COMPUTATIONAL METHODS FOR THE FUNCTIONAL ANALYSIS OF DNA SEQUENCE VARIANTS

by

Lucas Santana dos Santos

BS, Universidade Federal de Minas Gerais, 2008

MS, University of Pittsburgh, 2012

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Lucas Santana dos Santos

It was defended on

April 4, 2017

and approved by

Richard Duerr, MD, Medicine

Vanathi Gopalakrishnan, PhD, Associate Professor, Biomedical Informatics

Xia Jiang, PhD, Associate Professor, Biomedical Informatics

Dissertation Director: Panayiotis Benos, PhD, Professor, Computational and Systems Biology

COMPUTATIONAL METHODS FOR THE FUNCTIONAL ANALYSIS OF DNA SEQUENCE VARIANTS

Lucas Santana dos Santos, PhD

University of Pittsburgh, 2017

Complex diseases, such as cancer and inflammatory bowel disease, are caused by a combination of genetic and environmental factors. The advent of next-generation sequencing (NGS) technology allowed the genome-wide investigation of the underlying genetic causes of complex disorders. Analysis of the large amount of data generated by NGS is computationally intensive and require new computational methods. One of the current problems in genomic data analysis is the lack of computational methods for functional annotation of DNA sequence variants (DSVs), especially regulatory DNA sequence variants (rDSVs). In recent years, rDSVs have been shown to be the primary cause of complex diseases, supported by the fact that functional regulatory sites are more polymorphic than coding regions, and that rDSVs vastly outnumber coding variants. Also, GWAS studies of complex traits have shown that SNPs with the strongest association signals lie outside known genes in non-coding regions of the genome.

This dissertation contributes to a solution for the lack of computational methods for the analysis of DNA sequence variants. Two novel computational methods for the analysis of DSVs are proposed here: 1) an algorithm, called is-miRSNP, for the prediction of the effect of 3'UTR DSVs on miRNA binding, 2) a pipeline for the functional annotation of DSVs using NGS. The is-miRSNP algorithm uses a binding-energy approach for the prediction of DSVs effects on miRNA binding. The algorithm is flexible enough to process large amounts of data and can be easily integrated in existing pipelines. Experiments using a manually curated set of experimentally validated DSVs-miRNA showed that is-miRSNP outperforms all most popular existing methods. The pipeline for functional annotation of functional DSVs utilizes state-of-the-art existing computational methods. The pipeline has been applied to an effector memory T cell RNA-Seq dataset that is related to inflammatory bowel disease and has identified biologically relevant genes and isoforms that are differentially expressed upon treatment with Prostaglandin E2. Important pathways and biologically relevant DSVs were also identified and recovered. These methods have the potential to help clinicians and researchers analyze and interpret genomic datasets, and might in the future help the development of new diagnostics methods and treatments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

This dissertation is dedicated to my wife Michelle and my daughter Lavinia. I could not have done it without their unconditional love and support. A special recognition to my parents Claudia and Arquimedes, and my siblings Luíza and Eduardo. I thank them for their encouragement in pursuing a career in the United States and their infinite love.

A special thanks to my advisor Dr. Takis Benos, I could have not asked for a better mentor. Your patience, support and encouragement helped me to grow as a research scientist.

I would like to thank my committee members, Dr. Richard Duerr, Dr. Vanathi Gopalakrishnan and, Dr. Jia Xiang for serving as my committee members. I also want to thank you for your scientific and academic advice and support.

I will always be grateful to Dr. Bennett Van Houten for the opportunity to join his lab at the early stage of my career, for all the knowledge, advice, support and encouragement.

I thank my classmates and friends Dr. Arturo Lopez Pineda, Dr. Charalambos Floudas and Dr. Fernado Suárez Obando for all the joyous moments, your ideas and scientific discussions. I also thank all the current and past members of the Benos Lab for their scientific input.

I also thank Dr. Elise Fouquerel and Dr. Shikhar Uttam for all the joyous moments, support and care.

