancer Genome Atlas (TCGA) dataset, which included 10,900 non-synonymous SNVs on 6,188 genes involved in eight types of cancer: breast, colorectal, liver, lung, ovarian, and pancreatic cancer, as well as glioblastoma (two datasets) and leukemia. Among these genes, 1,259 protein products (~12%) or protein fragments were covered by the experimentally solved structures. An additional set of 2,029 protein products could be resolved through comparative modeling, thus expanding the number of structurally resolved proteins to 30%. Such a significant expansion of the dataset for the analysis prompted us to use COBRA approach for the annotation of bSNVs based on the predicted protein binding sites. When applying our most accurate COBRA-RF method to the combined set of 3,288 structurally characterized proteins, we predicted at least one bSNV for each of 1,203 proteins. Most genes had only one bSNV annotated, while a few of these genes had multiple bSNVs, as many as 16 different bSNVs on TP53 or 10 on EPHA3. Both of these genes have been well known for their roles in cancer and tumorigenicity (361-363). Interestingly, these two genes were among the top 5 most well represented bSNVs in the individual samples of specific cancers, together with KRAS, EFGR, and CDKN2A.

The analysis of the prevalence of bSNVs across the eight cancer types revealed several interesting patterns. First, bSNVs constituted a large part of the mutations across all eight types, between 23% and 67% (41% average) of all SNVs. bSNVs affected a number of genes, 1,137 out of 6,187 (18%) with a range of 68-685 in all cancer types, but only 4 out of 41 (10%) for leukemia and 116 out of 1171 (10%) for glioblastoma. The absolute numbers of bSNVs and affected gene were dependent on the number of patients. Normalization over the patient sample size revealed a range of 2-22 genes affected per patient in most cancer types except liver, which were 79. Considering the average number of bSNVs per patient ranged between 2 to 90 versus 8 to 116 for non-bSNV mutations. Next, when analyzing genes that had bSNVs in three or more cancer types, we found that some of them, such as TP53, EPHA3, NTRK3, and KRAS had been previously associated with cancer (241), while others, such as TNI3K and IGSF9, had not (241). The total number of bSNVs observed for genes not associated with the cancer progression (347) were significantly higher than expected (173), based on the average frequency of this mutation type across all genes ($p < 2.2*10^{-16}$).

Genes with bSNVs are involved in extracellular molecular interactions and are linked to olfactory function

The comparison of Gene Ontology (GO) molecular functions enriched among genes carrying at least one bSNV and genes carrying at least one non-bSNV mutation, found several patterns shared between these two gene types as well as patterns exclusive to bSNV genes. Both bSNV and non-bSNV genes were involved in many types of macromolecular binding and kinase activities, partially since 737 genes carried both bSNV and non-bSNV mutations. Of interest were the binding functions of bSNV genes that were associated with the transmembrane receptor activities. Strikingly, after removing the functions common to both types, we found that bSNV genes were uniquely enriched for GO terms associated with DNA repair and extracellular receptor activity in contrast to non-bSNVs that were enriched in classical intra-cellular binding activity (Figure 3.2.3.3).

Perhaps the most unexpected was the finding that the top enriched GO term unique to bSNVs was the olfactory receptor activity term (GO:0004984). Olfactory receptors, the genes that are expressed tin the cellular membranes and are implicated in the sense of smell, were thought by some studies to be unrelated to cancer (364). However, while the OR are named due to their expression in the sensory neurons of the olfactory epithelium, they were nevertheless found to be present in various other tissues such as brain, heart, kidney, testis, muscle, where their potential functions are largely unknown (365). Most recently, several studies have suggested that OR genes play an important role in several cancers (241, 366-369).

We finally compared the enrichment of existing functional pathways with bSNVs and non-bSNVs. While both datasets shared the enrichment in pathways associated with cancer, such as Signaling Pathways in Glioblastoma term and a more general Pathways in Cancer term, several cancer-related and signaling pathways were found to be uniquely enriched with bSNVs. These pathways included Non-Small Cell Lung Cancer as well as Small Cell Lung Cancer and Signaling by GPCR. Based on these results, we further investigated if any of the currently known cancer pathways are enriched with bSNVs. We considered main cancer pathways of three types: cell survival, cell fate, and genome maintenance (241). The 12 specific pathways included in this analysis were: RAS, cell cycle apoptosis, PI3K, Stat, Mapk, TGF-β, DNA damage control, Notch, HH, APC, chromatin modification, and transcriptional regulation. Surprisingly, only 4 out of 12 cancer pathways were enriched with bSNVs, while 10 out of 12 pathways were enriched with non-bSNVs.

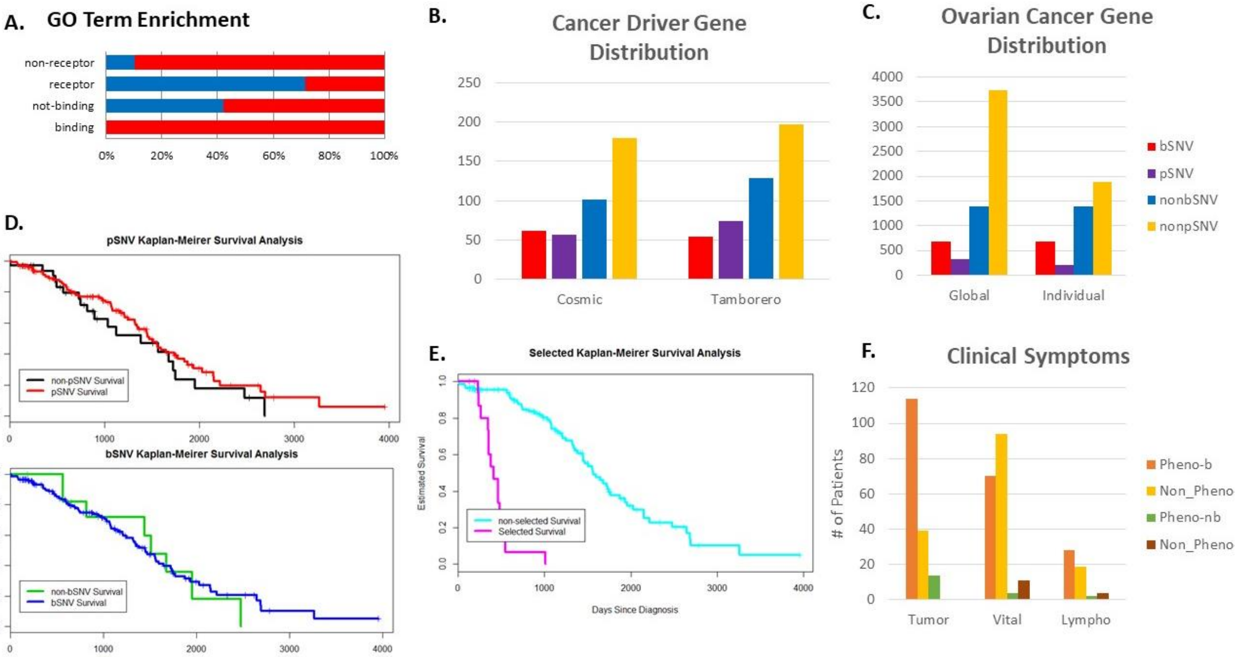


Figure 3.2.3.3. Summary of Functional analysis of bSNV and pSNV. A. Gene Oncology terms summarized into related to bSNV (blue) and nonbSNV (red) B. Assessment of the distribution of bSNV, nonbSNV, pSNV, and nonpSNV related to cancer driver genes and on ovarian cancer specific patients (C) D. The general trends on survival of patients with at least one pSNV or bSNV, which were not considered significant. However, the general trend is

pSNV are associated with increased survival whereas bSNV were associated with decreased survival. E. Survival analysis was instead conducted to look for genes with bSNV that were associated with survival. This caused the analysis to discover 40 genes that have a significant impact on survival. F. After exploring clinical factors, found only tumor presence and absence to be statistically significant with presence of bSNV.

Relationship between the SNVs affecting binding sites and phosphorylation sites

We next compared the distribution of bSNVs with another large-scale functionally annotated subset of mutations recently obtained from the same TCGA dataset of SNVs—phosphorylation –associated mutations, or pSNVs (241). The list of cancer genes with high confidence was collected from Cancer Gene Census (http://cancer.sanger.ac.uk/cancergenome/projects/census/). The same gene list was also previously used in the pSNV analysis (241). In total, 522 cancer genes were annotated. 61 cancer genes were observed in our whole gene list, while 32 are expected based on the number of detected genes in our dataset ($p < 1.2*10^{-6}$, hypergeometric test). The comparison of distributions of bSNVs and pSNVs across cancers in terms of either absolute numbers or normalized numbers, to remove the effect of the dataset size, has revealed remarkably similar contributions of bSNVs and pSNVs to the cancer-specific mutations.

Given that both mutation type's bSNVs and pSNVs affect many genes across multiple pathways, we next wanted to characterize the system-wide distribution of bSNV and pSNV mutation sets. To do so, we constructed a protein-protein interaction network centered around the proteins associated with each cancer type using the known experimental information, and mapped bSNVs and pSNVs on this network (Figure 3.2.3.4). We found that the overall number of genes containing bSNV mutations whose protein products are a part of a PPI cancer network across each cancer type is significantly larger than the number of genes with mapped pSNVs. Intriguingly, the bSNVs and pSNVs have a complementary cumulative effect on the cancer interactome: for each cancer type considered in this work, the dataset of genes that are a part of the interactome and that carry bSNVs have little to no overlap with genes carrying pSNVs with 40 – 80% overlap. For example, of the 114 patients in pancreatic cancer, there are 80 genes with bSNVs and 37 genes with pSNVs, but 32 containing both pSNVs and bSNVs.

Case study: KRAS, a gene with depleted pSNVs and enriched bSNVs

Not surprisingly, the highest number of unique (w.r.t their locations on the gene) somatic bSNVs across all samples of all different cancer types were found in TP53. However, TP53 did not have the highest total number of somatic bSNVs—another gene, KRAS, which was widely reported to link with cancer (241). We therefore focused on investigating the role of diverse types of somatic mutations for KRAS. Unlike TP53, the majority of SNVs (186) contributed to bSNVs (98.4%), but only to five positions on the gene. None of these mutations were reported to be associated with a phosphorylation site in a recent comprehensive analysis (241). Strikingly, only three of the total 189 SNVs extracted from TCGA dataset for KRAS were located on the positions 117 and 146, which were not predicted to be the binding site residues. Thus, most of the mutations detected in this gene was predicted to lie on the protein's putative binding site. We note that the KRAS function of normal tissue signaling had been reported to be compromised during cancer development (241). This was consistent with the bSNV functional annotations reported above.

Cancer Driver- bSNV Enrichment

Since a hallmark of cancer is the loss of DNA repair, a higher than normal genomic mutation rate is expected. This leads to genes being classified as either cancer drivers (part of the cancer process) or a byproduct of cancer (passengers) (370). We wished to explore whether genes with bSNVs were associated with cancer drivers or passengers. The Catalogue of Somatic Mutations in Cancer (Cosmic) is an ongoing effort to identify cancer causing genes (371). Within this database, genes are identified into two groups: Tier 1 (T1) and Tier 2 (T2). Tier 1 cancer driver genes include documented activity relevant to cancer. Tier 2 cancer driver genes include genes with strong indication of a role in cancer, but with less extensive evidence available. Out of the 699 genes identified as cancer driver, 48 and 81 overlaps with bSNV and non-bSNV; respectively. Splitting the cancer driver genes into T1 and T2 leads to a slight difference. T1 genes with 34 and 67 bSNV and non-bSNV genes; respectively. T2 genes with 14 for both bSNV and non-bSNV. The difference between bSNV and non-bSNV is considered not statistically significant (0.5801 p-value) based on a Fischer's exact test.

Cancer Survivability – bSNV

We wanted to explore our hypothesis that bSNV would be critical at a certain survival stage of cancer. Similarly, an analysis previously conducted on pSNVs, we analyzed survival data from TCGA (The Cancer Genome Atlas) for ovarian cancer patients on bSNVs using the Kaplan-Meier survival curve. This analysis was conducted from two different perspectives: global and individual scale. The global scale refers to a patient is considered to have a bSNV if there is at least one gene with a bSNV. Whereas, the individual scale considers patients with the presence or absence of a specific bSNV gene. The survival data on a global scale for both bSNV and pSNV are not significant when considering at the global scale. However, the global trend is that bSNVs has a negative impact on survival whereas pSNVs has a positive impact. However, when considering presence of bSNV on the individual gene level there are 40 out of 2080 genes that have a statistically significant difference. Performing gene enrichment analysis on these 40 genes leads only to olfactory receptors as an enriched function. However, this only accounts for 4 out of 40 genes present. There is not an enriched function or pathway associated when these 4 olfactory receptor genes are excluded from the analysis. Furthermore, there is only one cancer driver (*Akd1*) associated with these 40 genes. All 40 genes have a negative impact on survival.

Clinical Data - bSNV

To further explore any potential association of bSNV on clinical data, all clinical information available for the ovarian cancer patients was tested using the Fischer t-test for statistical significant for bSNV enrichment. Patients were defined as bSNV associated if at least one gene with a bSNV was considered, which considered 164 patients as bSNV and 16 as non-bSNV. In total, 53 clinical factors were assessed. Examples of clinical data are race, tumor grade, and lymphovascular invasion indicator. All clinical factors were not statistically significant except for the presence or absence of a tumor (0.0203 p-value). This is due to the observation that all patients that did not have a tumor had bSNVs.

3.2.4 Methods

Rational behind a new computational method for prediction of protein binding site. The objective of this work is to determine how many disease-associated SNVs are located within protein binding sites and characterize them. To obtain a comprehensive atlas of genetic variation implicated in protein binding, we develop a new method that predicts the protein binding sites. The method is designed to have significantly greater coverage, compared to state-of-the-art structure-based methods, without significant loss of the method's accuracy. This is done by expanding the set of target proteins with experimentally solved structures to the set of proteins with the structures modeled by a comparative modeling approach. A supervised learning method is trained to dynamically adjust importance of the information coming from the sequence- and structure-based features for a comparative model of the query protein. The model training intrinsically depends on the modeled quality of candidate region of the query protein and follows the following simple rationale. Depending on the sequence identity between the query sequence and template structure, the model's quality can range from very poor to excellent, often with some regions modeled better than others. In the former case, the new method will rely primarily on the sequence-based features, since the extracted structural information may be unreliable, while in the latter case, the structural features will be produced from a near-perfect comparative model, thus those will be the preferred source of information for the method.

The basic stages of our computational approach are organized as follows (Figure 3.2.3.1). We first assemble a comprehensive non-redundant set of protein structures, together with the comparative models of varied quality for each protein. Each structure or model in this dataset also has at least one experimentally defined protein binding site. The structures and their corresponding models are used for both training and testing of several supervised classifiers. Second, for each comparative model, we generate region-specific energy scores, each score corresponding to the model quality of a small protein region. Third, for each residue we generate a feature vector by combining the model energy scores with other sequence- and structure-based properties. Fourth, the feature vectors are used to train a supervised learning classifier that predicts whether a residue belongs to a protein binding site. Finally, after the model is trained and applied to predict the binding site residues, we post-process the prediction results by using a density-based clustering to screen the outliers.

A dataset of structurally resolved PPI complexes and their comparative models. The data to train and test a protein binding site classifier consist of protein structures known to participate in protein-protein interactions (PPIs), their comparative models, and experimentally known protein binding sites extracted from PPI complex structures. The complexes are collected from DOMMINO, a comprehensive database of macromolecular interactions (316). First, we collect all dimers, trimers, and tetramers solved by X-ray with resolution ≤ 3.0 Å. Second, we group hetero- and homo-oligomeric structures and map the protein binding sites onto protein chains comprising each oligomer. The homo-oligomers are distinguished from hetero-oligomers through the pairwise sequence similarity: if sequence identity between each pair of chains comprising an oligomer is greater or equal to 90%, it is defined as a homo-oligomer, otherwise it is defined as a hetero-oligomer. The procedure results in 22,800 dimers (16,076 homodimers and 6,724 heterodimers), 3,703 trimers (1,511 homotrimers and 2,192 heterotrimers) and 4,121 tetramers (2,741 homotetramers and 1,380 heterotetramers).

For each protein chain in a PPI complex, we identify the binding residues that constitute its protein binding site. Given an interaction between two proteins, a residue on one protein is defined as a binding site residue if at least one of its heavy atoms is within 6Å of any heavy atom in the other protein. For a protein chain in a homo-oligomer, we extracted protein binding sites from all binary PPIs involving this chain, identifying all residues involved in these binding sites. In a hetero-oligomeric complex, there may be some homodimeric PPIs shared between two identical subunits, as defined above. Thus, for a protein chain in a hetero-oligomeric complex, we consider only those protein binding sites that are involved in the heterodimeric PPIs. For both types of interactions, protein binding sites of less than 3 residues are considered as artifacts and discarded. As a result, we collect two sets of protein chains extracted from hetero- and homo-oligomers respectively, where, each of the chains is annotated with at least one protein binding site.

Next, a set of filters is applied to remove redundancy and decrease the error rate during testing. First, to reduce the number of disordered proteins in the dataset, protein chains with length less than 30 residues are removed. Second, we exclude another stand-alone class of proteins, trans-membrane proteins, from the consideration using PDB-TM database (241). Third, we exclude proteins whose structure is incomplete. Specifically, if the missing residue ratio (241) for a protein chain is greater than 10%, the protein chain will be excluded from the dataset. Finally, we reduce the training bias from the homologous proteins: we apply BLATClust to cluster the remaining protein chains using pair-wise sequence identity of 30% as a threshold. From each cluster, we pick up a representative protein chain with the highest structural resolution and the longest sequence length. After applying the above filters, the final dataset includes 1,160 protein chains involved in heteromeric interactions and 3,883 protein chains involved in homomeric interactions. Using this final dataset, we collect 51,887 binding site residues (positive set) and 215,512 non-binding residues (negative set) for heteromeric interaction dataset, and 203,709 binding site residues (positive set) and 819,762 non-binding residues (negative set) for homomeric interaction dataset.

Comparative models are obtained using MODELLER-9 software (372). Each comparative model was assessed using DOPE, a distance-dependent statistical potential calculated from known protein structures and available through MODELLER-9 (373). Models are generated for each sequence in the dataset through the following five steps: (1) Generate a sequence profile by running three rounds of PSI-BLAST (223) (2007 release) against non-redundant (NR) database (2008 release) with E-value cutoff 0.0001 (2) Generate Hidden Markov Model (HMM) using (374) with the sequence profile from the previous step; the secondary structure for HHSearch is predicted by (375) (3) Generate a sequence alignment between each sequence and the best PDB template found using PSIBLAST and HHSearch (4) Generate 500 models for each sequence using MODELLER-9 (5) Select the best model using DOPE scoring function provided by MODDELLER-9.

COBRA: An integrated approach to *de novo* prediction of protein binding sites. Our novel approach is applicable to any protein sequence for which a comparative model can be built. It leverages numerical features obtained from a sequence–based prediction method, and a structure-based prediction method applied to a comparative model rather than to a native protein structure. Specifically, we develop and compare four new classifiers that integrate results of the sequence-based and structure-based classifiers, adjusting their relative contribution for the different regions of comparative model, depending on the regions' quality. Unfortunately, most of the state-of-the-art sequence- and structure-based methods are either unavailable or presented only as web-servers (241). Thus, we have also designed two sequence-based and two structure-based protein binding site predictors of heteromeric and homomeric binding sites, correspondingly, using feature-based supervised learning classifiers and following the feature description from the two top-performing binding site prediction methods (241). In total, this comprehensive study resulted in training, evaluation, and comparative analysis of 8 machine learning classifiers.

The Sequence-based Binding Residue Annotation (SeqBRA) methods are feature-based classifiers whose features are obtained from the protein sequence using a sequential sliding window approach. For a target residue, a sequential sliding window of size 9 is defined as a sequentially continuous segment of protein sequence, with the target residue in the middle (the fifth position), with four residues before and four residues after the target residue. The size of the sliding windows is selected from the analysis of performance of the state-of-the-art current methods that use a similar sliding-window approach to generate features (241). The window is moved, one residue at a time, from N- to C-terminus, generating a feature vector for a new target residue at each step. For each candidate residue and its sliding window, a 10-feature vector is determined, with the first nine features corresponding to the residue type of each position of the sliding window. Each of the nine features is encoded as a standard 20-bit binary vector representing 20 residue types. The last feature represents the length of the protein. Because of the sliding window design, binding residues occurring in the first four (N-terminal) and the last four (C-terminal) positions of the protein sequence cannot be predicted. To resolve this, we add four ‘decoy’ residues before the first residue and after the last residue in the protein sequence. Each of these residues contributes ‘NULL’ values to the feature vector.

Similar to SeqBRA methods, the Structure-based Binding Residue Annotation (StrucBRA) classifiers employ a 9-residue sliding window with the candidate residue in the middle of the window. A vector of 27 features is calculated from the sequence and structure information of the residues in the window. The features include: (1) residue type of each position in the window encoded the same as in SeqBRA (9 features), (2) secondary structure of the candidate residue, (3) average hydrophobicity, (4) average accessible surface area (ASA), (5) average relative ASA, (6) average backbone ASA, (7) average relative backbone ASA, (8) average backbone ASA, (9) average side chain ASA, (10) average relative side chain ASA, (11) average relative non-polar ASA, (12) average polar ASA, (13) average relative polar ASA, (14) average depth index, (15) average protrusion index, (16) minimal protrusion index, (17) maximal protrusion index, (18) maximal depth index, and (19) length of the sequence. The secondary structure is calculated using DSSP package (376) and features (3)–(19) are determined using PSAIA software (377).

Both SeqBRA and StrucBRA methods leverage the random forest supervised classifier implemented in a scikit-learn library (335). Random forest classifiers have been consistently among the top performing methods for a number of bioinformatics tasks (378, 379). When generating the Random Forest models, the numbers of estimators (number of trees) are 200 for the heteromeric binding site classifier and 250 for the homomeric classifier. In both approaches, a feature vector is labeled positive if the candidate residue belongs to a protein binding site and negative otherwise.