My gratitude to Ms. Toni Porterfield for her support and advice, without her I would have been lost amidst graduate school responsibilities and bureaucracy.

I thank all the members of UPMC Molecular Genomic Pathology lab. A special thanks to Dr. Marina Nikiforova and Dr. Yury Nikiforov for the opportunity of working in such a great lab. I thank the current and past members MGP-bioinformatics team, Dr. Keith Callenbergh, Mr. Mehool Patel and Mr. Liang Chen, it has been a pleasure working with you. My gratitude to Dr. Abigail Wald for her support and understanding.

Finally, I want to thank all my family and friends in Brazil and in the USA. They have been an important source of love and support.

1.0 INTRODUCTION

Genetic variation among individuals can be classified into three categories: 1) single nucleotide variation (SNV); 2) intragenic insertion and deletion (INDEL); 3) structural variants that include translocations and copy number variation (Lee, Arvai, & Jones, 2015). The most common type of genetic variation is SNV. Whole genome sequencing of 2,636 individuals identified 20 million SNVs, whereas only 1.5 million INDELs were found (Gudbjartsson et al., 2015). This dissertation will focus on the study of SNVs, due to their higher frequency and bigger relevance to the understanding of human genetic variation.

SNVs can be further classified based on two criteria: population frequency and genomic location. Based on the first criteria SNVs can be split into single nucleotide polymorphisms (SNPs) and mutations. The distinction between these two is based on allelic frequency in the general population. SNPs have an allelic frequency greater or equal to 1%, whereas mutations are rarer with a frequency of less than 1% (Brookes, 1999). It is important to note that in this study, the term SNV will be used as no distinction is being made regarding the allelic frequency of a variant in the population. Using genomic location SNVs can be divided into coding and regulatory DNA Sequence Variants (DSVs). Coding variants are located only in exons, whereas regulatory variants (rDSV) can be found in the rest of the genome (promoter regions, 3’UTRs, 5’UTRs, introns and intergenic regions).

Coding variants can have a direct effect on an individual's phenotype. It can act by changing the quantity of the quality of a protein (Buckland, 2006). Alterations in protein quality are the result of changes in the mRNA sequence, which in turn translate into modifications of the protein's amino acid sequence and/or protein structure. Changes in quantity are the result of modifications on the rate which the mRNA is translated into protein.

Regulatory coding variants do not directly affect the protein quality, but instead alter mRNA abundance by changing gene expression. There are 3 ways that a rDSV can affect gene expression: 1) by altering transcription factor binding sites (Deplancke, Alpern, & Gardeux, 2016); 2) by changing microRNA binding sites (Mishra, Mishra, Banerjee, & Bertino, 2008); 3) by altering splice-sites (Baralle & Baralle, 2005).

Transcription factors (TF) are DNA-binding proteins, which bind to specific short DNA sequences usually located in the promoter regions of genes. The binding of these proteins can lead to activation or repression of the target gene (Latchman, 1997). DSVs can potentially alter the DNA sequence of a gene promoter region and consequently, create or destroy a TF binding site. One of the first accounts of a DSV effecting the binding of a TF was hypothesized by Orkin and coworkers in 1982 (Deplancke et al., 2016; Orkin et al., 1982). The authors suggested that a DSV (C to T, at position 87 of the transcription start site of the Beta Hemoglobin gene) was affecting the transcriptional recruitment of some unknown TF. Finally, in 1993, Miller and Biecker showed that the unknown TF affected by the DSV was the Kruppel like Factor 1 protein (Miller & Bieker, 1993).

MicroRNAs (miRNAs) are short (~22 nt) endogenous RNAs that control gene-expression in eukaryotes (Ambros, 2004; Bartel, 2004). Binding of these short non-coding RNAs to the 3'UTR of a messenger RNA (mRNA) leads to the mRNA's degradation. Recently, several polymorphisms that affect miRNA binding (miRSNPs) have been identified (Mishra et al., 2008). One example is the DSV C to T, at position 829 of the 3'UTR of the human dihydrofolate reductase (DHFR) gene, which impairs the binding of miR-24 at this particular site and leads to the overexpression of the DHFR gene. The DHFR gene is responsible for metabolizing the chemotherapeutic drug methotrexate. Such finding has important clinical implications as tumors which have this DSV are resistance to methotrexate and should be treated with an alternative chemotherapeutic regimen (Mishra et al., 2007).