Finally, two COmparative model Binding Residue Annotation (COBRA) methods were developed and compared, each integrating sequence- and structure-based predictions made on comparative models. Similar to the previously described methods, each of these methods is trained on two separate datasets of binding residues, one coming from heteromeric interaction dataset and another from homomeric interaction dataset, resulting in a total of four classifiers. The first method, $\text{COBRA}_{\text{LOG}}$, trains a simple L1-regularized logistic regression classifier implemented in LPS-v2.2 (380), with no further post-processing. The classifier has been widely used for a number of supervised learning and feature selection tasks and has demonstrated good generalization performance (381, 382). Here we consider the classifier with a regularization term defined as L_1 norm constraint on the vector of parameters. The constraint is introduced as a part of the optimization problem at the training stage, with the goal to avoid over-fitting. The second method, COBRA_{RF}, trains a more advanced, random forest, classifier implemented using the same parameters and package as for SeqBRA and StrucBRA classifiers. For both $\text{COBRA}_{\text{LOG}}$ and COBRA_{RF}, each residue is represented as a 39-dimensional feature vector. These features are grouped into four categories and are calculated using the linear and spatial sliding windows.

The first category of features consists of nine variables, each variable encoding the results of the sequence-based binding site prediction for each of the nine residues from the sequential sliding window of the candidate residue. Similarly, the second category of features includes nine variables that encode the results of the structure-based binding site prediction of nine residues from the spatial sliding window containing eight closest neighbors of the candidate residue using Euclidian distance from a neighbor to the candidate residue. Those neighbor residues are sorted based on their proximity to the candidate residue. The third category represents another set of nine variables that are the DOPE scores of the nine residues from the structural sliding window in the same order (241). The fourth category includes only one variable corresponding to the solvent accessibility of the target residue calculated using NACCESS software applied to the comparative model [66]. While our approach can take as input predictions from an arbitrary sequence-based and structure-based binding site prediction methods, here we selected the best performing methods, according to our evaluation.

Spatial clustering of predicted binding site residues. The results of residue classification are post-processed by spatial clustering of the predicted binding residues and filtering the outliers. For clustering, we use a density-based clustering algorithm DBSCAN (241). Being among the top performing clustering algorithms, DBSCAN also has several properties that make it suitable for our task. First, during the clustering procedure it defines the outliers, the data points that do not belong to any cluster. These outliers are structurally segregated and are unlikely to be a part of a protein binding site. Therefore, they are removed from the set of predicted binding residues. Second, the input for DBSCAN is the distance matrix, which allows one to use a custom distance measure. Here, we use the Minkowski distance defined between the PDB coordinates $\left(\left[x^i, y^i, z^i\right], \left[x^j, y^j, z^j\right]\right)$ of the two closest heavy atoms for each pair of predicted binding residues, *i* and *j*, calculated as:

$$D_{i,j} = \left[\left(x^i - x^j\right)^4 + \left(y^i - y^j\right)^4 + \left(z^i - z^j\right)^4\right]^{1/4}.$$

Two parameters of DBSCAN, *min_samples* and *eps*, defining the compactness of the clusters and minimum number of data points populating one cluster, are selected that optimize the performance of the clustering method during the training procedure (see next section for more detail).

Assessment of the approaches. The performance of binding site residue annotation methods is evaluated using 10-fold cross validation independently on homomeric and heteromeric interaction datasets. Eight classifiers are evaluated: (1) SeqBRA using protein sequences as input,

(2) StrucBRA using protein native structures as input, (3) StrucBRA using protein comparative models as input, (4) COBRA$_{LOG}$ optimized on f-measure, (5) COBRA$_{LOG}$ optimized on balanced accuracy, (6) COBRA$_{RF}$ optimized on f-measure, (7) COBRA$_{RF}$ optimized on balanced accuracy, and (8) COBRA$_{RF}$ optimized on f-measure and post-processed using residue clustering (all COBRA classifiers were also tested using protein comparative models as input).

The data sets are evenly split into 10 folds and use each subset of eight folds for training, one-fold for optimization (to get the probability threshold) and the other fold—for testing. During evaluation we also study how the prediction quality depends on the quality of comparative models tested in one StrucBRA and all COBRA methods. For that purpose, the comparative models are assigned to one of eight bins based on the sequence identity between the target sequence and template structure. In total, 138 protein chains and their comparative models are assigned to each bin in the heterogenic interaction dataset, and 241—to each bin in the homomeric interaction dataset. To avoid the training bias, datasets for training the sequence-based and structure-based classifiers are different from the dataset used to train, optimize, and test COBRA, since predictions of the former methods are included to the input feature vectors for the latter methods. Once the evaluation is completed, the probability thresholds for SeqBRA and StrucBRA methods used for new predictions is defined as an average of the 10 thresholds obtained from the 10-fold cross validation. As result, SeqBRA probability thresholds for homo- and heterometic protein binding sites are 0.19 and 0.18, and StrucBRA thresholds are 0.27 and 0.28, correspondingly.

In total, seven measures are used to optimize and evaluate the performance of the binding site prediction methods, accuracy (*Ac*), precision (*Pr*), recall (*Re*), specificity (*Sp*), balanced accuracy (*BA*), f-measure (*F*), and Matthews correlation coefficient (*MCC*):

$$Ac = \frac{TP + TN}{TP + FP + TN + \ FN}; Pr = \frac{TP}{TP\ x\ FP}; Re = \frac{TP}{TP + FN}; Sp = \frac{TN}{FP + TN};$$

$$BA = \sqrt{Sp\ x\ Re};\ F = \frac{2\ x\ Pr\ x\ Re}{Pr + Re}\ ; MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positive, and *FN* is the number of false negatives.

Overview of the large-scale annotation of cancer somatic mutations with protein binding. In a large-scale application of the protein binding site prediction method, we have annotated and analyzed SNVs from eight different cancer types. In this application, we apply StrucBRA for PDB with known experimental structures and COBRA_{RF} optimized on f-measure with residue clustering for the proteins whose structures were modeled. The annotation pipeline includes five basic stages. First, the SNVs, the corresponding genes and their protein products are collected from TCGA repository (241). Second, for each protein we attempt to retrieve one or several structures or structural models. Third, COBRA-RF is applied to those protein sequences with at least one structure or models. Fourth, a SNVs is annotated as a binding site SNV (bSNV) if the corresponding residue belongs to the predicted protein binding site. Last, all annotated SNVs are further analyzed with respect to the enriched function, potential role in cancer, and relationship with patient-specific clinical parameters.

Data collection and preprocessing. The processed TCGA dataset includes in total 6,188 genes and 10,900 SNVs. Among those genes, 1,259 genes have protein products that are covered by the experimentally solved structures retrieved from the Protein Data Bank (PDB) (241), either fully or partially. These structurally annotated protein regions carry 2,242 SNVs in total. However, during the structural characterization of a protein carrying an SNV, the sequence information corresponding to the PDB structure (and stored as a part of the corresponding PDB entry) comes from a source different than the source of the protein product sequence for the same gene stored in TCGA. As a result, changes between the two reference sequences (*e.g.* due to different alternative splicing isoforms) may result in an incorrect mutated residue match on the structure. Indeed, out of 1,259 genes, 188 genes with 407 SNVs have two mismatched reference sequences each. These sequences are then matched using a pairwise sequence alignment (241). However, after applying such protocol to 407 SNVs on the genes with inconsistent sequences, there are still 62 SNVS for which we cannot identify the position on the corresponding sequence by the sequence alignment. As a result, only 345 SNVs of the 407 SNVs with inconsistent sequence information have been structurally annotated. In total, 1,878 SNVs of 1,259 genes have been annotated in the experimentally structurally characterized regions.

In addition, 2,029 genes could be structurally characterized through comparative modeling. The structurally modeled regions contain 3,583 SNVs. The comparative models were obtained using a MODELLER-based automated computational pipeline ModPipe (372, 383).

Enrichment of somatic SNVs on protein binding sites. The set of genes containing at least one somatic SNV and at least one binding site region is selected for calculating the enrichment of mutations on the protein binding sites. Each sequence is divided into regions of two types: (1) binding sites region and (2) outside region. If somatic mutations are enriched on the protein binding sites, we expect their frequency to be significantly higher than the frequency of mutations on the outside regions. The fraction of mutations expected by chance in each region, f_E, is calculated by adding the total sequence length of each region in all proteins, and dividing it by the length of all proteins combined. The number of observed mutations, f_O, in each region over all proteins is also added together and divided by the total number of mutations. The odds ratio, θ, is calculated using these expected and observed fractions: $q = \frac{f_O(1- f_O)}{f_E(1- f_E)}$. Standard error, SE_θ, and Z-score, Z, are calculated using the log odds ratios:

$$SE_q = OR \times SE_{\log q}, \quad Z = \frac{\log(q)}{SE_{\log q}}, \text{ where}$$

$$SE_{\log q} = \sqrt{\frac{1}{n_{SNV\_REG}} + \frac{1}{n_{SNV\_TOT} - n_{SNV\_REG}} + \frac{1}{n_{RES\_REG}} + \frac{1}{n_{RES\_TOT} - n_{RES\_REG}}}, \text{ and}$$

where n_{SNV_REG} and n_{RES_REG} are the numbers of bSNVs and all residues in a region of one of the two types, and n_{SNV_TOT} and n_{RES_TOT} are the total numbers of bSNVs and all residues, respectively.

Enrichment of known "cancer genes" in genes carrying somatic bSNVs. We next determine if the set of genes each carrying at least one somatic bSNV is enriched with the known cancer-associated genes. The dataset of cancer genes includes 522 genes and is collected from the Cancer Gene Census, an ongoing effort to catalogue those genes for which mutations have been causally implicated in cancer (241). The same gene list was used in a recent work on annotation SNVs from the same TCGA dataset with phosphorylation sites (241). The enrichment is calculated using the hypergeometric test:

$$(X \geq n_0) = \sum_{k=n0}^{N} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Here, *N* is the number of total genes in human genome, *K* is the number of all cancer genes, *n* is the number of the bSNV carrying genes; n_0 is the number of the cancer genes observed among the bSNV carrying gens.

Functional Enrichment Analysis. We next determine if bSNV-carrying genes are enriched or depleted with specific functions. To do so, four gene lists were created (1) genes with at least one bSNVs, (2) genes with no bSNVs, (3) genes with at least one pSNVs, and (4) genes with no pSNVs. For the four gene lists, the gene enrichment analysis is performed using the ToppFun method from the ToppGene Suite (241). This analysis allows for multiple comparisons to be performed within the same computational framework from multiple database sources. It explores gene associations with GO molecular function, GO biological process, GO cellular component, pathways, human and/or mouse phenotypes, protein domain content, protein interactions, miRNA interactions, cytoband, transcription factor binding sites, and other factors. Parameters used for the program include an FDR correction factor (241) with a p-value cutoff of 0.05.

Cancer Survival Analysis. The comprehensive patient clinical data, available only for ovarian cancer, is downloaded from TCGA. The clinically relevant attributes are extracted for each sample ID the bSNV annotation analysis has been done for. To enable comparison with a recently performed survival analysis for somatic mutations involved in phosphorylation, the survival analysis is performed using the same protocol (241). Specifically, we implement Kaplan-Meier estimate of survival, where the input for the estimate is the time of survival, patients are stratified using two separate protocols, and whether the patient is survived within the observed timeframe. Patients are stratified and tested for a significant survival difference either from the SNV or gene perspective. The SNV perspective stratifies a patience as based on presence or absence of a bSNV for the patient. The gene perspective selects for only patients that have a SNV within that gene. Then patients are stratified by either a presence or absence of a bSNV. This implementation tests whether there is a difference between the survival curves using the *G-rho* family of tests by Harrington and Fleming (241) with weights on the death using Kaplan-Meier estimate of survival and log-rank or Mantel-Haenszel test. The analysis is implemented in R using Olsurv, data.table, stringr, and Exact packages. The same analysis is repeated on the individual gene basis.

Clinical Significance. The significance of the difference of ratios between bSNVs and non-bSNVs with respect to each of the clinically relevant attributes, including ethnicity, tumor_status, tumor_grade, residual_tumor, residual_disease_largest_nodule, vascular_invasion_indicator, lymphovascular_invasion_indicator, karnofsky_score, ecog_score, performance_status_timing, radiation_treatment_adjuvant, pharmaceutical_tx_adjuvant, treatment_outcome_first_course, new_tumor_event_dx_indicator, initial_pathologic_diagnosis, anatomic_neoplasm_subdivision, clinical_M, clinical_N, clinical_T, clinical_stage, days_to_initial_pathologic_diagnosis, extranodal_involvement, histological_type, icd_10, icd_o_3_histology, icd_o_3_site, pathologic_M, pathologic_N, pathologic_T, pathologic_stage, tissue_source_site, and tumor_tissue_site, is tested using the Fisher's exact t-test implemented in R. This simple implementation uses the hypergeometric distribution with three possible alternatives for a 2 by 2 contingency table. The null hypothesis is there is no statistically significant difference between the two ratios. The three alternatives tested with respect to this hypothesis include (1) the odds ratio is greater than 1, (2) the odds ratio is less than 1, and (3) equal to 1. Testing each of the three alternatives is important, since the effect of a bSNV on a clinical variable is generally unknown.

Cancer gene centered PPI network reconstruction. To gain further insights of bSNV-targeted genes we constructed a PPI network using BioGRID (241) dataset (version *BIOGRID-ORGANISM-Homo_sapiens-3.2.105*). The complete PPI network retrieved from BioGRID includes 15,854 proteins and 133,755 PPIs. A subnetwork, containing all proteins with at least one bSNV or pSNV was created and analyzed using this work's SNV annotation and the previously published dataset (241), yielding 4,994 proteins and 21,958 interactions. Based on the newly constructed subnetwork, we then extracted cancer-specific subnetworks for each of the cancer types.

3.2.5 Discussion

The main objective of this work is to investigate whether cancer SNVs have a prevalence for a protein's binding site. The rationale of this objective stems from two main observations. The first observation is a major challenge in genomics is the functional interpretation of genomic mutations (273). Functional mutations typically affect protein residues, however customary mutation evaluation methods focus mainly on gene and protein specific information such as comparative protein analysis for evolutionary or disease specific conservation (384). However, we can gain additional insight about functional mutations by considering particular protein sites related to molecular interactions such as phosphorylation (358). The second observation stems from disruption and/or dysregulation of protein interactions may be fundamental to allowing cancer progression (360). The typical phenotypic observation of cancer is uncontrollable growth, failure of immune cells to control, and lack of ability to undergo apoptosis. Signaling pathways or lack thereof all involve protein interactions thus modifications of the normal protein interaction allow for cancer to progress.

We achieved our objective by first designing a new *de novo* protein binding site prediction method called **Co**mparative **B**inding **R**egion **A**nnotator (COBRA). We decided to design a new method due to the desire to increase the number of protein binding sites that can be annotated without losing prediction accuracy. The main types of protein binding site prediction methods rely on a protein's sequence or structure. Sequence based methods are capable of assessing any protein sequence thus having a high coverage, but suffer from lower accuracy when compared to structure based methods, which are limited to resolved protein structures. Our novel approach relies on the ability for any protein sequence for which a comparative model can be leveraged. The approach integrates information from a sequence based prediction method and a structure based, but applied to a comparative model rather than the natural protein structure thus allowing an increase to the number of proteins that can be assessed. In summary, we trained, evaluated and compared 8 machine learning classifiers on an experimentally validated protein binding site dataset with the best result of 0.37 precision, 0.60 recall, 0.74 specificity, 72% accuracy, 0.29 MCC, and 0.46 F-measure for COBRA.

We achieved our main objective by applying COBRA for a large-scale annotation of SNVs across eight different cancer types from the cancer genome atlas (TCGA) for their overlap with protein binding sites (bSNVs). bSNVs were assessed for their enrichment of known cancer drivers, functional enrichment, patient survivability, clinical significance, and protein-protein interaction (PPI) network. Furthermore, to provide context to the patterns discovered we compared bSNVs to SNVs that affect the phosphorylation site (pSNVs) using the same data (241).

Our first observation was on average a high number of bSNVs (41%), but in contrast for pSNVs (9.5%). Furthermore, the particular bSNV mutation locations were typically repeated. For example, KRAS, a well-known cancer associated gene, had the highest number of bSNVs associated (98.4%). However, despite the 186 bSNVs, they only correlated with 5 mutational positions on the gene. This is an interesting observation because the binding site regions are typically less than 5% of the total length of a protein yet account for a large portion of SNVs.

Our second observation was for both protein function mutation types, bSNVs and pSNVs affect similar genes across multiple pathways with 40 – 80% overlap dependent on the cancer type. We found that the overall number of genes containing bSNV mutations whose protein products are a part of a PPI cancer network across each cancer type is significantly larger than the number of genes with mapped pSNVs. Intriguingly, the bSNVs and pSNVs have a complementary cumulative effect on the cancer interactome: for each cancer type considered in this work, the dataset of genes that are a part of the interactome and that carry bSNVs have little to no overlap with genes carrying pSNVs.

Our third main observation came from any functional enrichment of genes associated with at least one bSNV or not (non-bSNV) for top enriched GO terms and functional pathways. bSNV genes were uniquely enriched for GO terms associated with DNA repair and extracellular receptor activity in contrast to intra-cellular binding activity for non-bSNV genes. A hallmark of cancer is a failure of DNA repair. What is unexpected is the split for bSNVs genes to affect cellular signaling, but non-bSNVs affect protein binding within the cell as protein binding occurs throughout the cell as well as extracellular. A possible hypothesis for this observation is the type of protein binding being targeted: temporary versus permanent, which will require additional analysis to explore. Another unexpected finding, which is related to extra-cellular binding activity, was bSNVs are enriched with olfactory receptors (OR). From one perspective this is unusual as OR were labeled due to their implicated role in the sense of smell that were thought as being unrelated to cancer (364). However, recent studies are increasingly demonstrating a role of OR in cancer as part of a failure of cell to cell communication (241, 366-369). Lastly, we considered 12 specific pathways related to cell survival, cell fate, and genome maintenance typically thought to be associated with cancer (241). Interestingly, only 4 out of the 12 cancer pathways were associated with genes with bSNVs. Taking these findings together suggests bSNVs play a role in cell to cell communication within cancer progression.

Finally, we analyzed the clinical survival with bSNV. When we considered patients with or without a bSNV there was not statistically significance, but a general negative trend. This caused us to rephrase the analysis from the perspective of patients with genes enriched with bSNVs, which led to 40 bSNV genes having a negative impact on survival. These genes functions are consistent with our enriched terms as associated with olfactory receptor genes, which are thought to be involved in cell to cell communication (241, 366-369). This provides a direct link with the phenotypic observations of cancer of the disruption to cell to cell communication that is essential to allow cancerous cells to circumvent the immune system, allow metastasis, and a failure to apoptosis.

This analysis was conducted from two different perspectives: global and individual scale. The global scale refers to a patient is considered to have a bSNV if there is at least one gene with a bSNV. Whereas, the individual scale considers patients with the presence or absence of a specific bSNV gene. The survival data on a global scale for both bSNV and pSNV are not significant when considering at the global scale. However, the global trend is that bSNVs has a negative impact on survival whereas pSNVs has a positive impact. However, when considering presence of bSNV on the individual gene level there are 40 out of 2080 genes that have a statistically significant difference. Performing gene enrichment analysis on these 40 genes leads only to olfactory receptors as an enriched function. However, this only accounts for 4 out of 40 genes present.

There is not an enriched function or pathway associated when these 4 olfactory receptor genes are excluded from the analysis. Furthermore, there is only one cancer driver (*Akd1*) associated with these 40 genes. All 40 genes have a negative impact on survival.

In summary, we achieve our goal of assessing the prevalence of cancer SNVs with protein binding sites by using our novel *de novo* protein binding prediction tool COBRA, we were able to assess 23% more proteins then typically would be allowed without losing prediction accuracy. From our analysis of over 8 cancer types, we conclude that 1) bSNVs to be highly prevalent as they are close to 10 times more bSNVs than expected, 2) enriched functions for genes affected by bSNVs are related to cell to cell communication, and 3) interestingly, genes enriched with bSNVs were associated with a negative impact on cancer survival. These findings give support to other analyses which suggest that disruption and dysregulation of protein interactions may be fundamental for cancer to progress (360).

3.3 Survey of Alternative Splicing Impact On Protein Interaction Landscape In Human

3.3.1 Abstract

Alternative splicing of mRNA precursors is known to expand the diversity of protein isoforms from the majority of genes in humans. What is unknown is whether there are patterns associated with the functional aspects of a protein from alternative splicing selection. This makes assessment of the functional impact of alternative splicing difficult to access. In order to create a systematic structural analysis of this landscape, we analyzed protein domains and binding sites with their propensity with alternative splicing types and RNA-Seq expression patterns. To achieve this analysis, we incorporated transcript data across six different databases then selected transcripts from the agreement of at least four databases, which produce a protein product from humans. These transcripts had binding sites identified by two methodologies and protein domains identified through SUPERFAMILY. Furthermore, alternative splicing types such as intron retention and exon skipping were integrated with RNA-Seq expression analysis across multiple developmental time points and tissue types to assess for patterns of alternative splicing selection. In summary, we analyzed 16,682 genes with 218,222 isoforms in human, which cause on average 83% of binding sites and a range of 0 - 85% of domains, N-terminus, or C-terminus to be rearranged as a result of alternative splicing. Furthermore, we propose that genes be quantified with an 'alternative splicing impact factor' to summarize the impact alternative splicing has on possible protein function for future studies of influence. Our results suggest that while alternative splicing can drastically remove or alter (>50%) important components of a protein such as domains, N-terminus, or C-terminus it maintains the majority of binding sites (>73%) demonstrating alternative spliced proteins play a functional role.

3.3.2 Introduction

Upon sequencing of the first draft of the human genome, a paradox was highlighted where the number of genes did not correspond to the number of proteins, which are considered the main players in the molecular realm (385, 386). Given that humans are regarded as being more complex then organisms such as bacteria, worms, or fruit flies, it is considered a 'remarkable' observation the number of protein coding genes do not correlate with complexity (241, 387). This has led to a scientific revolution in expanding annotation of the genome from a focus on individual protein-coding genes to an increasing more complete view of the complex reality such as alternative splicing, pseudogenes, and noncoding RNAs (388).

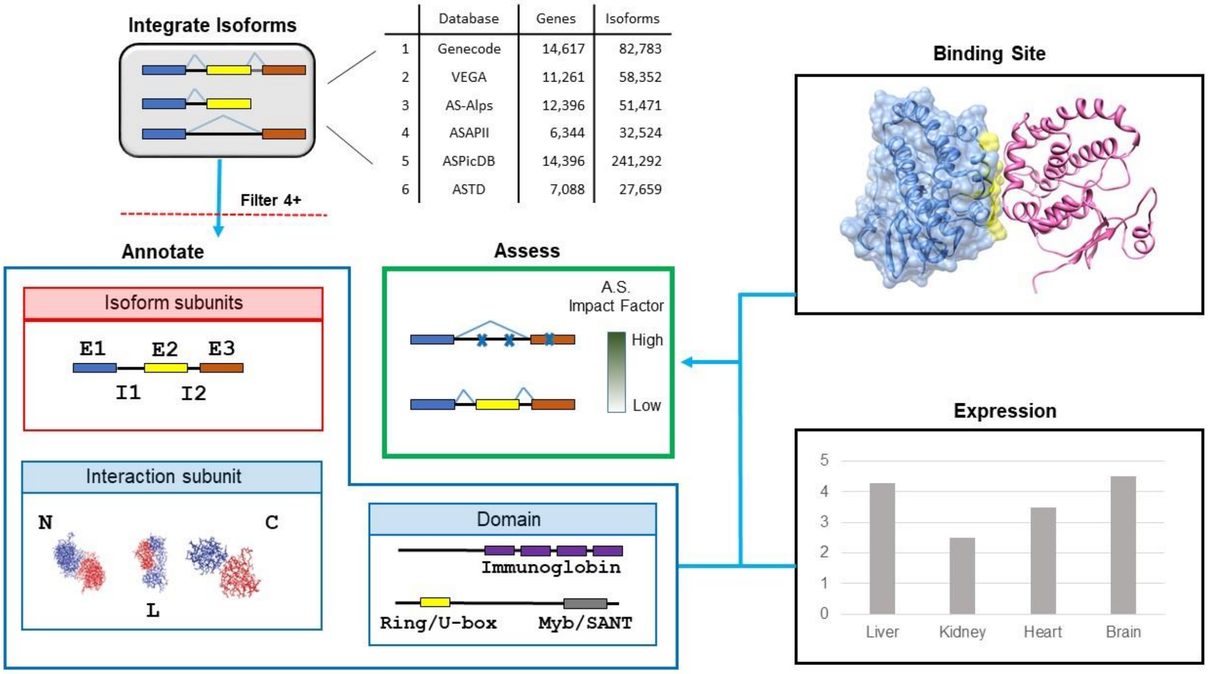
Alternative splicing has largely been marketed as being the main player for expanding the diversity of protein isoforms for the majority of genes in human. While this mechanism was discovered in the late 1970's in viruses (42, 43, 389), the frequency it occurs in humans was not demonstrated until much later due to the advent of next generation sequencing (51, 52). The observation that a single gene can produce on average 10 different 'versions' or isoforms for ~95% of human genes resulting in ~70% altering a protein function presents an interesting solution to the paradox (41, 47, 84, 390) (52, 391). While it is still an ongoing debate the frequency alternative splicing affects protein function (69, 71), it has been experimentally demonstrated on many case-by-case bases (74) and a large scale protein-protein interaction (84) that alternative splicing can 'rewire' its protein function and interaction (83).

What is essentially unknown is how proteomic interaction complexity is related to alternative splicing selection. This makes assessment of the functional impact of alternative splicing on proteins difficult to access. There have been some systematic studies into the functional role of alternative protein isoforms, but focused on the functions they can or cannot perform relative to their ‘reference’ counterpart typically the longest known protein transcript associated for the particular gene (84, 392, 393). Understanding the patterns of alternative spliced proteins affecting the interaction network has been defined as a critical step in studying complex genetic disorders (74, 82, 394, 395). Previously patterns such as protein domain architecture with alternative splicing have been studied limitedly (396-398), but have suggested that alternative splicing largely does not affect the protein domain architecture.

This work’s goal is to achieve a systematic analysis alternative splicing modifies on protein interaction. Specifically, we analyzed protein domains, N-terminus, C-terminus, linker region, and binding sites with alternative splicing and RNA-Seq expression patterns. We integrated six databases and filtered for protein producing transcripts, which had agreement across four databased to achieve a comprehensive alternative spliced database (COMP-AS). These transcripts had binding sites identified by two methodologies, protein domains identified through SUPERFAMILY, and used a comprehensive database of macromolecular interactions (DOMMINO v. 2.0 (316)) that includes the interactions between protein domains, interdomain linkers, N- and C-terminal regions, and protein peptides. Furthermore, alternative splicing types such as intron retention and exon skipping were integrated with RNA-Seq expression analysis across multiple developmental time points and tissue types to assess for patterns of alternative splicing selection. In summary, we analyzed 16,682 genes with 218,222 isoforms in human, which cause ‘rewiring’ on average 83% of binding sites and a range of 0 - 85% of domains, N-terminus, or C-terminus to be rearranged as a result of alternative splicing. Furthermore, we propose that genes be quantified with an ‘alternative splicing impact factor’ to summarize the impact alternative splicing has on possible protein function for future studies of influence. Our results suggest that while alternative splicing can drastically remove or alter (>50%) important components of a protein such as domains, N-terminus, or C-terminus it maintains the majority of binding sites (>73-97%) demonstrating alternative spliced proteins have the potential to play a functional role.

3.3.3 Results

This work’s goal is to achieve a systematic analysis alternative splicing modifies on protein interaction. Specifically, we analyzed protein domains, N-terminus, C-terminus, linker region, and binding sites with alternative splicing and RNA-Seq expression patterns. We integrated six databases and filtered for protein producing transcripts, which had agreement across four databased to achieve a comprehensive alternative spliced database (COMP-AS). These transcripts had binding sites identified by two methodologies, protein domains identified through SUPERFAMILY, and used a comprehensive database of macromolecular interactions (DOMMINO v. 2.0 (316)) that includes the interactions between protein domains, interdomain linkers, N- and C-terminal regions, and protein peptides. Furthermore, alternative splicing types such as intron retention and exon skipping were integrated with RNA-Seq expression analysis across multiple developmental time points and tissue types to assess for patterns of alternative splicing selection. In summary, we analyzed 16,682 genes with 218,222 isoforms in human, which cause ‘rewiring’ on average 83% of binding sites and a range of 0 - 85% of domains, N-terminus, or C-terminus to be rearranged as a result of alternative splicing. Furthermore, we propose that genes be quantified with an ‘alternative splicing impact factor’ to summarize the impact alternative splicing has on possible protein function for future studies of influence. Our results suggest that while alternative splicing can drastically remove or alter (>50%) important components of a protein such as domains, N-terminus, or C-terminus it maintains the majority of binding sites (>73-97%) demonstrating alternative spliced proteins have the potential to play a functional role.



	Database	Genes	Isoforms
1	Genecode	14,617	82,783
2	VEGA	11,261	58,352
3	AS-Alps	12,396	51,471
4	ASAPII	6,344	32,524
5	ASPicDB	14,396	241,292
6	ASTD	7,088	27,659

Figure 3.3.3.1. Methodology for Alternative Splicing Protein Functionality Assessment. Alternative Spliced Protein Isoforms were integrated from six databased then filtered for isoforms with 100% agreement between at least four databases. These isoforms were annotated for interaction subunit, protein domain, and isoform. Binding sites were then annotated. RNA-Seq was then conducted across 120 tissue types, 40 age groups, and both genders to give context to the prevalence of alternative spliced transcripts.

Furthermore, alternative splicing types such as intron retention and exon skipping were integrated with RNA-Seq expression analysis across multiple developmental time points and tissue types to assess for patterns of alternative splicing selection. In summary, as a result of alternative splicing 'rewiring' can drastically remove or alter (>50%) important components of a protein such as domains, N-terminus, or C-terminus, but maintains the majority of binding sites (>73-97%) demonstrating alternative spliced proteins have the potential to play a functional role.

Dataset Statistics

After filtering for at least four database agreement across Genecode, VEGA, AS-Alps, ASAPII, ASPicDB, and ASTD, resulting in 16,682 genes with 218,222 alternative spliced protein isoforms. Upon annotation these isoforms had, 160,776 (74%) domains, 164,223 (75%) C terminus, 151,091 (69%) N-terminus, and 82,937 (38%) linker regions. On average 83% of binding sites are modified across alternative spliced isoforms with the range of (73-97%).

Example

Alternative splicing has been well established as having a tissue context change. What is not well reflected is how well this correlates with protein domain structure and binding site changes. Using the example of HMBOX (Fig 3.3.3.2) it is highlighted that protein-binding site changes do not correlate with protein expression or alternative splicing rearrangement. Primarily the C terminal region are the regions that are heavily modified for this particular gene, which interestingly does not have a large impact on the protein domains. Additionally, when considering the expression changes the transcripts with heavy modifications by C terminal, protein domain, or binding sites do not correlate with transcript usage. This may suggest that alternative splicing is used more to exclude functional uses rather than expand functions.

Summary of Trends

Expanding this observation across the entire alternative splicing landscape alternative splicing represents wide versatility. The first analysis looks at the top 50 protein families in terms of what does alternative splicing do (Figure 3.3.3.1). The most prevalent change is no change in terms of distributions across all possible modifications. However, when modifications do affect it is by deletions rather then mutations. The distribution across protein families is quite variable suggesting further exploration is required. Expanding the analysis by looking at just type of modifications by protein structure region (U,N,L,D,C) demonstrates that in fact deletions, alterations, and single nucleotide changes are the most common across all types. Changes are generally in the majority within protein domains.

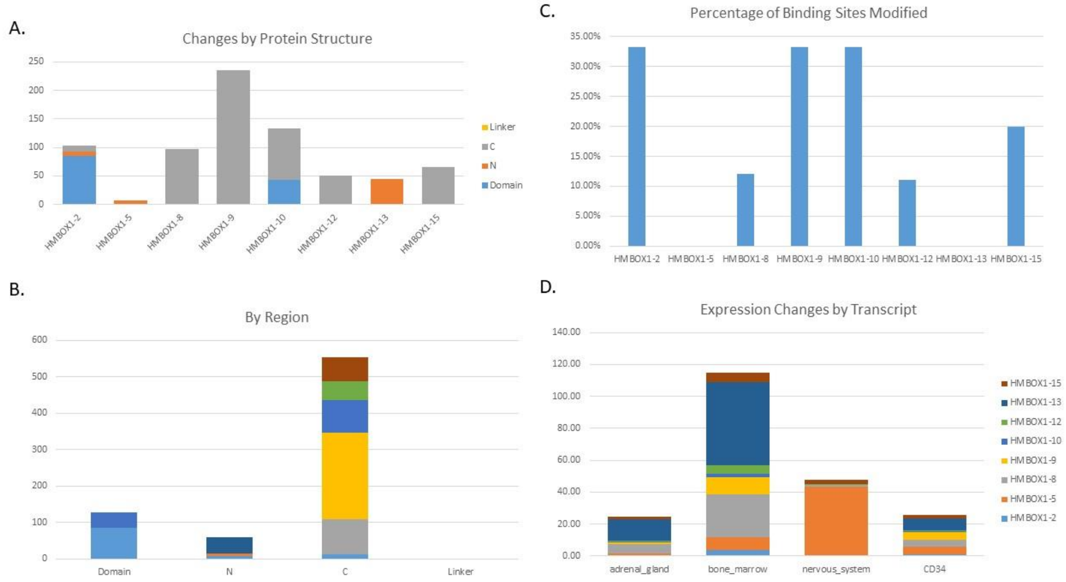


Figure 3.3.3.2 HMBOX Alternative Splicing (AS) Diversity Changes. A) 8 transcripts for HMBOX with the number of nucleotide changes by alternative spliced transcript based on protein structure definitions. B) Same information as A except information is plotted by protein structure. C) Percent of protein binding sites modified as a result of AS. D) 4 represented tissues demonstrating the veracity of expression changes, which does not correlate with protein structure modification.

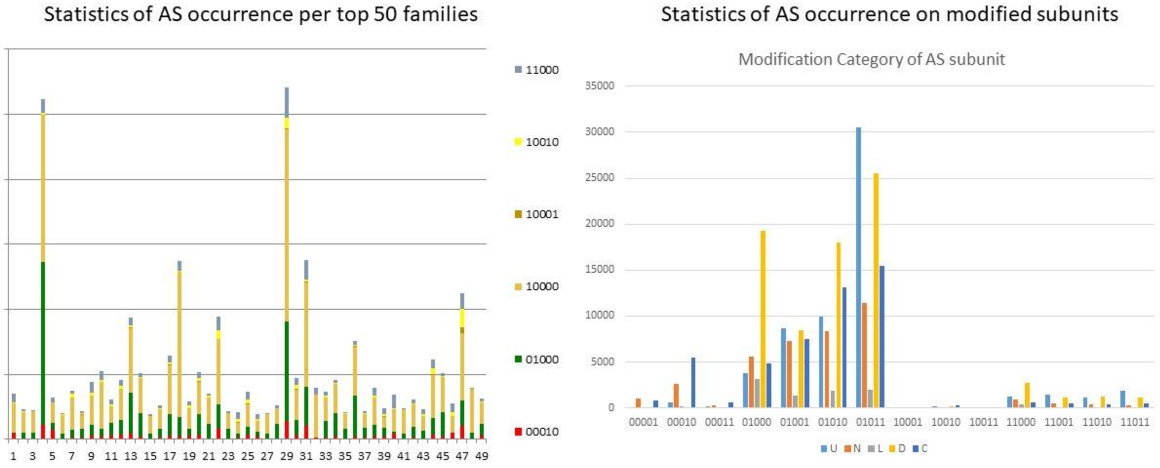


Figure 3.3.3.3. Global Statistics of Alternative Splicing Occurrence. Alteration patterns: unchanged (u), deleted (d), inserted (i), altered (a), single mutated (s) are code by (u, d, i, a, s). (e.g. u=1, d=1, i=0, a=1, s=1 results in a pattern 11011) A) Alterations by top 50 families. B) Alterations by protein structure subunits.

Summary of Protein Binding Site Modifications

Analyzing alternative splicing effect on protein binding sites (Figures 3.3.3.4) demonstrates that the majority of changes are modified with the protein domain having the highest changes. However, focusing on the extent of modification demonstrates that majority of changes are less then 20% of the entire protein binding site.

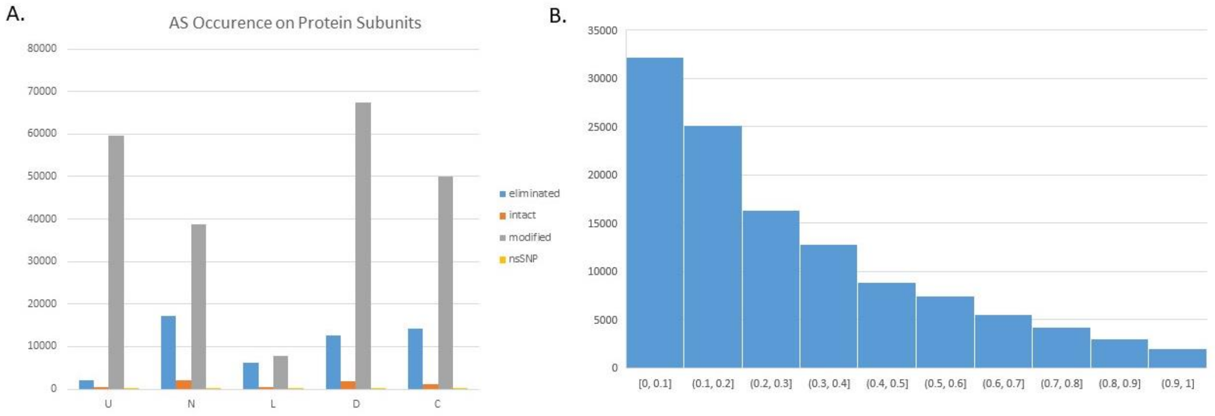


Figure 3.3.3.4 Global Summary of Alternative Splicing Effect on Protein Binding Sites. A) Reflects the frequency of changes based on definition of eliminated, left intact, modified, and SNPs. B) The percentage of protein binding sites changes implicating that most binding sites are less than 20% modified.

3.3.4 Discussion

This work systemically assesses the impact alternative splicing has on protein interaction in human. Specifically, we integrated six alternative splice databases then filtered to only those isoforms that are protein coding and have 100% agreement between at least four databases to achieve a high consensus (COMP-AS). This resulted in 16,682 genes with 218,222 alternative splice protein isoforms. These isoforms had binding sites identified by two methodologies, protein domains identified through SUPERFAMILY, and used a comprehensive database of macromolecular interactions (DOMMINO v. 2.0 (316)) that includes the interactions between protein domains, interdomain linkers, N- and C-terminal regions, and protein peptides.

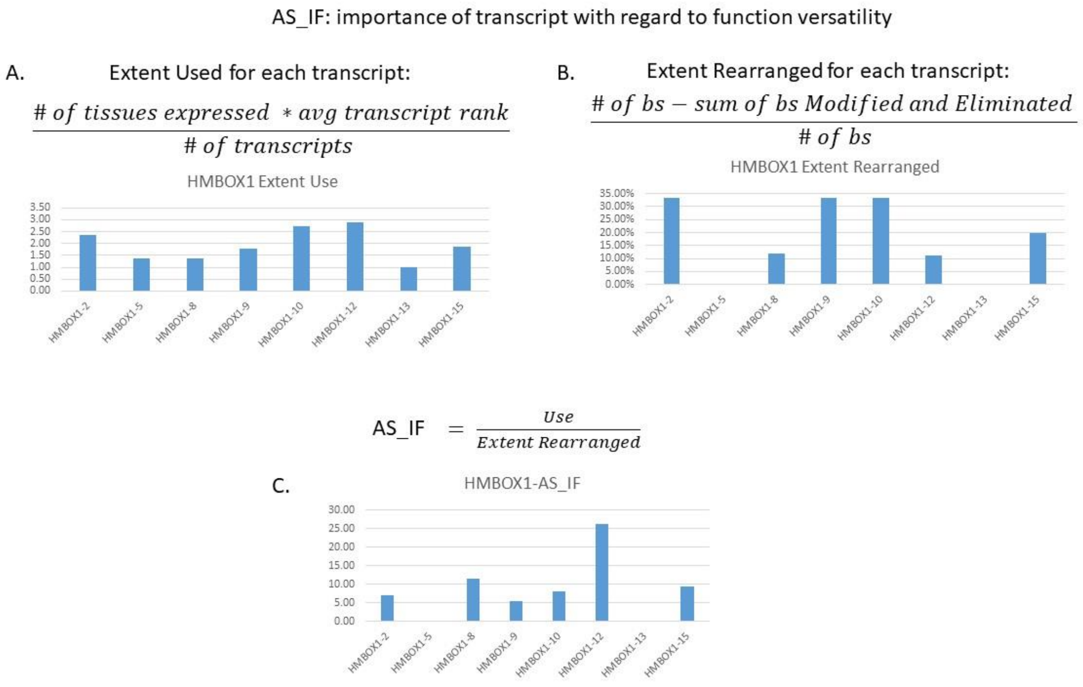


Figure 3.3.3.5. Alternative Splicing Impact Factor. In order to quantify the functional versality of a protein as a result of alternative splicing the metric Alteriatve Splicing Impact Factor (AS-IF) was developed. AS-IF reflects the veracity of protein function change as a reflection of alternative splicing rearrangement. There are two components that make up AS-IF. A) One component attempts to quantify the extent a transcript is used based on the frequency of expression across tissues as well as modified by the transcript expression rank. B) The second component is to quantify the extent of protein rearrangement of transcripts as a reflection of binding sites deletions and modifications. C) These two components divided to emphasis the use of a transcript by the extent alternative splicing rearranges the protein. As an example, the three bar plots calculated AS_IF on HMBOX1.

To reflect the impact of modification alternative splicing has on protein structure, we propose an Alternative Splicing Impact Factor (ASIF) (Figure 3.3.3.5). In order to attempt to quantify how impactful ASIF is on function changes, a single value would be useful. Our proposed value represents both the expression modifications in addition to protein binding site changes. Combining both of these components allows for each transcript to be represented by the impact AS as on the functional changes.

3.3.5 Methods

The methodology used within this study integrated six databases, which profile alternative splicing, filtered for an agreement of at least four databases, annotated protein binding sites, domains, N-terminus, C-terminus, and linker regions. Using this data, alternative splicing patterns were analyzed in context of alternative splicing types and RNA-Seq expression patterns across 140 tissue types, 40 age points, and both genders. Our approach made use of two protein binding site methodologies, SUPERFAMILY protein annotation, and DOMMINO a macromolecular interaction database.

Dataset Integration

The six databases integrated were VEGA (399) , AS-Alps (400), ASAPII (401) , ASPicDB (402), and ASTD (403). These databases incorporate manual, automated, computationally predicted and experimentally derived alternative splicing annotation for human. Due to this, we only used alternative spliced annotation for proteins that were in 100% agreement across four or more databases (COMP-AS).

Binding Site Prediction

Two protein binding site methodologies were used. The first relies on template based. The second relies on PSI-BLAST, which makes use of SCOP protein domain annotation in order to extend the coverage. Since the first methodology relies on template based it is more reliable. Due to this, when there is overlap between the methods the first method is preferable.

Protein Domain and Subunit Annotation

Using DOMMINO v2.0 (316) was used to annotate the protein domains, N-terminus, C-terminus, and linker regions (protein regions between domina, N, and C).