Transcription in eukaryotes is a multi-step process: it begins with the transcription of a DNA template into an unprocessed RNA molecule called the pre-mRNA. Next, the pre-mRNA is processed; its introns and sometimes exons are removed by a process called splicing. The spliced pre-mRNA is the mature form of the mRNA, and will be later translated into a protein (Clancy, 2008).

Splicing is an essential step in the formation of mRNA and will consequently impact the protein that will be synthetized. Therefore, missplicing of genes lead to diseases like leukemia, hematolymphoid neoplasias, retinitis pigmentosa and microcephalic osteodysplastic primordial dwarfism type 1 (MODPD1) (Mohamed et al., 2014; Singh & Cooper, 2012). A common cause for missplicing is the presence of DSVs that will either create or destroy a splice-site. It is estimated that from those point mutations that alter splice-sites, 15% will result in some type of genetic disorder (Baralle & Baralle, 2005).

Monogenic Mendelian disorders, such as cystic fibrosis, Tay-Sachs disease, and sickle cell anemia, have been linked to specific protein coding DSVs. However, only very few polygenic diseases have been liked to specific protein coding DSVs in a single gene (this is the case of BRCA1 mutations in breast cancer, which only comprises of a very small percentage of all breast cancer cases).

Monogenic Mendelian disorders only comprise a small portion of all human disorders with most human diseases being classified as ‘complex disorders’. Complex diseases are caused by a combination of genetic and environmental factors. Examples of complex diseases are cancers, inflammatory bowel disease, asthma, Parkison’s disease, Alzheimer’s diseases (Buckland, 2006; Hunter, 2005).

The small number of causal protein-coding variants associated with complex diseases led to the hypothesis that majority of disease-causing variants are probably located in the non-coding DNA (Cobb, Busst, Petrou, Harrap, & Ellis, 2008). The hypothesis that rDSVs are the main cause for complex diseases is further support by the fact that functional regulatory sites are more polymorphic than coding regions (Pampin & Rodriguez-Rey, 2007). Also, rDSVs vastly outnumber coding variants, as less than two percent of the genome is composed by exons (Scacheri & Scacheri, 2015). Therefore, is likely that phenotypes are a consequence of variation in the expression of genes rather than the expression of different genes, and that rDSVs are the main source of genetic variation (Buckland, 2006; Hudson, 2003; Knight, 2005; Scacheri & Scacheri, 2015; Tak & Farnham, 2015).

1.1 THE PROBLEM

As discussed in the introduction, various complex diseases are likely caused by rDSVs that have some effect on the regulation of gene expression. Unfortunately, the identification and functional interpretation of rDSVs is hard (Buckland, 2006; Hudson, 2003; Knight, 2005; Pampin &

Rodriguez-Rey, 2007). There are two main challenges in identifying, evaluating and prioritizing rDSVs: 1) lack of methods to accurately evaluate and prioritize rDSVs, especially variants that affect miRNA binding sites; 2) need for genomic data analysis pipelines that can identify, evaluate and prioritize rDSVs.

The first challenge deals with identification and prioritization of rDSVs that affect miRNA binding sites. Various methods and datasets for predicting rDSVs that alter transcription factor binding sites and splice-sites already exist. However, available computational methods for predicting the effect of DSVs that alter miRNA binding sites are not accurate and results are hard to interpret.

The second problem relates to lack of analysis pipeline that focus on the identification and prioritization of DSVs found in genomic datasets. With the abundance of public genomic datasets and the popularization of NGS, methods that can properly integrate and analyze various genomic datasets are in great need.