RNA-Seq Quantification

To determine the expression values for alternative splicing, we quantified across 120 tissue types, 40 age groups, and both genders. The RNA-Seq samples came from publically available Expression Atlas (404), which includes 1,170 normal samples for human. Kallisto (405) using default settings was used to quantify the samples against COMP-AS as the sequences to quantify against.

CHAPTER 4: Conclusions and Future Directions

The main aim of this dissertation is to leverage omics data to expand the value and understanding of alternative splicing (A.S.). Specifically, four types of ‘omics’ data were utilized across the six projects described here: transcriptomics, proteomics, genomics, and epigenomics. The context used to describe alternative splicing was usually within a complex genetic disorder such as diabetes or cancer. However, there were projects described within a ‘normal’ as well as crop production. While each chapter and subsection have a wide variety of conclusions specific to the problem context, there are many general trends that can be highlighted from this dissertation.

When alternative splicing produces RNAs for translation, these can be considered functionally unique proteins. The usual perspective is to consider a gene as the main base unit of hereditary. What is increasingly being demonstrated is through alternative splicing of a gene, the range of proteins produced can be considered functionally unique as they interact within their molecular environment. This suggests that rather than organizing a systemic viewpoint around a gene, it should rather be done at individual transcripts and their protein products. Changing this viewpoint to a transcript centric perspective, most likely will change how molecular pathways are considered to act within the normal or abnormal context (406).

Alternative splicing creates more noncoding RNA then coding RNA. This is both an observation and a suggestion for future directions. It was observed quite often by myself when conducting the various analyses that far more transcripts were removed then preserved upon filtering for protein. It has been hypothesized that these transcripts are not functionally useful while some may be used as regulators such as miRNA through the DICER complex (407). This is suggested as the expression level of these transcripts are low and often below a given filter threshold. This raises a question why they are expressed at all as there are mechanisms for silencing (408). While the general opinion trend suggests an evolutionary artifact or in the process of being degraded, the volume I have observed (sometimes up to 50% of all unique transcripts expressed) I would hypothesize that there is something functionally relevant to this observation. I would hypothesize that these noncoding RNAs have far more functions then currently realized. This is based on noncoding RNAs such as miRNA, lncRNA, and circRNA have only been recently discovered and analyzed using the same 'omic' type data discussed in this dissertation from the past decade (409-411). I would suggest developing analytical tools such as from Chapter 3.3, that would focus on the noncoding RNA rather than coding.

Alternative splicing needs more system-wide functional analysis. Much that is understood about alternative splicing is on a case-by-case basis rather than via systemic pattern analysis. To this end, a future direction of this dissertation is combining the various projects and development of new tools to systemically assess the effect of alternative splicing regarding coding RNA. Specifically, combining the analysis from chapters 3.1 and 3.3 to extend this analysis to explore what are the functional consequences of alternative splicing from function and interaction perspective. This would extend the understanding into the possible molecular rationale for particular transcripts existence under certain normal or abnormal conditions.

In summary, this work brings together different ‘omics’ data to expand our understanding and promote the use of A.S. Specifically, there are six projects described here which make use of transcriptomics, proteomics, genomics, and epigenomics, which often overlap, on the focus in a couple of complex genetic diseases as well as analyzing a parasite, which affects soybeans.   The projects range from systemically profiling machine learning methods utilizing RNA-Seq based alternative splicing expression data to promote its use, development of a method to predict whether alternative splicing occurs affects its interaction, a systematic analysis across the transcriptome for comparing binding sites and domains with alternative splicing and expression patterns, assessment of single nucleotide variation on protein binding sites, assessment of epigenomics with transcriptomics within the context of acute lymphoblastic leukemia, and looking for patterns of alternative splicing on parasites infecting soybeans.

Appendix A: Supplementary 2.3

A.1 Figures

Figures Appendix A1.1-3. Distribution of Classes for each Dataset. Y-axis is the number of samples. The X-axis lists the classes used for binary and multiclass (MC) classifications. Colors group classes of similar type. Each figure depicts a different dataset for RBM (S9), NCBI (S10), TCGA (S11)

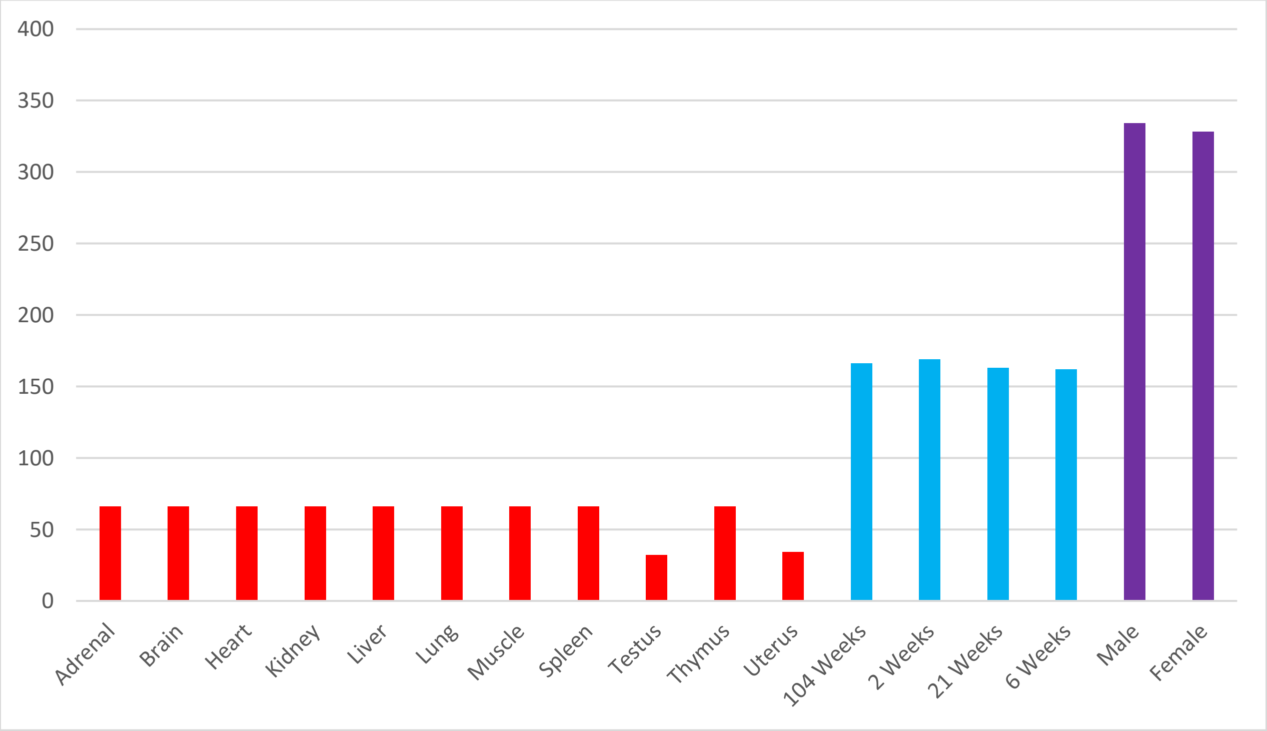


Figure Appendix A1.1- RBM

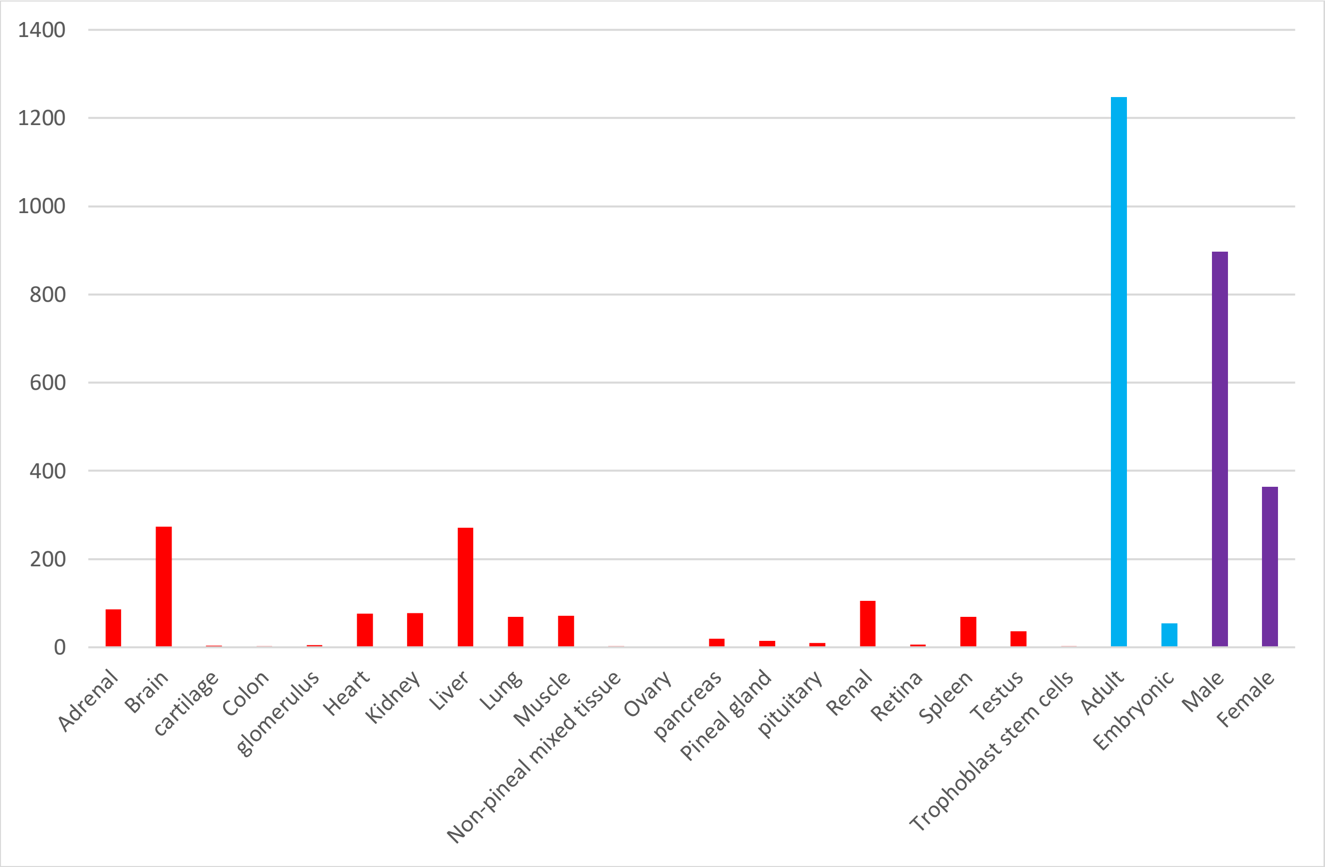


Figure Appendix A1.2-NCBI

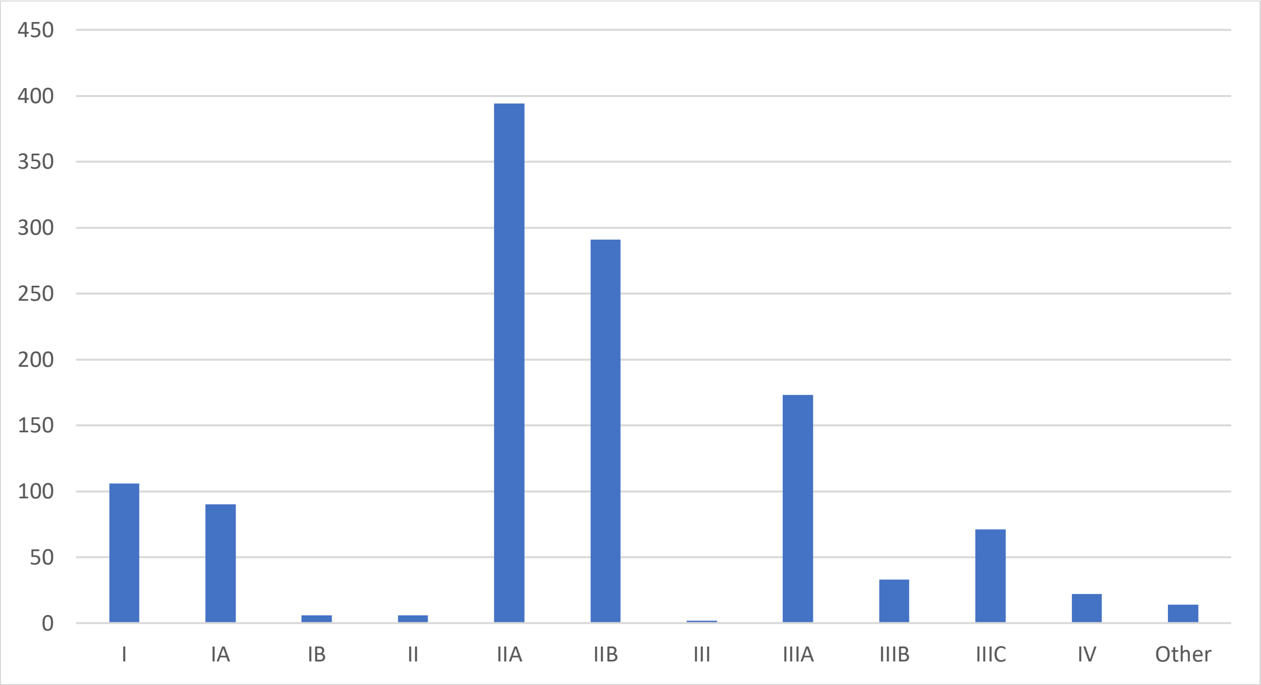


Figure Appendix A1.3- TCGA

Figures Appendix A1.4-7. Comparing the Number of Features Selected During Feature Selection Protocol for Gene vs Isoform Based Classification. The Y-axis is the number of features. The X-axis lists the classes used for binary and multiclass (MC) classifications. Each figure depicts an individual dataset: RBM (S4), NCBI (S5), TCGA-log$_2$ Normalized (S6), and TCGA-Raw Count (S7).

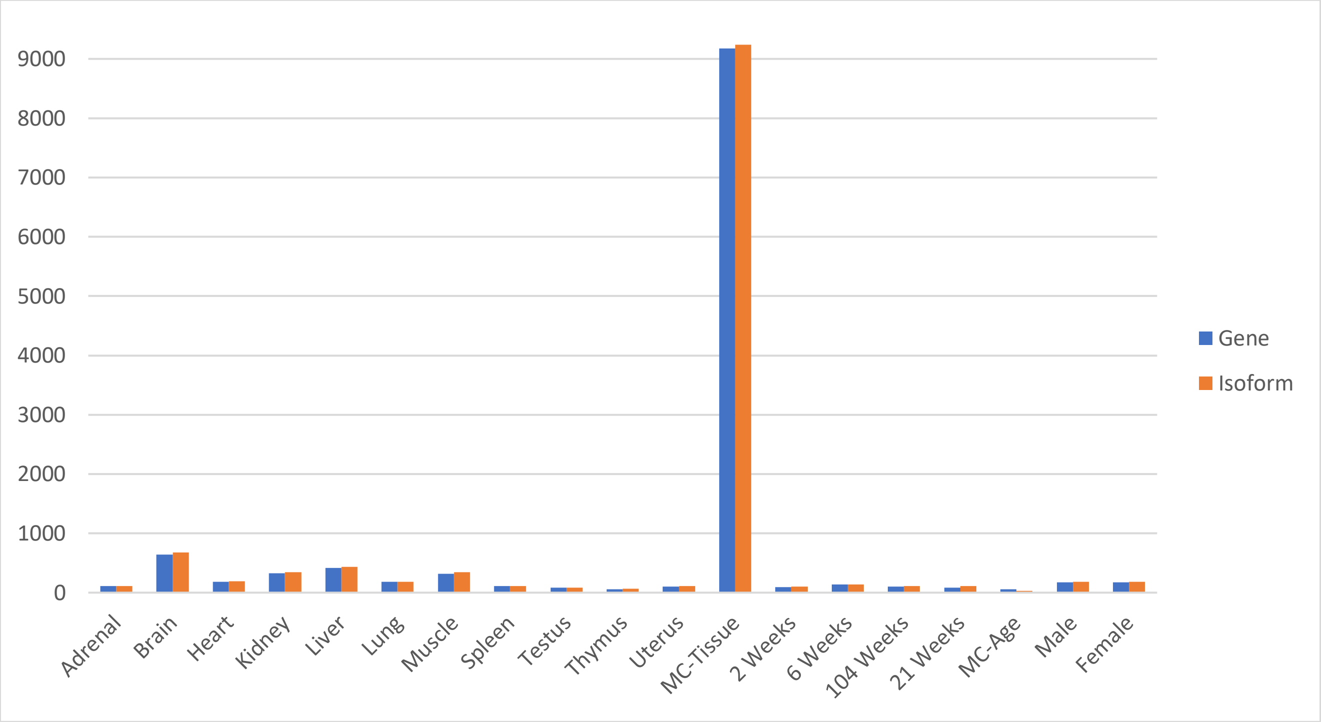


Figure Appendix A1.4-RBM

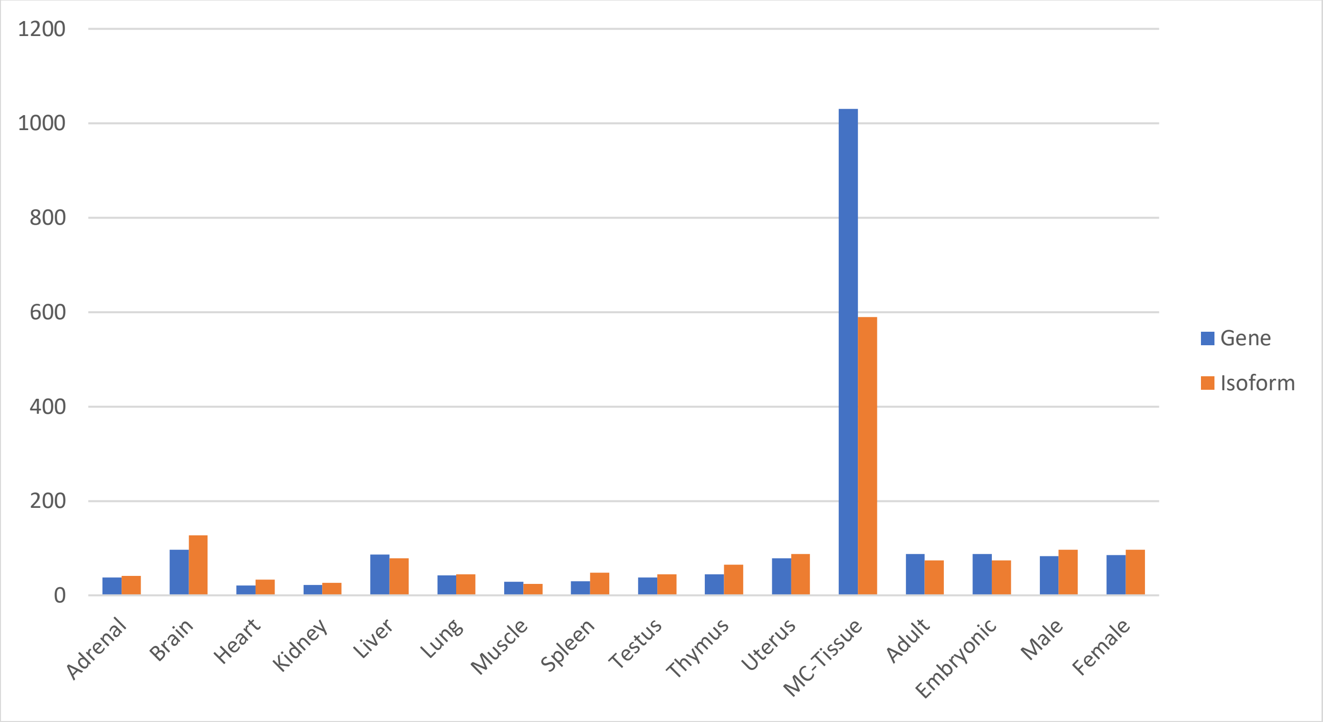


Figure Appendix A1.5-NCBI

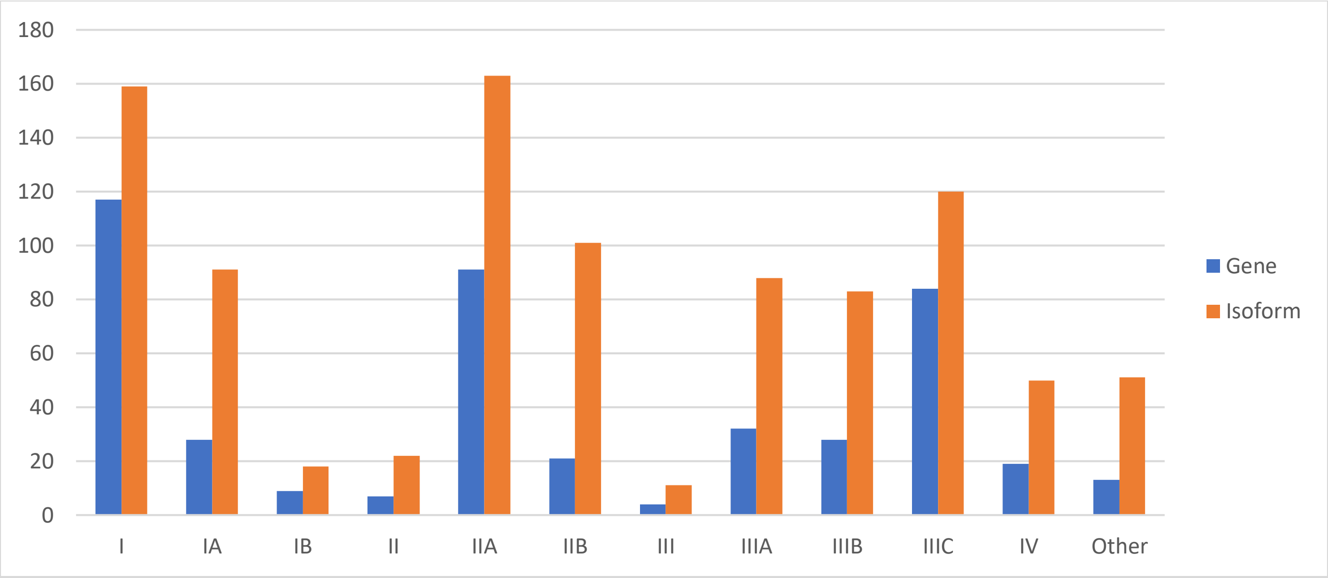


Figure Appendix A1.6- TCGA Log_2 Normalized

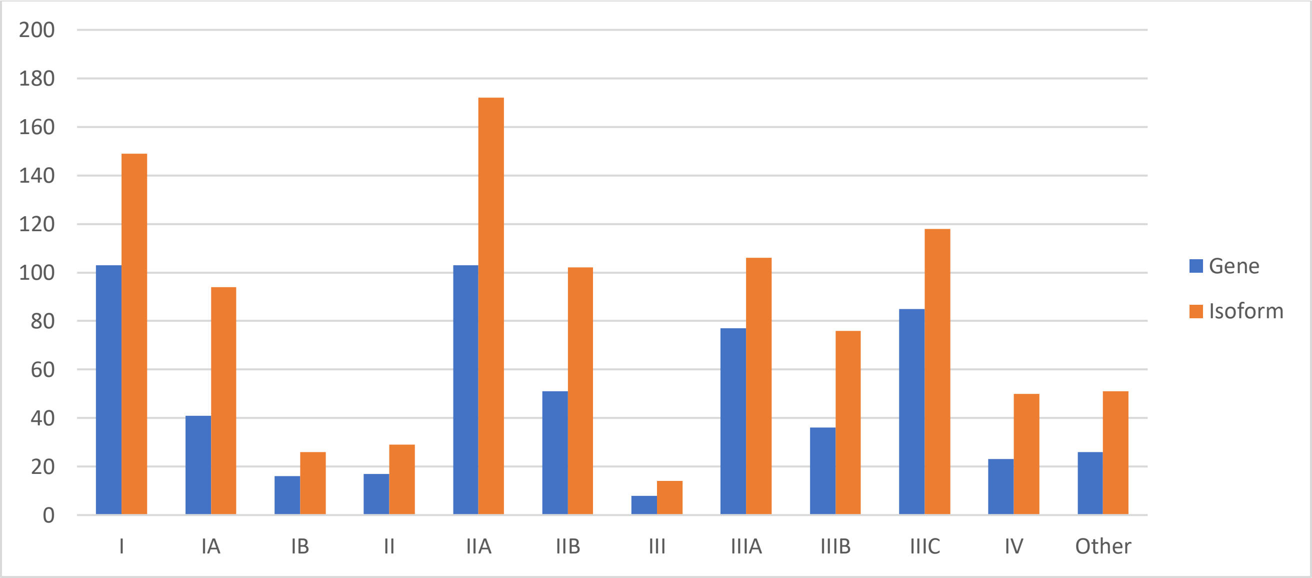


Figure Appendix A1.7 – TCGA Raw Counts

Figures Appendix A1.8–11 Comparing the Number of Features Selected Post-Feature Selection across Normalization Techniques for Gene Features. The Y-axis is the number of features. The X-axis lists the classes used for binary and multiclass (MC) classifications. Each figure depicts a different dataset for RBM (S8), NCBI (S9), TCGA-$\log_2$ Normalized (S10), and TCGA-Raw Count (S11).

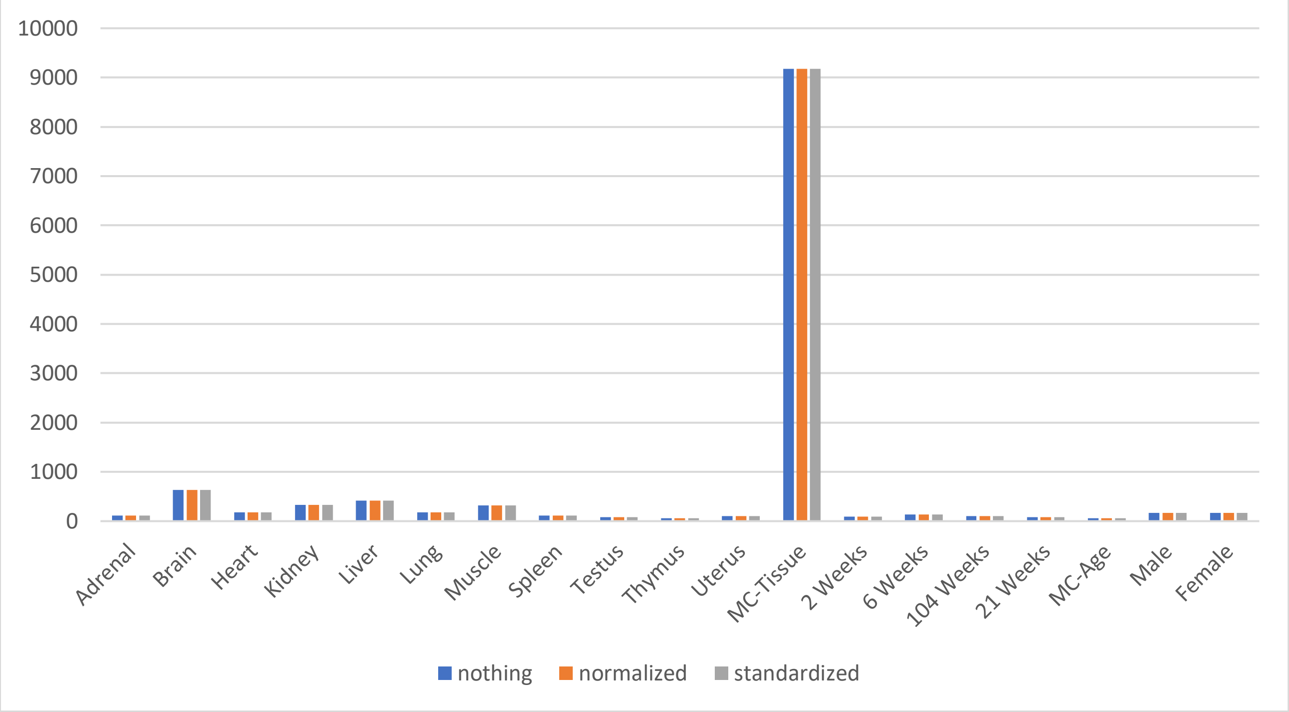


Figure S8 – RBM

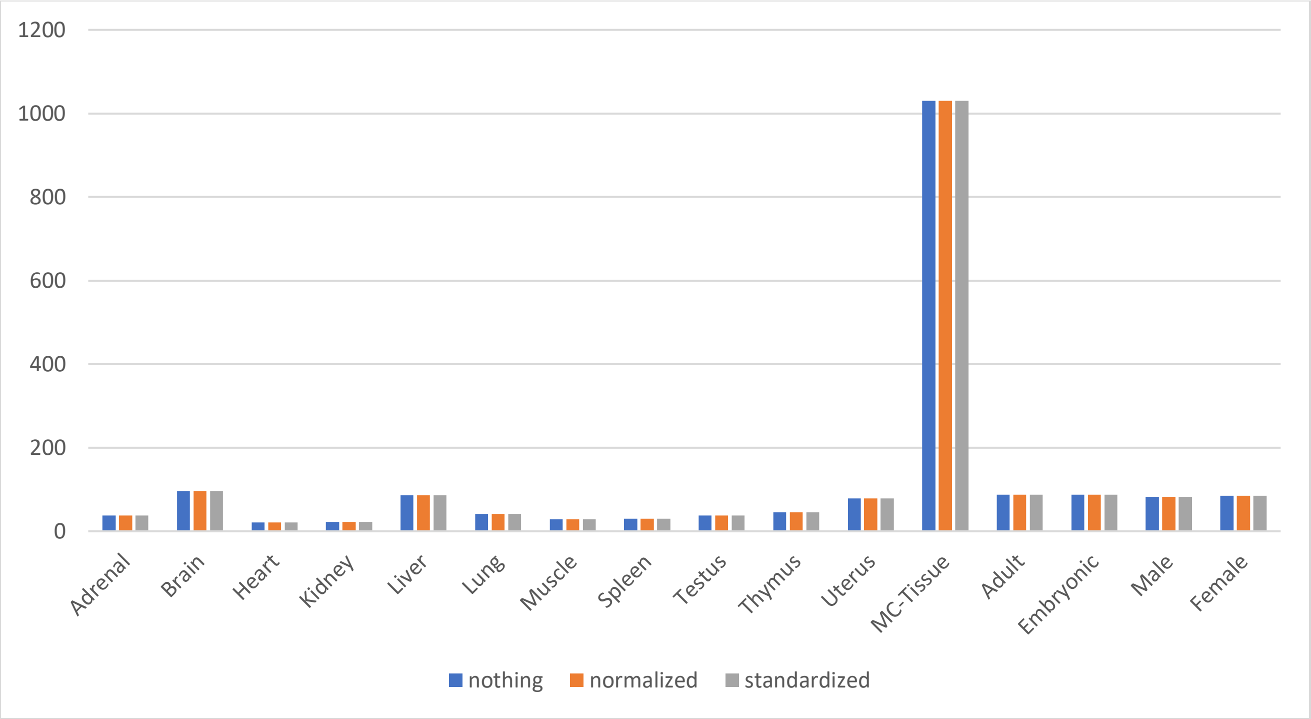


Figure Appendix A1.9-NCBI

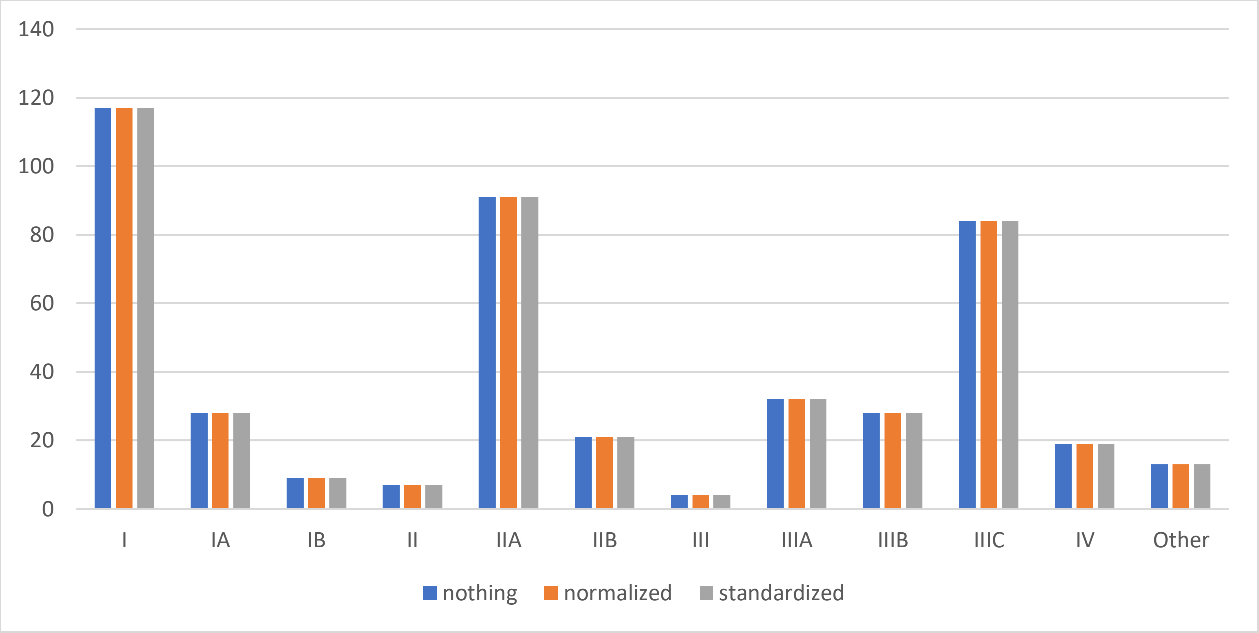


Figure Appendix A1.10-TCGA Log$_2$ Normalization

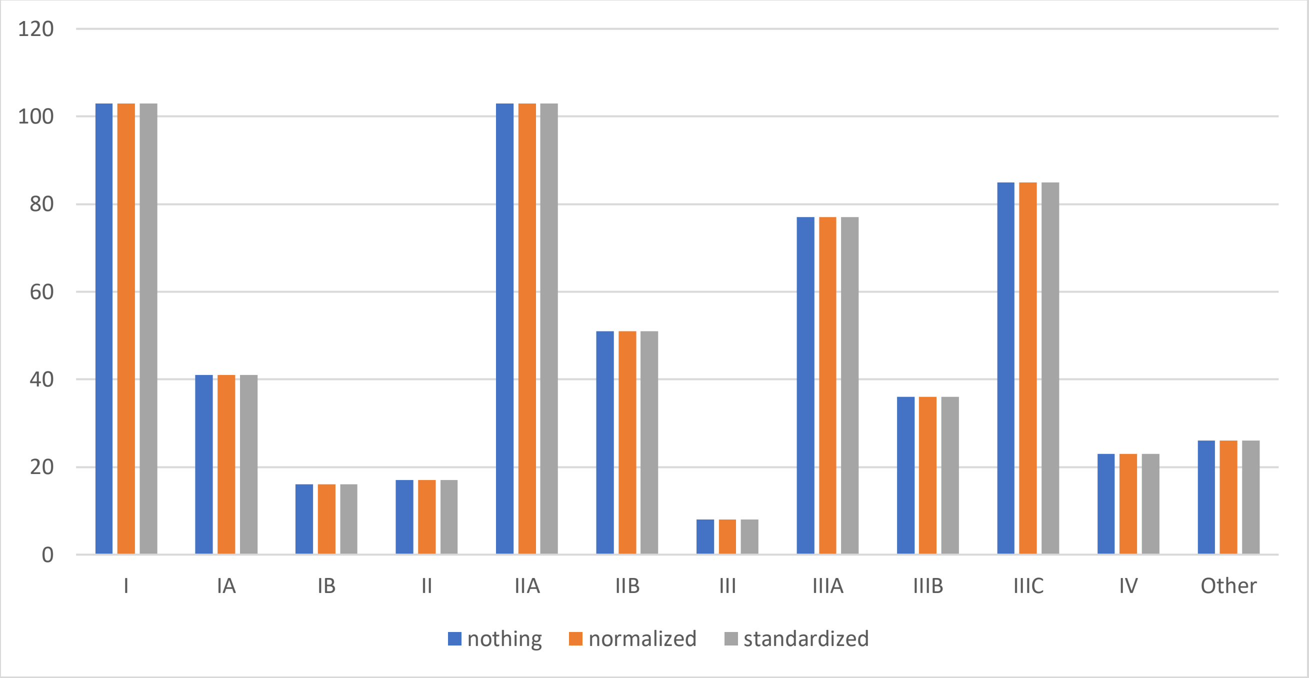


Figure Appendix A1.11-TCGA Raw Count

Figures Appendix A1.12-13. Heat Map representation of *f* -measure standard deviation across the 10-Fold Cross Validation across machine learning methods, classes, datasets, and normalization techniques. For the majority of classification tasks, the standard deviation was less than 1% for both Gene (S12) and Transcript (S13). The top x-axis represents normalization techniques including Nothing (no normalization), Standardized, and Normalized. The bottom x-axis represents the machine learning techniques (DT = Decision Table, J48 = J48 Decision Tree, LR = Linear Regression, NB = Naïve Bayes, RF = Random Forest, SVM = Support Vector Machine). The y-axis represents the classes where MC stands for multiclass. Datasets for each panel are (A) RBM, (B) NCBI, (C) TCGA – $\log_2$ normalized counts, (D) TCGA – raw counts.

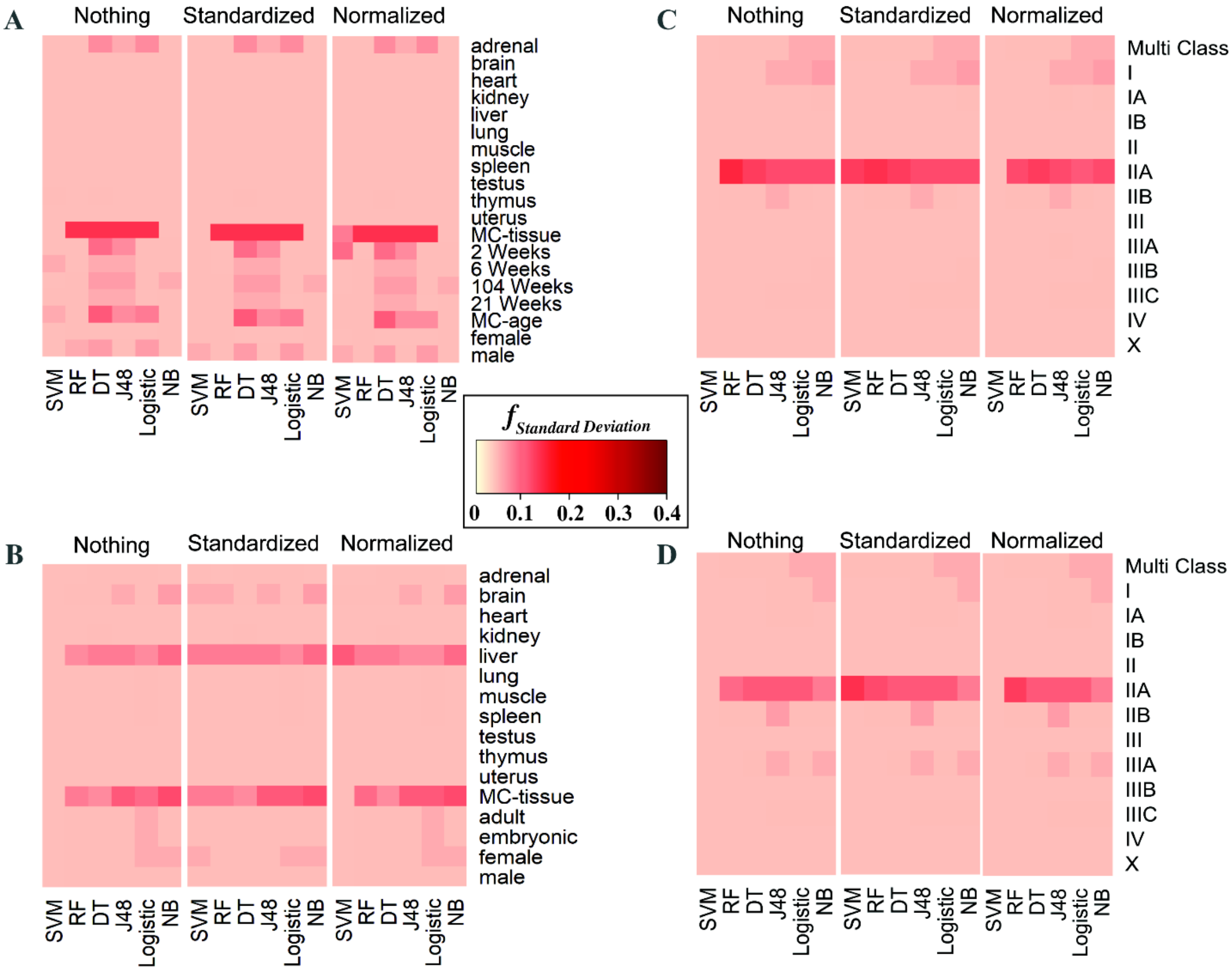


Figure S12-Gene

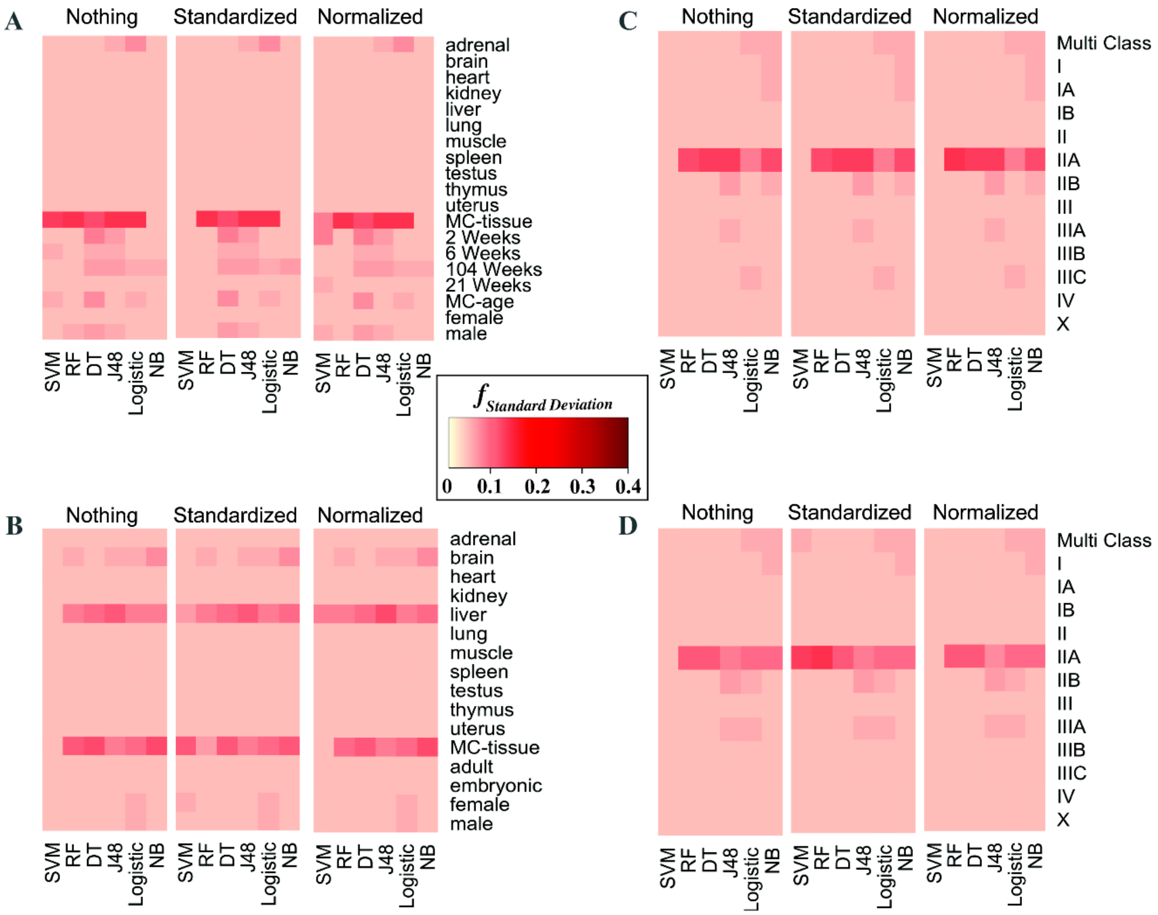


Figure Appendix A1.13- Transcript

A.2 Tables

Table Appendix A2.1- List of all NCBI SRA Project IDs for NCBI Dataset

Accession Number	Number of Samples
SRP037986	662
SRP023266	144
SRP041131	125
SRP039021	116
SRP028932	48
SRP016501	27
SRP036442	24
SRP041119	16
SRP045777	16
SRP021090	14
SRP028515	12
SRP029760	12
SRP041920	12
SRP055430	12
SRP042370	10
SRP021119	8
SRP046247	8
SRP018407	6
SRP044684	6
SRP046248	6
SRP035358	4
SRP041741	4
SRP045117	4
SRP051483	4