1.1.1 Predicting and prioritizing rDSVs that affect miRNA binding sites

The Human Genome Project revealed that ~99% of the genome is composed of non-coding regions. The importance of these non-coding regions became self-evident when comparative genomics revealed that a great number of these regions are conserved among mammalians. In recent years, genome-wide association studies (GWAS) showed that the majority of complex trait-associated loci are in non-coding regions (Kellis et al., 2014).

The lack of understanding of the biology of non-coding regions associated with its evolutionary and medical importance lead to the creation of Encyclopedia of DNA Elements (ENCODE) project. The ENCODE project is a big international initiative that has as the main goal to identify and characterize functional elements within the human genome (Consortium, 2012; Diehl & Boyle, 2016; Kellis et al., 2014; Pazin, 2015; Qu & Fang, 2013).

Using high-throughput NGS-based assays (RNA-seq, CAGE, RNA-PET, ChIP-Seq, DNase-seq, FAIRE-seq, histone ChIP-Seq, MNase-seq, RRBS), the ENCODE project has identified and characterized the following functional elements from 147 different cell types: RNA transcribed regions, protein-coding regions, TF binding sites, chromatin structure and DNA methylation (Consortium, 2012; Qu & Fang, 2013).

The ENCODE project also investigated the potential effect of DSVs in the function of genomic functional elements. For protein-coding genes, DSVs that are likely to disrupt splice-sites, introduce frame-shift or lead to protein truncation (premature stop of translation) were found. Variants that are in TF binding regions and that are likely to disrupt binding were also identified (Consortium, 2012).

The functional understanding of rDSVs that affect splice-sites and TF binding was greatly enhanced by the ENCODE project. Furthermore, these two categories of rDSVs have been widely investigated by other groups and several computational tools to predict their effect have been created.

Despite the ample efforts to characterize rDSVs, variants that can potentially alter miRNA binding have not been widely studied. The reason for such are two: 1) complicated nature of miRNA binding, which don't follow as strict rules as TF binding; 2) technology difficulties in experimentally determining miRNA biding sites, despite the newly developed NGS-based assays PAR-CLIP (Hafner et al., 2010), HITS-CLIP (Licatalosi et al., 2008) and iCLIP (Konig et al., 2010).

Recently, computational tools for predicting the effect of rDSV's in miRNA binding sites were created. However, these tools have the following problems: 1) they are mostly databases that host a collection of pre-computed predictions. Thus, they do not allow users to run predictions in a set of customized rDSVs, making them not practical for the analysis of large genomics datasets; 2) results of available methods are difficult to interpret, making rDSV prioritization a hard task.

1.1.2 Functional evaluation and prioritizing of DSVs in genomic datasets

Genomic and clinical genomic datasets focus on protein-coding variants, mainly because functional interpretation of coding variants is easier. (Consortium, 2012; Macarthur, 2012). In contrast to coding DSVs, variants in the non-coding regions of the genome usually have small or undetectable impacts on gene expression due to functional redundancy. Therefore, interpretation of rDSVs in a genomic-scale data poses great challenge (Macarthur, 2012).

The decreasing cost of genomic assays as well major public initiatives like The Cancer Genome Atlas (TCGA), ENCODE, and The International Cancer Genome Consortium (ICGC) lead to a vertiginous increase in the amount of publicly available genomic datasets (Kannan et al., 2016).

One of the findings of the ENCODE project is that RNA production and processing is quantitatively correlated with transcription factor binding at gene promoters and with the state of the chromatin (Consortium, 2012). These findings suggest that regulation of gene expression is a complex process, and that correct functional prediction of the effect of a DSV should consider the current state of the cell (i.e. which TFs, genes and miRNAs are expressed as well as the state of chromatin). The most accurate way to determine the current state of a cell is to integrate various types of genomic data, which is, fortunately, possible today due to the wealth of public data available. However, most of analysis methods and pipelines available today are designed to take into consideration only one type of genomic data, which lead in many cases to wrong functional interpretation of a DSV.