SRP009272	2
SRP013262	2
SRP017140	2
SRP029980	2
SRP047494	1

Table Appendix A2.2. Initial number of features, and the average number of features after feature selection procedure across different classification problems

	Initial		After Feature Selection	
	Gene	Transcript	Gene	Transcript
RBM	25,538	29,130	659	671
NCBI	10,711	17,506	119	97
TCGA-Raw	20,524	73,592	49	82
TCGA-$\log_2$ Normalized	20,524	73,592	38	80

Table Appendix A2.3- Number of Features Selected: Binary classification (average) and Multiclass classification

	Binary-Average		Multiclass	
	Gene	Transcript	Gene	Transcript
RBM-Tissue	230	242	9173	9237
RBM-Age	102	114	60	27
NCBI-Tissue	48	97	1030	589
TCGA-Raw-Cancer Stage	38	80	20	73
TCGA-$\log_2$ Normalized-Cancer Stage	49	82	20	73

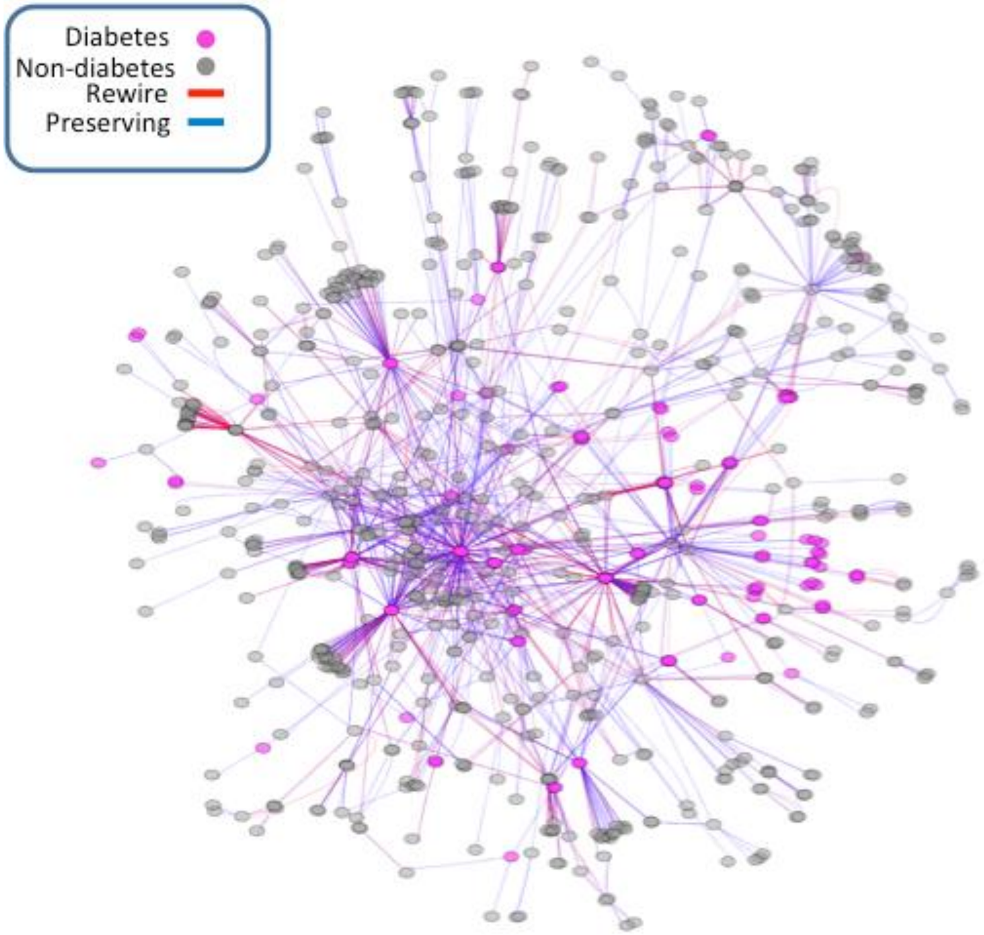
Appendix B: Supplementary 3.1

Supplementary Data

B1. Figures

```
$X_{train}$ = train set data samples
$Y_{train}$ = train set labels
$U$ = unlabeled data samples
$N$ = number of elements to add to train set
$T = False$
$\varepsilon$ = threshold value
$X_{best} = X_{train}$
$RF_{model}$ ←run Random Forest classifier on train set $(X_{train}, Y_{train})$
$Fscore_{best}$ ← F-score of $RF_{model}$ based on 10-fold CV on $(X_{train}, Y_{train})$
$RF_{best\ model} \leftarrow RF_{model}$
**while not** $T$ **do:**
    $Y_U$ ← Classification result of $RF_{model}$on $U$
    $\left(U_{ordered}, Y_{U_{ordered}}\right)$ ←order $U$ along with corresponding $Y_U$ according to the probability of $Y_{U_i}$ to be correct label for sample $U_i$ in descending order
    $(K, Y_K)$ ← first $N$ elements from set $\left(U_{ordered}, Y_{U_{ordered}}\right)$
    $(X_{new}, Y_{new})$ ← merge $(X_{train}, Y_{train})$ and $(K, Y_K)$
    $RF_{newmodel}$ ← run Random Forest classifier on train set $(X_{new}, Y_{new})$
    $Fscore_{new}$ ← F-score of $RF_{newmodel}$ based on 10-fold CV on $(X_{train}, Y_{train})$
    **if** $Fscore_{new} > Fscore_{best}$:
        $Fscore_{best} \leftarrow Fscore_{new}$
            $RF_{best\ model} \leftarrow RF_{newmodel}$
        $(X_{train}, Y_{train}) \leftarrow (X_{new}, Y_{new})$
    **remove** $K$ from $U$
    **if** $|Fscore_{old} - Fscore_{new}| < \varepsilon$**:**
        $T$ ←True
```

Appendix B1.1. A pseudocode of iterative self-learning random forest algorithm used in AS-IN Tool.



Appendix B1.2. Alternative Splicing Focused Protein–Protein Interaction Network. This dataset (D1) was developed using yeast-two hybrid to explore the effect alternative splicing has on protein-protein interactions. We used this dataset to train and test the supervised machine learning models. As an example, we highlighted nodes corresponding to the genes associated with diabetes (magenta); all other nodes are colored in grey. Edges that correspond to the PPIs that are experimentally confirmed to be eliminated by an alternatively spliced isoform are colored red; those edges that correspond to PPIs that are not affected by the alternative isoforms are colored blue.

B.2 Tables

Appendix B2.1. List of features used in a feature-based machine learning classifier. The list includes 3 main groups – biochemical features, statistical potentials and AS-related “delta” features. Related Proteins column indicates which proteins among the reference isoform A_1, its interacting partner B, and alternatively spliced isoform A_2 are involved. Two stand-alone proteins indicate that features described in the corresponding group were obtained for each protein independently. (X,Y) grouping of proteins indicate that both proteins X and Y are required to obtain the corresponding features. X-Y indicates that the corresponding features reflect the difference between the individual features of proteins X and Y.

Feature Group		**Related Proteins**	**Feature List**
Biochemical		$\mathbf{A_1}$ $\mathbf{B}$	Molecular weight Number of residues Average residue weight Charge Isoelectric point A280 molecular extinction coefficient for reduced and cystine bridges Frequency, Molarity, DayhoffStat for each residue and residue property (Tiny, Small, Aliphatic, Aromatic, Non-polar, Polar, Charged, Basic, and Acidic)
Statistical Potentials		$\mathbf{(A_1,B)}$	3 largest statistical potentials among all combinations of domain from protein and protein
		$\mathbf{A_1}$	2 largest statistical potentials for individual domains of protein
Delta Features	Biochemical	$\mathbf{A_1\text{-}A_2}$	*See feature list for Biochemical*
	Statistical potentials	$\mathbf{(A_1,B)\text{-}(A_2,B)}$	*See feature list for Statistical Potentials*
		$\mathbf{A_1\text{-}A_2}$	
	Sequence Alignment Based	$\mathbf{(A_1,A_2)}$	Length change (ratio) Length change (absolute) N-termini C-termini Maximum alignment gap size Mean alignment gap size Number of alignment gaps Number of large gaps (>=10 bases) Number of small gaps (<10 bases)
	Domain Based	$\mathbf{(A_1,A_2)}$	Domains lost Domains changed Domain or linker

Appendix B2.2. Comparison of alternative splicing-specific machine learning models and general *ab initio* PPI prediction methods. Our top performing machine learning model is the semi-supervised random forest. It has the best scores for each metric except recall. PPI prediction methods fairs poorly for our problem. As both F1-score and MCC are also low our conclusion is that M1, M2 and M3 in its current states are unfit for our problem. Low AUC also suggests that we cannot raise other metrics much by simply varying the probability cutoff threshold.

Algorithm	Feature Selection	Accuracy	Precision	Recall	F1-score	MCC	AUC
Semi-Supervised RF	None	**0.88**	**0.92**	**0.92**	**0.92**	**0.70**	**0.84**
Random Forest	None	0.86	0.90	0.92	0.91	0.65	0.81
SVM-RBF	RFE	0.84	0.87	0.93	0.89	0.55	0.75
SVM-Linear	Lasso	0.82	0.84	0.93	0.88	0.48	0.71
PPI Prediction Methods							
M1		0.58	0.29	0.50	0.36	0	0.51
M2		0.58	0.29	0.50	0.36	0	0.41
M3		0.46	0.50	0.52	0.40	0.05	0.42

REFERENCES

1. Weatherall D, Greenwood B, Chee HL, & Wasi P (2006) Science and technology for disease control: past, present, and future. *Disease control priorities in developing countries* 2:119-138.
2. Bateson W & Mendel G (1913) *Mendel's principles of heredity* (University press).
3. Bateson W & Mendel G (2013) *Mendel's principles of heredity* (Courier Corporation).
4. Avery OT, MacLeod CM, & 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. *Journal of experimental medicine* 79(2):137-158.
5. Watson JD & Crick FH (1953) The structure of DNA. *Cold Spring Harbor symposia on quantitative biology*, (Cold Spring Harbor Laboratory Press), pp 123-131.
6. Crick F (1970) Central dogma of molecular biology. *Nature* 227(5258):561-563.
7. Hamosh A, Scott AF, Amberger JS, Bocchini CA, & McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* 33(suppl 1):D514-D517.
8. McKusick VA (1998) *Mendelian inheritance in man: a catalog of human genes and genetic disorders* (JHU Press).
9. Siegel RL, Miller KD, & Jemal A (2018) Cancer statistics, 2018. *CA: a cancer journal for clinicians* 68(1):7-30.
10. Control CfD & Prevention (2017) National diabetes statistics report, 2017. *Atlanta, GA: Centers for Disease Control and Prevention, US Dept of Health and Human Services*.
11. Cohen JE (1995) How many people can the earth support? *The Sciences* 35(6):18-23.
12. Tubiello FN, *et al.* (2015) The contribution of agriculture, forestry and other land use activities to global warming, 1990–2012. *Global change biology* 21(7):2655-2660.
13. Crow JF (1998) 90 years ago: the beginning of hybrid maize. *Genetics* 148(3):923-928.
14. Smil V (2004) *Enriching the earth: Fritz Haber, Carl Bosch, and the transformation of world food production* (MIT press).
15. Khoury CK, *et al.* (2014) Increasing homogeneity in global food supplies and the implications for food security. *Proc Natl Acad Sci U S A* 111(11):4001-4006.
16. Allen TW, *et al.* (2017) Soybean yield loss estimates due to disease in the United States and Ontario, Canada, from 2010 to 2014. *Plant Health Prog.* 18:19-27.
17. Bebber DP, Holmes T, Smith D, & Gurr SJ (2014) Economic and physical determinants of the global distributions of crop pests and pathogens. *New Phytol* 202(3):901-910.
18. Alexandratos N, *et al.* (2012) World agriculture towards 2030/2050: ESA working paper no. 12–03. *UN Food and Agriculture Organization: Rome*.
19. Hayden EC (2014) Technology: the $1,000 genome. *Nature* 507(7492):294-295.
20. Metzker ML (2009) Sequencing technologies—the next generation. *Nature Reviews Genetics* 11(1):31-46.
21. Bentley DR (2006) Whole-genome re-sequencing. *Current opinion in genetics & development* 16(6):545-552.
22. Bamshad MJ, *et al.* (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics* 12(11):745-755.
23. Ozsolak F & Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 12(2):87-98.

24. Saliba AE, Westermann AJ, Gorski SA, & Vogel J (2014) Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* 42(14):8845-8860.
25. Boguski MS, Arnaout R, & Hill C (2009) Customized care 2020: how medical sequencing and network biology will enable personalized medicine. *F1000 biology reports* 1.
26. McShane LM*, et al.* (2013) Criteria for the use of omics-based predictors in clinical trials. *Nature* 502(7471):317-320.
27. Mason CE, Porter SG, & Smith TM (2014) Characterizing Multi-omic Data in Systems Biology. *Systems Analysis of Human Multigene Disorders*, (Springer), pp 15-38.
28. Domany E (2014) Using High-Throughput Transcriptomic Data for Prognosis: A Critical Overview and Perspectives. *Cancer research* 74(17):4612-4621.
29. Wang WY, Barratt BJ, Clayton DG, & Todd JA (2005) Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics* 6(2):109.
30. Parker LA, GómezSaez N, Lumbreras B, Porta M, & Hernández-Aguado I (2010) Methodological deficits in diagnostic research using '-omics' technologies: evaluation of the QUADOMICS tool and quality of recently published studies. *PloS one* 5(7):e11419.
31. Ferlay J*, et al.* (2015) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer* 136(5).
32. Perrin JM, Bloom SR, & Gortmaker SL (2007) The increase of childhood chronic conditions in the United States. *Jama* 297(24):2755-2759.
33. Eyre-Walker A (2010) Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences* 107(suppl 1):1752-1756.
34. Kilpinen H & Barrett JC (2013) How next-generation sequencing is transforming complex disease genetics. *Trends in Genetics* 29(1):23-30.
35. Goldstein DB*, et al.* (2013) Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics* 14(7):460-470.
36. Mikhail FM (2014) Copy number variations and human genetic disease. *Current opinion in pediatrics* 26(6):646-652.
37. Chen M & Manley JL (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* 10(11):741-754.
38. Singh RK & Cooper TA (2012) Pre-mRNA splicing in disease and therapeutics. *Trends in molecular medicine* 18(8):472-482.
39. Portela A & Esteller M (2010) Epigenetic modifications and human disease. *Nature biotechnology* 28(10):1057-1068.
40. Chen L & Zheng S (2009) Studying alternative splicing regulatory networks through partial correlation analysis. *Genome Biol* 10(1):R3.
41. Baralle FE & Giudice J (2017) Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology* 18(7):437-451.
42. Chow LT, Gelinas RE, Broker TR, & Roberts RJ (1977) An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell* 12(1):1-8.
43. Berget SM, Moore C, & Sharp PA (1977) Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences* 74(8):3171-3175.
44. Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* 103(3):367-370.
45. Pohl M, Bortfeldt RH, Grützmann K, & Schuster S (2013) Alternative splicing of mutually exclusive exons—A review. *Biosystems* 114(1):31-38.

46. Blencowe BJ (2017) The relationship between alternative splicing and proteomic complexity. *Trends in biochemical sciences* 42(6):407-408.

47. Bush SJ, Chen L, Tovar-Corona JM, & Urrutia AO (2017) Alternative splicing and the evolution of phenotypic novelty. *Phil. Trans. R. Soc. B* 372(1713):20150474.

48. Keren H, Lev-Maor G, & Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet* 11(5):345-355.

49. Barbosa-Morais NL, *et al.* (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338(6114):1587-1593.

50. Reddy AS, Marquez Y, Kalyna M, & Barta A (2013) Complexity of the alternative splicing landscape in plants. *Plant Cell* 25(10):3657-3683.

51. Pan Q, Shai O, Lee LJ, Frey BJ, & Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40(12):1413-1415.

52. Wang ET, *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470-476.

53. Hu Z, *et al.* (2015) Revealing Missing Human Protein Isoforms Based on Ab Initio Prediction, RNA-seq and Proteomics. *Scientific reports* 5:10940-10940.

54. Pruitt KD, *et al.* (2013) RefSeq: an update on mammalian reference sequences. *Nucleic acids research* 42(D1):D756-D763.

55. Soergel DA, Lareau LF, & Brenner SE (2013) Regulation of gene expression by coupling of alternative splicing and NMD.

56. Dvinge H, Kim E, Abdel-Wahab O, & Bradley RK (2016) RNA splicing factors as oncoproteins and tumor suppressors. *Nature reviews. Cancer* 16(7):413.

57. Wang Y, *et al.* (2015) Mechanism of alternative splicing and its regulation. *Biomedical reports* 3(2):152-158.

58. Scotti MM & Swanson MS (2016) RNA mis-splicing in disease. *Nature Reviews Genetics* 17(1):19-32.

59. Rhinn H, *et al.* (2012) Alternative α-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. *Nature communications* 3:1084.

60. Tammaro C, Raponi M, Wilson DI, & Baralle D (2012) BRCA1 exon 11 alternative splicing, multiple functions and the association with cancer. (Portland Press Limited).

61. Schwerk C & Schulze-Osthoff K (2005) Regulation of apoptosis by alternative pre-mRNA splicing. *Molecular cell* 19(1):1-13.

62. Ram DR, *et al.* (2016) Balance between short and long isoforms of cFLIP regulates Fas-mediated apoptosis in vivo. *Proceedings of the National Academy of Sciences* 113(6):1606-1611.

63. Varey A, *et al.* (2008) VEGF165b, an antiangiogenic VEGF-A isoform, binds and inhibits bevacizumab treatment in experimental colorectal carcinoma: balance of pro-and antiangiogenic VEGF-A isoforms has implications for therapy. *British journal of cancer* 98(8):1366.

64. Plowman SJ, *et al.* (2006) The K-Ras 4A isoform promotes apoptosis but does not affect either lifespan or spontaneous tumor incidence in aging mice. *Experimental cell research* 312(1):16-26.

65. Kim SS, *et al.* (2006) Hyperplasia and spontaneous tumor development in the gynecologic system in mice lacking the BRCA1-Δ11 isoform. *Molecular and cellular biology* 26(18):6983-6992.

66. Djebali S*, et al.* (2012) Landscape of transcription in human cells. *Nature* 489(7414):101-108.

67. Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, & Brazma A (2013) Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome biology* 14(7):R70.

68. Hoskins AA & Moore MJ (2012) The spliceosome: a flexible, reversible macromolecular machine. *Trends in biochemical sciences* 37(5):179-188.

69. Tress ML, Abascal F, & Valencia A (2017) Alternative splicing may not be the key to proteome complexity. *Trends in biochemical sciences* 42(2):98-110.

70. Uhlén M*, et al.* (2015) Tissue-based map of the human proteome. *Science* 347(6220):1260419.

71. Kim M-S*, et al.* (2014) A draft map of the human proteome. *Nature* 509(7502):575-581.

72. Hao Y*, et al.* (2015) Semi-supervised learning predicts approximately one third of the alternative splicing isoforms as functional proteins. *Cell reports* 12(2):183-189.

73. Wang X*, et al.* (2018) Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Molecular & Cellular Proteomics* 17(3):422-430.

74. Kelemen O*, et al.* (2013) Function of alternative splicing. *Gene* 514(1):1-30.

75. Singh B & Eyras E (2017) The role of alternative splicing in cancer. *Transcription* 8(2):91-98.

76. Kim HK, Pham MHC, Ko KS, Rhee BD, & Han J (2018) Alternative splicing isoforms in health and disease. *Pflügers Archiv-European Journal of Physiology*:1-22.

77. David A, Razali R, Wass MN, & Sternberg MJ (2012) Protein–protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human mutation* 33(2):359-363.

78. Wang X*, et al.* (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology* 30(2):159-164.

79. Sahni N*, et al.* (2015) Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161(3):647-660.

80. Bergholdt R*, et al.* (2012) Identification of novel type 1 diabetes candidate genes by integrating genome-wide association data, protein-protein interactions, and human pancreatic islet gene expression. *Diabetes* 61(4):954-962.

81. O'Roak BJ*, et al.* (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* 485(7397):246.

82. Wu G, Feng X, & Stein L (2010) A human functional protein interaction network and its application to cancer data analysis. *Genome biology* 11(5):R53.