1.2 THE APPROACH

This dissertation address two main problems in the current field of computational genomics: 1) the lack of computational methods capable of accurately predicting the effect of rDSVs on miRNA binding and 2) the inexistence of computational pipelines that can functionally predict the effect of DSVs considering the current state of cells.

The first problem is addressed by proposing a new algorithm called is-mirSNP. It uses binding energy as a measure of how strong a mRNA will bind to miRNA. The binding energy of two different sequences with a given miRNA are calculated. One of sequences contains the reference allele, whereas the other has the DSV of interest. The binding energy difference between the two mRNA-miRNA pairs are compared, and if statistically significant the DSVs is considered to alter miRNA binding. The is-mirSNP algorithm can be divided in two main parts: a) estimation of background binding energy distributions, b) scoring the effect of a DSVs in miRNA binding.

In the first step, two different background distributions are empirically estimated: background binding distribution, and the distribution of log ratios of binding energy p-values. Each of the two background distributions are calculated for each miRNA. The estimation of the background distributions need to happen only once, and are required by the scoring of a DSV.

First, the background binding distribution is calculated. This distribution is comprised of the binding energies of a given miRNA with all 3'UTR sequences in the reference genome. After this is completed, the distribution of log ratios of p-values is computed. First, two binding energies are calculated: a) the binding energy of a miRNA with a sequence containing the reference alleles, and b) the binding energy of a miRNA with a sequence containing the DSVs of interest. For each of these energies, a p-value is obtained using the previously computed background energy distribution. Then, the log-ratio of the two p-values is computed. For the computation of the distribution of log-ratios of binding energy p-values, this process is repeated 50,000 different known 3'UTR DSVs.

The second step of the algorithm consists of scoring a DSV. The main purpose of scoring a DSV is to evaluate how it would affect miRNA. For such, the binding energy of a DSV is compared to the binding energy of a reference allele. The magnitude of the energy difference between alleles is important, but not meaningful if it does not translate into actual binding. A great proportion of DSVs will change miRNA binding energy to some extent, but only a small portion will have a positive or negative effect on binding. The scoring step of the algorithm works in the following way: 1) Sequences for all possible k-mers are generated: one containing the reference allele and another containing the alternative allele; 2) Binding energy for all k-mers are calculated; 3) The lowest binding energy for k-mers containing reference allele, and lowest binding energy for k-mers containing the DSV allele are kept; 3) P-values for lowest binding energy for reference and DSV is computed from background distribution; 4) Log p-value ratio is calculated; 5) P-value of log p-value ratio is calculated using the distribution of log ratios of binding energy p-values. This process is initially done for the following k-mers: 7-mers, 8-mers and l-mers (where l is the length of the miRNA being tested). The algorithm stops whenever the p-value of the log-pvalue ratio is significant (p-value < 0.05), otherwise results for the last step (l-mer evaluation) are reported. The is-mirSNP algorithm gives the users a p-value for each DSV that is evaluated. In addition, the algorithm reports the binding energies for each allele and their p-values, which greatly helps the interpretation of the results.

The second problem addressed in this dissertation is the lack of analysis pipelines that can functionally annotate DSVs. The pipeline proposed here integrates GWAS with gene expression data (RNA-Seq), and functionally annotates inferred DSVs. The pipeline has 3 steps: 1) RNA-Seq analysis; 2) DSV Annotation; 3) Integration of GWAS with expression data.

The first step in the pipeline is RNA-Seq analysis. RNA-Seq analysis starts by assessing the quality of data. Metrics utilized for quality-control are number of reads, and overall reads PHRED scores. Reads with low quality are either trimmed (if low quality regions are only in the end and/or beginning of reads) or discarded. At this step, removal of adapter sequences is done if applicable. After the quality-control step, reads are mapped to the reference genome, and gene and isoform expressions are quantified. Next, tests for differentially expressed genes are performed. The number of tests or which tests that need to be performed are defined by the biological questions that need to be answered as well as the experimental design.

The second step consists of finding DSVs that are inside or in the surroundings of differentially expressed genes. DSVs in the exons, introns, 3'UTR, 5'UTR and promoter regions (5kp upstream and 5kb downstream) are identified.

Next, previously identified DSVs are further filtered. Tag-DSVs (DSVs that are statistically associated with the complex disease in question) are obtained from relevant existing GWAS studies. Then, linkage-disequilibrium (LD) analysis are run between tag-DSVs and identified DSVs. DSVs with a LD score > 0.8 are considered to be in LD with a Tag-SNP and are functionally annotated.

Functional annotation of DSVs will depend on the region where a DSV is found: DSVs in exons are functionally annotated for their potential impact in protein sequence and structure, DSVs in introns are annotated for splice-sites and effect in TF binding, DSVs in 5'UTR and promoter regions are annotated for effect in TF binding, and DSVs in 3'UTR are annotated for effect in miRNA binding. In the assessment of TF binding only expressed in the RNA-Seq dataset TFs are considered. DSVs that are in LD with a Tag-SNP and are predicted to have any function are reported.

After implementation, the analysis pipeline just described will be tested on a RNA-seq datasets that mimics the behavior of Th17 lymphocytes in inflammatory bowel disease patients.

1.2.1 THESIS

The main thesis of this dissertation is that the new algorithms and methods presented here can correctly identify, assign function to, and prioritize DSVs.

Using manually curated validation datasets and a novel RNA-seq dataset, the following conjectures will be evaluated:

Claim 1. The is-miRSNP algorithm is capable of correctly identifying validated DSVs known to affect miRNA binding. The algorithm presented here not only performs better than existing tools, but the results obtained are easier to interpret. The algorithm can be used to evaluate novel and existing data.

Claim 2. Our pipeline for identification of functional DSVs can identify functional relevant variants from RNA-Seq data, when integrated to existing GWAS data. Results obtained from our pipeline when applied to a novel RNA-Seq experiment leads to biologically meaningful results.

1.3 SIGNIFICANCE

This section discusses the significance of the work if the above hypothesis and claims are supported by experimental evidence. The is-mirSNP algorithm deals with the problem of predicting whether a DSV can affect a miRNA binding site. From a bioinformatics stand point, we present a novel algorithm that uses binding energy to predict the effect of a given DSV. The novelty of our approach consists of computing the background binding energy distributions and the background distributions of the log-ratio of the p-values of a pair of binding energies. This approach allows for the calculation of p-values which are easy to interpret. The algorithm is fast and robust. It can be easily integrated into existing analysis pipelines, and can be used to predict DSVs found by large genomic datasets. From a medical perspective is-mirSNP will allow scientists to accurately assign function to a vast number of rDSVs that would not have been functionally annotated otherwise. Therefore, predictions done by is-mirSNP algorithm will allow for more accurate interpretation of clinical and research genomic datasets, which in turn can lead to the elucidation of molecular mechanisms of diseases.

A pipeline that analyzes and functionally annotates DSVs is also proposed. It consists of a new bioinformatics approach to the analysis and integration of genomic datasets. The pipeline puts together existing computational tools, and utilizes existing methods in a novel way. It can leverage and interpret existing data. From a biomedical stand-point the analysis pipeline described can elucidate and ascribe function to DSVs that would otherwise not have been associated to disease. Identification of such variants have several potential clinical benefits as such DSVs can be used for new diagnostic tests, to better estimate prognostics or even be targets of new therapeutics.

1.4 DISSERTATION OVERVIEW

This dissertation has the following structure:

Chapter 2 describes in detail our algorithm for identification and prioritization of DSVs that can affect miRNA binding sites. It also describes datasets used for algorithm evaluation and comparison to existing tools.

Chapter 3 provides detailed information about our pipeline for identification of functional relevant variants from RNA-Seq data, when integrated to existing GWAS data

Chapter 4 presents the application of the pipeline described in Chapter 3 to an effector memory T cell RNA-Seq dataset that is related to inflammatory bowel disease.