83. Cui H, Dhroso A, Johnson N, & Korkin D (2015) The variation game: Cracking complex genetic disorders with NGS and omics data. *Methods* 79-80:18-31.

84. Yang X*, et al.* (2016) Widespread expansion of protein interaction capabilities by alternative splicing. *Cell* 164(4):805-817.

85. Hawkins RD, Hon GC, & Ren B (2010) Next-generation genomics: an integrative approach. *Nat Rev Genet* 11(7):476-486.

86. Welter D*, et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* 42(D1):D1001-D1006.

87. Frazer KA, Murray SS, Schork NJ, & Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* 10(4):241-251.

88. Oksenberg JR, Baranzini SE, Sawcer S, & Hauser SL (2008) The genetics of multiple sclerosis: SNPs to pathways to pathogenesis. *Nature Reviews Genetics* 9(7):516-526.

89. Lettre G & Rioux JD (2008) Autoimmune diseases: insights from genome-wide association studies. *Human molecular genetics* 17(R2):R116-R121.

90. Subramanian A, *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102(43):15545-15550.

91. Ramanan VK, Shen L, Moore JH, & Saykin AJ (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development. *TRENDS in Genetics* 28(7):323-332.

92. Askland K, Read C, & Moore J (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Human genetics* 125(1):63.

93. Relton CL & Davey Smith G (2010) Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. *PLoS medicine* 7(10):e1000356.

94. Baxter E, Windloch K, Gannon F, & Lee JS (2014) Epigenetic regulation in cancer progression. *Cell Bioscience*.

95. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6-21.

96. Stirzaker C, *et al.* (1997) Extensive DNA methylation spanning the Rb promoter in retinoblastoma tumors. *Cancer Res* 57(11):2229-2237.

97. Papadopoulos N, *et al.* (1994) Mutation of a mutL homolog in hereditary colon cancer. *Science* 263(5153):1625-1629.

98. Delhommeau F, *et al.* (2009) Mutation in TET2 in myeloid cancers. *N Engl J Med* 360(22):2289-2301.

99. Kumar P, Henikoff S, & Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* 4(7):1073-1081.

100. Fradin D, *et al.* (2010) Parent-of-origin effects in autism identified through genome-wide linkage analysis of 16,000 SNPs. *PLoS One* 5(9).

101. Naik US, *et al.* (2011) A study of nuclear transcription factor-kappa B in childhood autism. *PLoS One* 6(5):e19488.

102. Chauhan A, *et al.* (2011) Brain region-specific deficit in mitochondrial electron transport chain complexes in children with autism. *Journal of neurochemistry* 117(2):209-220.

103. Ginsberg MR, Rubin RA, Falcone T, Ting AH, & Natowicz MR (2012) Brain transcriptional and epigenetic associations with autism. *PLoS One* 7(9):e44736.

104. Abrahams BS & Geschwind DH (2008) Advances in autism genetics: on the threshold of a new neurobiology. *Nature Reviews Genetics* 9(5):341-355.

105. Shulha HP, *et al.* (2012) Epigenetic signatures of autism: trimethylated H3K4 landscapes in prefrontal neurons. *Archives of general psychiatry* 69(3):314-324.

106. Berko ER, *et al.* (2014) Mosaic epigenetic dysregulation of ectodermal cells in autism spectrum disorder. *PLoS Genet* 10(5):e1004402.

107. Schmitz C, *et al.* (2004) Hippocampal Neuron Loss Exceeds Amyloid Plaque Load in a Transgenic Mouse Model of Alzheimer's Disease. *The American journal of pathology* 164(4):1495-1502.

108. Wang S-C, Oelze B, & Schumacher A (2008) Age-specific epigenetic drift in late-onset Alzheimer's disease. *PLoS One* 3(7):e2698.

109. Angrisano T, *et al.* (2014) Epigenetic Switch at Atp2a2 and Myh7 Gene Promoters in Pressure Overload-Induced Heart Failure. *PloS one* 9(9):e106024.

110. Reeves HL & Friedman SL (2002) Activation of hepatic stellate cells-a key issue in liver fibrosis. *Front Biosci* 7(4):808-826.

111. Mann DA (2014) Epigenetics in liver disease. *Hepatology*.

112. Anonymous (National cancer institute report.

113. Hystad ME*, et al.* (2007) Characterization of early stages of human B cell development by gene expression profiling. *The Journal of Immunology* 179(6):3662-3671.

114. van Zelm MC*, et al.* (2005) Ig gene rearrangement steps are initiated in early human precursor B cell subsets and correlate with specific transcription factor expression. *The Journal of Immunology* 175(9):5912-5922.

115. Almamun M*, et al.* (2014) Genome-wide DNA methylation analysis in precursor B-cells. *Epigenetics* 9(12):1588-1595.

116. Hodges E*, et al.* (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Molecular cell* 44(1):17-28.

117. Berdasco M & Esteller M (2010) Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Developmental cell* 19(5):698-711.

118. Figueroa ME*, et al.* (2013) Integrated genetic and epigenetic analysis of childhood acute lymphoblastic leukemia. *The Journal of clinical investigation* 123(7):3099-3111.

119. Feinberg AP & Vogelstein B (1983) Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* 301(5895):89.

120. Liang P*, et al.* (2011) Genome-wide survey reveals dynamic widespread tissue-specific changes in DNA methylation during development. *BMC genomics* 12(1):231.

121. Ji H*, et al.* (2010) Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* 467(7313):338.

122. Xu J*, et al.* (2007) Pioneer factor interactions and unmethylated CpG dinucleotides mark silent tissue-specific enhancers in embryonic stem cells. *Proceedings of the National Academy of Sciences* 104(30):12377-12382.

123. Aran D, Sabato S, & Hellman A (2013) DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome biology* 14(3):R21.

124. Maksakova I, Mager D, & Reiss D (2008) Endogenous retroviruses. *Cellular and molecular life sciences* 65(21):3329-3347.

125. Lee E*, et al.* (2012) Landscape of somatic retrotransposition in human cancers. *Science* 337(6097):967-971.

126. Lee S-T*, et al.* (2012) A global DNA methylation and gene expression analysis of early human B-cell development reveals a demethylation signature and transcription factor network. *Nucleic acids research* 40(22):11339-11351.

127. Pastor WA, Aravind L, & Rao A (2013) TETonic shift: biological roles of TET proteins in DNA demethylation and transcription. *Nature reviews Molecular cell biology* 14(6):341.

128. Nagano T & Fraser P (2011) No-nonsense functions for long noncoding RNAs. *Cell* 145(2):178-181.

129. Wang KC & Chang HY (2011) Molecular mechanisms of long noncoding RNAs. *Molecular cell* 43(6):904-914.

130. Gupta RA*, et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 464(7291):1071.

131. Trimarchi T*, et al.* (2014) Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell* 158(3):593-606.

132. Han L, *et al.* (2014) The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nat Commun* 5:3963.

133. Xiao-Jie L, Ai-Mei G, Li-Juan J, & Jiang X (2015) Pseudogene in cancer: real functions and promising signature. *Journal of medical genetics* 52(1):17-24.

134. Dorsam RT & Gutkind JS (2007) G-protein-coupled receptors and cancer. *Nature reviews cancer* 7(2):79.

135. Weng Y-R, Cui Y, & Fang J-Y (2012) Biological functions of cytokeratin 18 in cancer. *Molecular Cancer Research* 10(4):485-493.

136. Poliseno L, *et al.* (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 465(7301):1033.

137. Sharma A, Ray R, & Rajeswari MR (2008) Overexpression of high mobility group (HMG) B1 and B2 proteins directly correlates with the progression of squamous cell carcinoma in skin. *Cancer investigation* 26(8):843-851.

138. Kwon J-H, *et al.* (2010) Overexpression of high-mobility group box 2 is associated with tumor aggressiveness and prognosis of hepatocellular carcinoma. *Clinical Cancer Research* 16(22):5511-5521.

139. Takashima N, *et al.* (2006) Expression and prognostic roles of PABPC1 in esophageal cancer: correlation with tumor progression and postoperative survival. *Oncology reports* 15(3):667-671.

140. Busche S, *et al.* (2013) Integration of high-resolution methylome and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. *Cancer research* 73(14):4323-4336.

141. Lorincz MC, Dickerson DR, Schmitt M, & Groudine M (2004) Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature Structural and Molecular Biology* 11(11):1068.

142. Maunakea AK, *et al.* (2010) Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* 466(7303):253.

143. Maunakea AK, Chepelev I, Cui K, & Zhao K (2013) Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell research* 23(11):1256.

144. Wood AJ, *et al.* (2008) Regulation of alternative polyadenylation by genomic imprinting. *Genes & development* 22(9):1141-1146.

145. Vogelstein B, *et al.* (2013) Cancer genome landscapes. *science* 339(6127):1546-1558.

146. Nagai MA, *et al.* (2010) Down-regulation of the candidate tumor suppressor gene PAR-4 is associated with poor prognosis in breast cancer. *International journal of oncology* 37(1):41-49.

147. Okudela K, *et al.* (2014) Expression of HDAC9 in lung cancer–potential role in lung carcinogenesis. *International journal of clinical and experimental pathology* 7(1):213.

148. Vlaicu SI, *et al.* (2008) Role of response gene to complement 32 in diseases. *Archivum immunologiae et therapiae experimentalis* 56(2):115.

149. Valadez JA & Cuajungco MP (2015) PAX5 is the transcriptional activator of mucolipin-2 (MCOLN2) gene. *Gene* 555(2):194-202.

150. Van Limbergen EJ, *et al.* (2014) FLT1 kinase is a mediator of radioresistance and survival in head and neck squamous cell carcinoma. *Acta oncologica* 53(5):637-645.

151. Alachkar H, *et al.* (2014) Preclinical efficacy of maternal embryonic leucine-zipper kinase (MELK) inhibition in acute myeloid leukemia. *Oncotarget* 5(23):12371.

152. Lichtenberger BM, *et al.* (2010) Autocrine VEGF signaling synergizes with EGFR in tumor cells to promote epithelial cancer development. *Cell* 140(2):268-279.

153. Wild L & Flanagan JM (2010) Genome-wide hypomethylation in cancer may be a passive consequence of transformation. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1806(1):50-57.

154. Almamun M, Schnabel JL, Gater ST, Ning J, & Taylor KH (2013) Isolation of precursor B-cell subsets from umbilical cord blood. *Journal of visualized experiments: JoVE* (74).

155. Zhang Y, *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome biology* 9(9):R137.

156. Down TA, *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology* 26(7):779.

157. Heinz S, *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* 38(4):576-589.

158. Trapnell C, *et al.* (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* 31(1):46-53.

159. Huang da W, Sherman BT, & Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4(1):44-57.

160. Gardner M, *et al.* (2018) Novel global effector mining from the transcriptome of early life stages of the soybean cyst nematode Heterodera glycines. *Scientific reports* 8(1):2505.

161. Mitchum MG, *et al.* (2013) Nematode effector proteins: An emerging paradigm of parasitism. *New Phytologist* 199(4):879-894.

162. Gheysen G & Mitchum MG (2011) How nematodes manipulate plant development pathways for infection. *Current Opinion in Plant Biology* 14(4):415-421.

163. Quentin M, Abad P, & Favery B (2013) Plant parasitic nematode effectors target host defense and nuclear functions to establish feeding cells. *Frontiers in Plant Science* 4(MAR).

164. Toruño TY, Stergiopoulos I, & Coaker G (2016) Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners. in *Annual review of phytopathology*, pp 419-441.

165. Saucet SB & Shirasu K (2016) Molecular Parasitic Plant–Host Interactions. *PLoS Pathogens* 12(12).

166. Abad P, *et al.* (2008) Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. *Nature Biotechnology* 26(8):909-915.

167. Cotton JA, *et al.* (2014) The genome and life-stage specific transcriptomes of Globodera pallida elucidate key aspects of plant parasitism by a cyst nematode. *Genome Biology* 15(3).

168. Eves-van den Akker S, *et al.* (2016) The genome of the yellow potato cyst nematode, Globodera rostochiensis, reveals insights into the basis of parasitism and virulence. *Genome Biology* 17(1).

169. Kikuchi T, *et al.* (2011) Genomic insights into the origin of parasitism in the emerging plant pathogen bursaphelenchus xylophilus. *PLoS Pathogens* 7(9).

170. Zheng J, *et al.* (2016) The ditylenchus destructor genome provides new insights into the evolution of plant parasitic nematodes. *Proceedings of the Royal Society B: Biological Sciences* 283(1835).

171. Phillips WS, *et al.* (2017) The draft genome of globodera ellingtonae. *Journal of*

Nematology 49(2):127-128.

172. Fosu-Nyarko J, Nicol P, Naz F, Gill R, & Jones MGK (2016) Analysis of the transcriptome of the infective stage of the beet cyst nematode, H. schachtii. *PLoS ONE* 11(1).

173. Haegeman A, Bauters L, Kyndt T, Rahman MM, & Gheysen G (2013) Identification of candidate effector genes in the transcriptome of the rice root knot nematode Meloidogyne graminicola. *Molecular Plant Pathology* 14(4):379-390.

174. Kumar M*, et al.* (2014) De novo transcriptome sequencing and analysis of the cereal cyst nematode, Heterodera avenae. *PLoS ONE* 9(5).

175. Petitot AS*, et al.* (2016) Dual RNA-seq reveals Meloidogyne graminicola transcriptome and candidate effectors during the interaction with rice plants. *Molecular plant pathology* 17(6):860-874.

176. Bekal S*, et al.* (2015) A SNARE-like protein and biotin are implicated in soybean cyst nematode virulence. *PLoS ONE* 10(12).

177. Gao B*, et al.* (2003) The parasitome of the phytonematode Heterodera glycines. *Molecular Plant-Microbe Interactions* 16(8):720-726.

178. Noon JB*, et al.* (2015) Eighteen new candidate effectors of the phytonematode heterodera glycines produced specifically in the secretory esophageal gland cells during parasitism. *Phytopathology* 105(10):1362-1372.

179. Gardner M, Verma A, & Mitchum MG (2015) Emerging roles of cyst nematode effectors in exploiting plant cellular processes. in *Advances in Botanical Research*, pp 259-291.

180. Eves-van den Akker S, Lilley CJ, Jones JT, & Urwin PE (2014) Identification and Characterisation of a Hyper-Variable Apoplastic Effector Gene Family of the Potato Cyst Nematodes. *PLoS Pathogens* 10(9).

181. Jones JD & Dangl JL (2006) The plant immune system. *Nature* 444(7117):323-329.

182. Na R & Gijzen M (2016) Escaping Host Immunity: New Tricks for Plant Pathogens. *PLoS Pathogens* 12(7).

183. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, & Zdobnov EM (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210-3212.

184. Haas BJ*, et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8):1494-1512.

185. Bekal S*, et al.* (2014) A novel flavivirus in the soybean cyst nematode. *Journal of General Virology* 95(PART 6):1272-1280.

186. Bekal S, Domier LL, Niblack TL, & Lambert KN (2011) Discovery and initial analysis of novel viral genomes in the soybean cyst nematode. *Journal of General Virology* 92(8):1870-1879.

187. Noel GR & Atibalentja N (2006) 'Candidatus Paenicardinium endonii' an endosymbiont of the plant-parasitic nematode Heterodera glycines (Nemata: Tylenchida), affiliated to the phylum Bacteroidetes. *International Journal of Systematic and Evolutionary Microbiology* 56(7):1697-1702.

188. Penz T*, et al.* (2012) Comparative Genomics Suggests an Independent Origin of Cytoplasmic Incompatibility in Cardinium hertigii. *PLoS Genetics* 8(10).

189. Santos-Garcia D*, et al.* (2014) The genome of cardinium cBtQ1 provides insights into enome reduction, symbiontmotility, and its settlement n bemisia tabaci. *Genome Biology and Evolution* 6(4):1013-1030.

190. Endo BY (1979) The ultrastructure and distribution of an intracellular bacterium-like

microorganism in tissues of larvae of the soybean cyst nematode, Heterodera glycines. *Journal of Ultrasructure Research* 67(1):1-14.

191. Ruark CL*, et al.* (2017) Soybean cyst nematode culture collections and field populations from North Carolina and Missouri reveal high incidences of infection by viruses. *PLoS ONE* 12(1).

192. Finn RD*, et al.* (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research* 45(D1):D190-D199.

193. Wang J, Joshi S, Korkin D, & Mitchum MG (2010) Variable domain I of nematode CLEs directs post-translational targeting of CLE peptides to the extracellular space. *Plant Signaling and Behavior* 5(12).

194. Wang J*, et al.* (2010) Dual roles for the variable domain in protein trafficking and host-specific recognition of Heterodera glycines CLE effector proteins. *New Phytologist* 187(4):1003-1017.

195. Elling AA, Davis EL, Hussey RS, & Baum TJ (2007) Active uptake of cyst nematode parasitism proteins into the plant cell nucleus. *International Journal for Parasitology* 37(11):1269-1279.

196. Baldacci-Cresp F*, et al.* (2012) (Homo)glutathione deficiency impairs root-knot nematode development in Medicago truncatula. *PLoS Pathogens* 8(1).

197. Rehman S*, et al.* (2009) A secreted SPRY domain-containing protein (SPRYSEC) from the plant-parasitic nematode Globodera rostochiensis interacts with a CC-NB-LRR protein from a susceptible tomato. *Molecular Plant-Microbe Interactions* 22(3):330-340.

198. Nakamura Y*, et al.* (2009) Prevalence of Cardinium bacteria in planthoppers and spider mites and taxonomic revision of "Candidatus Cardinium hertigii" based on detection of a new Cardinium group from biting midges. *Applied and Environmental Microbiology* 75(21):6757-6763.

199. Eleftherianos L, Atri J, Accetta J, & Castillo JC (2013) Endosymbiotic bacteria in insects: Guardians of the immune system? *Frontiers in Physiology* 4 MAR.

200. Patel N*, et al.* (2010) A nematode effector protein similar to annexins in host plants. *Journal of experimental botany* 61(1):235-248.

201. Lu SW, Tian D, Borchardt-Wier HB, & Wang X (2008) Alternative splicing: A novel mechanism of regulation identified in the chorismate mutase gene of the potato cyst nematode Globodera rostochiensis. *Molecular and Biochemical Parasitology* 162(1):1-15.

202. Noon JB*, et al.* (2016) A Plasmodium-like virulence effector of the soybean cyst nematode suppresses plant innate immunity. *The New phytologist* 212(2):444-460.

203. Thorpe P*, et al.* (2014) Genomic characterisation of the effector complement of the potato cyst nematode Globodera pallida. *BMC Genomics* 15(1).

204. Robertson L, Robertson WM, & Jones JT (1999) Direct analysis of the secretions of the potato cyst nematode Globodera rostochiensis. *Parasitology* 119(2):167-176.

205. Robertson L*, et al.* (2000) Cloning, expression and functional characterisation of a peroxiredoxin from the potato cyst nematode Globodera rostochiensis. *Molecular and Biochemical Parasitology* 111(1):41-49.

206. Guo X*, et al.* (2017) Identification of cyst nematode B-type CLE peptides and modulation of the vascular stem cell pathway for feeding cell formation. *PLoS Pathogens* 13(2).

207. Ali S*, et al.* (2015) Analysis of Globodera rostochiensis effectors reveals conserved functions of SPRYSEC proteins in suppressing and eliciting plant immune responses. *Frontiers in Plant Science* 6(AUG).

208. Diaz-Granados A, Petrescu AJ, Goverse A, & Smant G (2016) SPRYSEC effectors: A versatile protein-binding platform to disrupt plant innate immunity. *Frontiers in Plant Science* 7(OCTOBER2016).

209. Jaouannet M, *et al.* (2013) The root-knot nematode calreticulin Mi-CRT is a key effector in plant defense suppression. *Molecular Plant-Microbe Interactions* 26(1):97-105.

210. Ganji S, Jenkins JN, & Wubben MJ (2014) Molecular characterization of the reniform nematode C-type lectin gene family reveals a likely role in mitigating environmental stresses during plant parasitism. *Gene* 537(2):269-278.

211. Guzmán P (2012) The prolific ATL family of RING-H2 ubiquitin ligases. *Plant Signaling and Behavior* 7(8).

212. Cosgrove DJ (2000) Loosening of plant cell walls by expansins. *Nature* 407(6802):321-326.

213. Minic Z & Jouanin L (2006) Plant glycoside hydrolases involved in cell wall polysaccharide degradation. *Plant Physiology and Biochemistry* 44(7-9):435-449.

214. Weinstein L & Albersheim P (1979) Structure of plant cell walls. *Plant Physiol.* 63:425-432.

215. Qin L, *et al.* (2004) A nematode expansin acting on plants. *Nature* 427(6969):30.

216. Wieczorek K, *et al.* (2006) Expansins are involved in the formation of nematode-induced syncytia in roots of Arabidopsis thaliana. *Plant Journal* 48(1):98-112.

217. De Boer JM, *et al.* (1999) Developmental expression of secretory β-1,4-endoglucanases in the subventral esophageal glands of Heterodera glycines. *Molecular Plant-Microbe Interactions* 12(8):663-669.

218. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.

219. Grabherr MG, *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29(7):644-652.

220. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4):357-359.

221. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, & Wheeler DL (2005) GenBank. *Nucleic Acids Research* 33(DATABASE ISS.):D34-D38.

222. Magrane M & Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database* 2011:bar009.

223. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25(17):3389-3402.

224. Finn RD, Clements J, & Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research* 39(SUPPL. 2):W29-W37.

225. Punta M, *et al.* (2012) The Pfam protein families database. *Nucleic Acids Research* 40(D1):D290-D301.

226. Petersen TN, Brunak S, Von Heijne G, & Nielsen H (2011) SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods* 8(10):785-786.

227. Sonnhammer EL, von Heijne G, & Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 6:175-182.

228. Howe KL, *et al.* (2016) WormBase 2016: Expanding to enable helminth genomic research. *Nucleic Acids Research* 44(D1):D774-D780.

229. Li B & Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12(1).

230. Chandramohan R, Wu P, Phan J, & Wang M (2016) Benchmarking RNA-Seq quantification tools. *Conference Proceedings of the IEEE Engineering in Medicine and Biology Society*.

231. Robinson MD, McCarthy DJ, & Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1):139-140.

232. Conesa A*, et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol* 17(1):13-13.

233. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5):1792-1797.

234. Kumar S, Stecher G, & Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* 33(7):1870-1874.

235. Bray NL, Pimentel H, Melsted P, & Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* 34(5):525-527.

236. Pimentel HJ, Bray N, Puente S, Melsted P, & Pachter L (2016) Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv*.

237. Madden T (2002) The BLAST sequence analysis tool. *The NCBI Handbook*.

238. Nguyen Ba AN, Pogoutse A, Provart N, & Moses AM (2009) NLStradamus: A simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10.

239. Dhroso A, Eidson S, & Korkin D (2018) Genome-wide prediction of bacterial effectors across six secretion system types using a feature-based supervised learning framework. *bioRxiv*.

240. Hall MA (1999) Correlation-based feature selection for machine learning. (The University of Waikato).

241. Weber L*, et al.* (2017) Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis. *PloS one* 12(3):e0172491.

242. Mortazavi A, Williams BA, McCue K, Schaeffer L, & Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* 5(7):621-628.

243. Zhang Z, Pal S, Bi Y, Tchou J, & Davuluri RV (2013) Isoform level expression profiles provide better cancer signatures than gene level expression profiles. *Genome medicine* 5(4):33-33.

244. Nilsen TW & Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463(7280):457-463.

245. Wei IH, Shi Y, Jiang H, Kumar-Sinha C, & Chinnaiyan AM (2014) RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia* 16(11):918-927.

246. Achim K*, et al.* (2015) High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol* 33(5):503-509.

247. Danielsson F, James T, Gomez-Cabrero D, & Huss M (2015) Assessing the consistency of public human tissue RNA-seq data sets. *Briefings in bioinformatics* 16(6):941-949.

248. Wan Y*, et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* 505(7485):706-709.

249. Mele M*, et al.* (2015) Human genomics. The human transcriptome across tissues and

individuals. *Science* 348(6235):660-665.

250. Trincado JL, Sebestyen E, Pages A, & Eyras E (2016) The prognostic potential of alternative transcript isoforms across human tumors. *Genome medicine* 8(1):85.
251. Shen S, Wang Y, Wang C, Wu YN, & Xing Y (2016) SURVIV for survival analysis of mRNA isoform variation. *Nature communications* 7.
252. Climente-González H, Porta-Pardo E, Godzik A, & Eyras E (2017) The functional impact of alternative splicing in cancer. *Cell reports* 20(9):2215-2226.
253. Edwards TM & Myers JP (2007) Environmental exposures and gene regulation in disease etiology. *Environmental health perspectives* 115(9):1264-1270.
254. Gamazon ER & Stranger BE (2014) Genomics of alternative splicing: evolution, development and pathophysiology. *Human genetics* 133(6):679-687.
255. Luco RF, Allo M, Schor IE, Kornblihtt AR, & Misteli T (2011) Epigenetics in alternative pre-mRNA splicing. *Cell* 144(1):16-26.
256. Sebestyen E, Zawisza M, & Eyras E (2015) Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res* 43(3):1345-1356.
257. Cáceres JF & Kornblihtt AR (2002) Alternative splicing: multiple control mechanisms and involvement in human disease. *TRENDS in Genetics* 18(4):186-193.
258. Tarca AL, Carey VJ, Chen XW, Romero R, & Draghici S (2007) Machine learning and its applications to biology. *PLoS computational biology* 3(6):e116.
259. Liu C, Che D, Liu X, & Song Y (2013) Applications of machine learning in genomics and systems biology. *Comput Math Methods Med* 2013:587492.
260. Neelima E & Babu MP (2017) A comparative Study of Machine Learning Classifiers over Gene expressions towards Cardio Vascular Diseases Prediction. *International Journal of Computational Intelligence Research* 13(3):403-424.
261. Libbrecht MW & Noble WS (2015) Machine learning applications in genetics and genomics. *Nature Reviews Genetics* 16(6):321-332.
262. Vandesompele J*, et al.* (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol* 3(7):RESEARCH0034.
263. Jagga Z & Gupta D (2014) Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms. *BMC proceedings* 8(Suppl 6 Proceedings of the Great Lakes Bioinformatics Confer):S2.
264. Costa IG, de Carvalho FdA, & de Souto MC (2004) Comparative analysis of clustering methods for gene expression time course data. *Genetics and Molecular Biology* 27(4):623-631.
265. Pirooznia M, Yang JY, Yang MQ, & Deng Y (2008) A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics* 9(1):S13.
266. Mudge JM, Frankish A, & Harrow J (2013) Functional transcriptomics in the post-ENCODE era. *Genome Res* 23(12):1961-1973.
267. Consortium G (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348(6235):648-660.
268. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, & Craig DW (2016) Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics* 17(5):257-271.

269. Thompson JA, Tan J, & Greene CS (2016) Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ* 4:e1621.
270. Labuzzetta CJ, *et al.* (2016) Complementary feature selection from alternative splicing events and gene expression for phenotype prediction. *Bioinformatics* 32(17):i421-i429.
271. Yu Y, *et al.* (2014) A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. *Nat Commun* 5:3230-3230.
272. Kodama Y, Shumway M, & Leinonen R (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic acids research* 40(D1):D54-D56.
273. Weinstein JN, *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nature genetics* 45(10):1113-1120.
274. Edge SB & Compton CC (2010) The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of surgical oncology* 17(6):1471-1474.
275. Hall MA & Smith LA (1998) Practical feature subset selection for machine learning. *Proceedings of the 21st Australasian Computer Science Conference*, pp 181-191.
276. Xiong HY, *et al.* (2015) The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347(6218):1254806.
277. Schirmer M, D'Amore R, Ijaz UZ, Hall N, & Quince C (2016) Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC bioinformatics* 17(1):125.
278. Boulesteix AL, Janitza S, Kruppa J, & König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2(6):493-507.
279. Goldstein LD, *et al.* (2016) Prediction and Quantification of Splice Events from RNA-Seq Data. *PloS one* 11(5):e0156132.
280. Alamancos GP, Pagès A, Trincado JL, Bellora N, & Eyras E (2015) Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* 21(9):1521-1531.
281. Zhu X (2005) Semi-supervised learning literature survey. (Computer Sciences, University of Wisconsin-Madison).
282. LeCun Y, Bengio Y, & Hinton G (2015) Deep learning. *Nature* 521(7553):436-444.
283. Vapnik V & Vashist A (2009) A new learning paradigm: learning using privileged information. *Neural networks : the official journal of the International Neural Network Society* 22(5-6):544-557.
284. Zhu Y, Qiu P, & Ji Y (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nature methods* 11(6):599-600.
285. Engstrom PG, *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 10(12):1185-1191.
286. Gordon A & Hannon G (2010) Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit*.
287. NCBI RC (2013) Database resources of the National Center for Biotechnology Information. *Nucleic acids research* 41(Database issue):D8.
288. Kim D, *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36-R36.
289. Trapnell C, *et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562-578.

290. Flicek P, *et al.* (2012) Ensembl 2012. *Nucleic Acids Res* 40(Database issue):D84-90.

291. Wang K, *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* 38(18):e178-e178.

292. Mark Hall EF, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1).

293. Rish I (2001) An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence*, (IBM New York), pp 41-46.

294. Li T, Zhang C, & Ogihara M (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20(15):2429-2437.

295. Liu H, Li J, & Wong L (2002) A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome informatics* 13:51-60.

296. McCallum A & Nigam K (1998) A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*, (Citeseer), pp 41-48.

297. Kim S-B, Han K-S, Rim H-C, & Myaeng SH (2006) Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering* 18(11):1457-1466.

298. Niculescu-Mizil A & Caruana R (2005) Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning*, (ACM), pp 625-632.

299. Ng AY & Jordan MI (2002) On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems* 2:841-848.

300. Shevade SK & Keerthi SS (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics* 19(17):2246-2253.

301. Wang L, *et al.* (2013) CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research* 41(6):e74-e74.

302. Sartor MA, Leikauf GD, & Medvedovic M (2009) LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics* 25(2):211-217.

303. Liao J & Chin K-V (2007) Logistic regression for disease classification using microarray data: model selection in a large p and small n case. *Bioinformatics* 23(15):1945-1951.

304. Asur S, Raman P, Otey ME, & Parthasarathy S (2006) A model-based approach for mining membrane protein crystallization trials. *Bioinformatics* 22(14):e40-e48.

305. Quinlan J (1993) C4. 5: Programs for Machine Learning. C4. 5-programs for machine learning/J. Ross Quinlan. (Morgan Kaufmann Publishers).

306. Quinlan JR (1979) *Discovering rules by induction from large collections of examples* (Expert systems in the micro electronic age. Edinburgh University Press).

307. Breiman L (2001) Random forests. *Machine learning* 45(1):5-32.

308. Vapnik VN (1998) *Statistical learning theory* (Wiley, New York) pp xxiv, 736 p.

309. Dou Y, Yao B, & Zhang C (2014) PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. *Amino acids* 46(6):1459-1469.

310. Zhao N, Pang B, Shyu CR, & Korkin D (2011) Feature-based classification of native and non-native protein–protein interactions: Comparing supervised and semi-supervised

learning approaches. *Proteomics* 11(22):4321-4330.

311. Hirose S, Shimizu K, Kanai S, Kuroda Y, & Noguchi T (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics* 23(16):2046-2053.
312. Alber F, Förster F, Korkin D, Topf M, & Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.* 77:443-477.
313. Corominas R, *et al.* (2014) Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature communications* 5:3650.
314. Wang X, *et al.* (2017) Detection of proteome diversity resulted from alternative splicing is limited by trypsin cleavage specificity. *Molecular & Cellular Proteomics*:mcp. RA117. 000155.
315. Hu Z, *et al.* (2015) Revealing missing human protein isoforms based on ab initio prediction, RNA-seq and proteomics. *Scientific reports* 5:10940.
316. Kuang X, Dhroso A, Han JG, Shyu C-R, & Korkin D (2016) DOMMINO 2.0: integrating structurally resolved protein-, RNA-, and DNA-mediated macromolecular interactions. *Database* 2016.
317. Berman HM, *et al.* (2006) The Protein Data Bank, 1999–. *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, (Springer), pp 675-684.
318. Stein A, Russell RB, & Aloy P (2005) 3did: interacting protein domains of known three-dimensional structure. *Nucleic acids research* 33(suppl_1):D413-D417.
319. Zhao N, Pang B, Shyu CR, & Korkin D (2011) Charged residues at protein interaction interfaces: unexpected conservation and orchestrated divergence. *Protein Science* 20(7):1275-1284.
320. Andreani J, Faure G, & Guerois R (2012) Versatility and invariance in the evolution of homologous heteromeric interfaces. *PLoS computational biology* 8(8):e1002677.
321. Zhao N, Han JG, Shyu C-R, & Korkin D (2014) Determining Effects of Non-synonymous SNPs on Protein-Protein Interactions using Supervised and Semi-supervised Learning. *PLoS computational biology* 10(5):e1003592.
322. Singh A, *et al.* (2007) MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic acids research* 36(suppl_1):D815-D819.
323. Keren H, Lev-Maor G, & Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* 11(5):345.
324. Cui H, Dhroso A, Johnson N, & Korkin D (2015) The variation game: Cracking complex genetic disorders with NGS and omics data. *Methods* 79:18-31.
325. Lara-Pezzi E, Gómez-Salinero J, Gatto A, & García-Pavía P (2013) The alternative heart: impact of alternative splicing in heart disease. *Journal of cardiovascular translational research* 6(6):945-955.
326. Chapelle O, Scholkopf B, & Zien A (2009) Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks* 20(3):542-542.
327. Xia Z, Wu L-Y, Zhou X, & Wong ST (2010) Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC systems biology*, (BioMed Central), p S6.
328. Rual J-F, *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 437(7062):1173.
329. Rolland T, *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*

159(5):1212-1226.
330. Venkatesan K*, et al.* (2009) An empirical framework for binary interactome mapping. *Nature methods* 6(1):83.
331. Yu H*, et al.* (2011) Next-generation sequencing to generate interactome datasets. *Nature methods* 8(6):478.
332. Zhong Q*, et al.* (2016) An inter-species protein–protein interaction network across vast evolutionary distance. *Molecular systems biology* 12(4):865.
333. Cortes C & Vapnik V (1995) Support-vector networks. *Machine learning* 20(3):273-297.
334. Chang C-C & Lin C-J (2011) LIBSVM: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)* 2(3):27.
335. Pedregosa F*, et al.* (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12(Oct):2825-2830.
336. Kuang X*, et al.* (2011) DOMMINO: a database of macromolecular interactions. *Nucleic acids research* 40(D1):D501-D506.
337. Andreeva A*, et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic acids research* 32(suppl_1):D226-D229.
338. Wilson D, Madera M, Vogel C, Chothia C, & Gough J (2006) The SUPERFAMILY database in 2007: families and functions. *Nucleic acids research* 35(suppl_1):D308-D313.
339. Hira ZM & Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics* 2015.
340. Perovic V*, et al.* (2017) TRI_tool: a web-tool for prediction of protein–protein interactions in human transcriptional regulation. *Bioinformatics* 33(2):289-291.
341. Pan X-Y, Zhang Y-N, & Shen H-B (2010) Large-Scale prediction of human protein−protein interactions from amino acid sequence based on latent topic features. *Journal of proteome research* 9(10):4992-5001.
342. Zerbino DR*, et al.* (2018) Ensembl 2018. *Nucleic Acids Research* 46(D1):D754-D761.
343. Wang C-Y & Liao JK (2012) A mouse model of diet-induced obesity and insulin resistance. *mTOR*, (Springer), pp 421-433.
344. Speakman J, Hambly C, Mitchell S, & Krol E (2007) *Animal models of obesity* pp 55-61.
345. Tremblay Fdr, Gagnon A, Veilleux A, Sorisky A, & Marette A (2005) Activation of the mammalian target of rapamycin pathway acutely inhibits insulin signaling to Akt and glucose transport in 3T3-L1 and human adipocytes. *Endocrinology* 146(3):1328-1337.
346. Kim D*, et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* 14(4):R36.
347. Trapnell C*, et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* 7(3):562.
348. Anonymous (Type 2 Diabetes Knowledge Portal.
349. Jason F*, et al.* (2017) Sequence data and association statistics from 12,940 type 2 diabetes cases and controls. *Scientific data* 4:170179.
350. Gaulton KJ*, et al.* (2015) Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. *Nature genetics* 47(12):1415.
351. Mercader JM*, et al.* (2017) A loss-of-function splice acceptor variant in IGF2 is protective for type 2 diabetes. *Diabetes*:db170187.
352. Von Mering C*, et al.* (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic acids research* 33(suppl_1):D433-D437.

353. Negrini S, Gorgoulis VG, & Halazonetis TD (2010) Genomic instability—an evolving hallmark of cancer. *Nature reviews Molecular cell biology* 11(3):220.

354. Hanahan D & Weinberg RA (2011) Hallmarks of cancer: the next generation. *cell* 144(5):646-674.

355. Dagogo-Jack I & Shaw AT (2017) Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*.

356. Supek F & Lehner B (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* 521(7550):81.

357. Poulos RC & Wong JW (2018) Finding cancer driver mutations in the era of big data research. *Biophysical Reviews*:1-9.

358. Reimand J & Bader GD (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular systems biology* 9(1).

359. Bailey MH*, et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell* 173(2):371-385. e318.

360. Buljan M, Blattmann P, Aebersold R, & Boutros M (2018) Systematic characterization of pan-cancer mutation clusters. *Molecular systems biology* 14(3):e7974.

361. Olivier M, Hollstein M, & Hainaut P (2010) TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology* 2(1):a001008.

362. Brognard J & Hunter T (2011) Protein kinase signaling networks in cancer. *Current opinion in genetics & development* 21(1):4-11.

363. Vail ME*, et al.* (2014) Targeting EphA3 inhibits cancer growth by disrupting the tumor stromal microenvironment. *Cancer research* 74(16):4470-4481.

364. Lawrence MS*, et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214-218.

365. Abaffy T (2015) Human olfactory receptors expression and their role in non-olfactory tissues-a mini-review. *Journal of Pharmacogenomics & Pharmacoproteomics* 6(4):1.

366. Neuhaus EM*, et al.* (2009) Activation of an olfactory receptor inhibits proliferation of prostate cancer cells. *J Biol Chem* 284(24):16218-16225.

367. Sanz G*, et al.* (2014) Promotion of cancer cell invasiveness and metastasis emergence caused by olfactory receptor stimulation. *PLoS One* 9(1):e85110.

368. Gelis L*, et al.* (2016) Functional characterization of the odorant receptor 51E2 in human melanocytes. *Journal of Biological Chemistry* 291(34):17772-17786.

369. Ranzani M*, et al.* (2017) Revisiting olfactory receptors as putative drivers of cancer. *Wellcome open research* 2.

370. Marx V (2014) Cancer genomes: discerning drivers from passengers. (Nature Publishing Group).

371. Forbes SA*, et al.* (2014) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research* 43(D1):D805-D811.

372. Fiser A & Šali A (2003) Modeller: generation and refinement of homology-based protein structure models. *Methods in enzymology,* (Elsevier), Vol 374, pp 461-491.

373. Korkin D*, et al.* (2006) Structural modeling of protein interactions by analogy: application to PSD-95. *PLoS computational biology* 2(11):e153.

374. Nordström KJ, Almén MS, Edstam MM, Fredriksson R, & Schiöth HB (2011) Independent HHsearch, Needleman–Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. *Molecular biology and evolution* 28(9):2471-2480.

375. McGuffin LJ, Bryson K, & Jones DT (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404-405.

376. Kabsch W & Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577-2637.

377. Mihel J, Šikić M, Tomić S, Jeren B, & Vlahoviček K (2008) PSAIA–protein structure and interaction analyzer. *BMC structural biology* 8(1):21.

378. Wu J, *et al.* (2008) Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* 25(1):30-35.

379. Ebina T, Toh H, & Kuroda Y (2010) DROP: an SVM domain linker predictor trained with optimal features selected by random forest. *Bioinformatics* 27(4):487-494.

380. Wright SJ (2012) Accelerated block-coordinate relaxation for regularized optimization. *SIAM Journal on Optimization* 22(1):159-186.

381. Goodman J (2004) Exponential priors for maximum entropy models. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

382. Lee S-I, Lee H, Abbeel P, & Ng AY (2006) Efficient l~ 1 regularized logistic regression. *AAAI*, pp 401-408.

383. Webb B & Sali A (2014) Protein structure modeling with MODELLER. *Protein Structure Prediction*:1-15.

384. Jordan DM, Ramensky VE, & Sunyaev SR (2010) Human allelic variation: perspective from protein function, structure, and evolution. *Current opinion in structural biology* 20(3):342-350.

385. Venter JC, *et al.* (2001) The sequence of the human genome. *Science* 291(5507):1304-1351.

386. Consortium IHGS (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860.

387. Blencowe BJ (2006) Alternative splicing: new insights from global analyses. *Cell* 126(1):37-47.

388. Mudge JM & Harrow J (2016) The state of play in higher eukaryote gene annotation. *Nature Reviews Genetics* 17(12):758.

389. Tress ML, *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences* 104(13):5495-5500.

390. Lynch KW (2015) Thoughts on NGS, alternative splicing and what we still need to know. *RNA* 21(4):683-684.

391. Hu Z, *et al.* (2015) Revealing missing human protein isoforms based on Ab Initio prediction, RNA-seq and proteomics. *Scientific reports* 5:srep10940.

392. Buljan M, *et al.* (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* 46(6):871-883.

393. Ellis JD, *et al.* (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell* 46(6):884-892.

394. Liu S & Cheng C (2013) Alternative RNA splicing and cancer. *Wiley interdisciplinary reviews. RNA* 4(5):547-566.

395. Pal S, Gupta R, & Davuluri RV (2012) Alternative transcription and alternative splicing in cancer. *Pharmacology & therapeutics* 136(3):283-294.

396. Light S & Elofsson A (2013) The impact of splicing on protein domain architecture. *Current opinion in structural biology* 23(3):451-458.

397. Hiller M, Huse K, Platzer M, & Backofen R (2005) Creation and disruption of protein features by alternative splicing-a novel mechanism to modulate function. *Genome biology* 6(7):R58.

398. Liu S & Altman RB (2003) Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic acids research* 31(16):4828-4835.

399. Harrow JL, *et al.* (2013) The vertebrate genome annotation browser 10 years on. *Nucleic acids research* 42(D1):D771-D779.

400. Shionyu M, Yamaguchi A, Shinoda K, Takahashi K-i, & Go M (2008) AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic acids research* 37(suppl_1):D305-D309.

401. Kim N, Alekseyenko AV, Roy M, & Lee C (2007) The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res* 35(Database issue):D93-98.

402. Martelli PL, *et al.* (2011) ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Res* 39(Database issue):D80-85.

403. Koscielny G, *et al.* (2009) ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics* 93(3):213-220.

404. Papatheodorou I, *et al.* (2017) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic acids research* 46(D1):D246-D251.

405. Weijers S, *et al.* (2012) KALLISTO: cost effective and integrated optimization of the urban wastewater system Eindhoven. *Water Practice and Technology* 7(2):wpt2012036.

406. Niklas KJ, Bondos SE, Dunker AK, & Newman SA (2015) Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. *Frontiers in cell and developmental biology* 3:8.

407. Chendrimada TP, *et al.* (2005) TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 436(7051):740.

408. Ferguson LR (2011) RNA silencing: Mechanism, biology and responses to environmental stress. *Mutat Res* 714(1-2):93-94.

409. Ponting CP, Oliver PL, & Reik W (2009) Evolution and functions of long noncoding RNAs. *Cell* 136(4):629-641.

410. Memczak S, *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495(7441):333-338.

411. Brodersen P & Voinnet O (2009) Revisiting the principles of microRNA target recognition and mode of action. *Nat Rev Mol Cell Biol* 10(2):141-148.