2.0 PREDICTION OF MIRNA BINDING SITES

2.1 BACKGROUND

MicroRNA (miRNAs) are endogenous non-coding small RNAs that play a very important role in the regulation of gene expression. The principal characteristics of miRNAs are their length of approximately 22 nucleotides and the fact that they are produced by two RNaseIII proteins, Drosha and Dicer (Ha & Kim, 2014).

The miRNA biogenesis generally follows a 'linear' pathway, although miRNA-specific modifications to this pathway are not uncommon (Winter, Jung, Keller, Gregory, & Diederichs, 2009). The 'linear' canonical pathway starts with the production of a primary miRNA transcript called the pri-miRNA. Then, this pri-miRNA is cleaved in the nucleous by the Pasha (Drosha-DGCR8) complex. The cleaved product, now called a pre-miRNA, is exported to the cytoplasm where it is cleaved one more by the RNase Dicer, resulting in a mature miRNA with a length of approximately 22 nucleotides. The miRNA is loaded together with the Argonaute (AGO2) protein into the RISC complex where it determines by complimentary which messenger RNA (mRNA) will be degraded (Winter et al., 2009).miRNAs play an important role in regulation of gene expression with at least 60% of all genes being regulated by these small non-coding RNAs

(Friedman, Farh, Burge, & Bartel, 2009). Several studies have showed the importance of miRNAs in diseases like cancer and neurodevelopmental disorders (Ha & Kim, 2014)

The recognition of miRNA binding-sites follows certain rules but several exceptions have been reported (Clark et al., 2014). miRNA binding usually occurs in the 3' untranslated region (3'UTR) of mRNAs. For canonical miRNA binding, maximum complementary of a short 5' region on the miRNA to its target is necessary (Ellwanger, Buttner, Mewes, & Stumpflen, 2011; Landi, Barale, Gemignani, & Landi, 2011). This short region usually spans positions 2-8 of the 5' end of the miRNA and is known as the 'seed' (Clark et al., 2014). A maximum of one mismatch between the seed and its target is tolerated. The pairing of G with T is also allowed in the mRNA-miRNA duplex (Landi et al., 2011). Ellwanger et al. investigated the minimal set of seeds needed for miRNA binding. They identified a set of 6 seed types, which consist of seeds from six to eight nucleotides long and with mismatches at different positions. Interestingly, the majority of functional seeds were six nucleotides long, but most target prediction algorithms focused on seven of eight nucleotide long seeds (Ellwanger et al., 2011). In addition to canonical seed interactions, non-standard binding in which contiguous base pairing is interrupted by bulges has been reported. 'Seed-less' interactions of targets located outside the 3'UTR which were non conserved amongst various species has also been observed (Clark et al., 2014) .

NGS technologies allowed the high-throughput study of miRNAs and their effects in a large scale. Utilizing methods like PAR-CLIP, HITS-CLIP and iCLIP researchers can investigate miRNA-target binding in a whole-genome scale (Danan, Manickavel, & Hafner, 2016; Darnell, 2010; Konig et al., 2011) . However, the enormous amount of data generated by these technologies creates data analysis challenges, which require novel computational tools (T. Wang et al., 2015).

Several algorithms that predict miRNA binding sites have been developed. These algorithms look for features such as conservation of binding sites, patterns in seed matching, and energetic stability of the miRNA-mRNA pair (Landi et al., 2011). TargetScan checks the conservation of sites among different species in conjunction with the thermodynamic stability of the miRNA-mRNA pair (Lewis, Burge, & Bartel, 2005). Miranda utilizes thermodynamic stability, base pairing information and conservation information to assess miRNA-mRNA pairing (Enright et al., 2003). RNAhybrid calculates the optimal energy between the two RNA sequences (Rehmsmeier, Steffen, Hochsmann, & Giegerich, 2004). MicroInspector uses a sliding window to check for complementary between the miRNA and its target (Rusinov, Baev, Minkov, & Tabler, 2005). Pictar checks for complementary in the miRNA seed region, calculates the binding energy and calculates the likelihood of binding using a Hidden Markov Model (Krek et al., 2005). Diana-microT not only calculates the binding energy, but also considers conservation and biological pathway information (Maragkakis et al., 2009).

Single-nucleotide polymorphisms can affect the binding between miRNA and mRNA by either increasing or decreasing the binding energy of a given miRNA and mRNA (Landi et al., 2011). Recently, several databases that contain SNPs, which can potentially alter miRNA binding sites, were created. The most widely used tools are PolymiRTS (Ziebarth, Bhattacharya, Chen, & Cui, 2012), mirSNP (C. Liu et al., 2012), mrSNP (Deveci, Catalyurek, & Toland, 2014), mirsnpscore (Thomas, Saito, & Saetrom, 2011) and miRNASNP2 (Gong et al., 2015). These databases utilize the previously described miRNA target prediction tools to detect the formation or disruption of binding sites. All the databases follow the same approach: they scan the entire genome against a SNP database like dbSNP and store the entire results. Although computationally efficient this approach only permits users to retrieve results for SNPs that are already known. Users cannot obtain predictions for novel DSVs that are not available in databases but present in their datasets. Pre-computation of results usually become outdated, if databases are not diligently updated. This is especially true as more genomic data is being generated at a faster pace. Another problem with pre-computed, web-tools is that querying large amounts of SNPs is usually difficult or impossible. Integration of such tools with existing pipelines for analysis of whole genome data is also not possible. Deveci et al. tried to address these issues by developing a tool capable of handling custom queries, so that new SNPs could be assessed (Deveci et al., 2014). However, mrSNP still suffers of main drawbacks. First, it does not provide users with a score or probability of a SNP altering a miRNA-binding site. Second, it is only capable of scanning 3'UTR and searching for SNPs that alter canonical binding. It is important to note that the mrSNP website has been down since the beginning of this work, therefore it has not been possible to obtain a copy of the code.

Most the algorithms and tools created for predicting the effect of DSVs in miRNA binding are web-based and work like a database where users can query pre-computed results. Usually, these algorithms are run for a specific version of a SNP database like dbSNP. There is also the problem with the interpretation of the results obtained from existing tools: most of the time the meaning of the predictions are not clear as no intuitive score is generated. This poses a challenge for users trying to prioritize and rank a great number of results obtained from genomic analysis.

2.2 METHODS

Our algorithm uses a principle that is similar to the one presented by is-rSNP (Macintyre, Bailey, Haviv, & Kowalczyk, 2010). Is-rSNP is an algorithm designed to predict the effect of regulatory DSVs on transcription factor binding sites. Briefly, it estimates the background binding distributions of a transcription factor from its position-scoring matrix (PWM). Sequences containing the alleles of interest are scored using the very same PWM. The two scores are compared and statistical significance is calculated.

The nature of TF binding and miRNA binding is different, with miRNAs showing more binding motif variation which does not allow for the computation of PWMs. The is-mirSNP algorithm takes a unique approach to the statistical estimation of miRNA binding variation. This approach allows the proper separation between binding and non-binding, and the accurate estimation of the effects of a DSV. Binding energy is used as a measure of direct miRNA-binding, and the background distributions are empirically estimated by a novel algorithm (Material and Methods sections 2.2.1 and 2.2.2). The scoring method of is-rSNP and our algorithm follow the same principle; both methods calculate scores for sequences with and without the DSV. The scores are then compared and statistical significance is calculated using the pre-computed background distributions (Material and Methods sections 2.2.3).

2.2.1 Computation of miRNA binding energy background distributions

The first step in the prediction of the functional effect of a DSV in miRNA binding is the calculation of the background binding energy distribution for a given miRNA. The background energy distribution consists of the probabilities that a particular binding energy is observed. The use of binding energy as a indicative of miRNA binding is a proven method and has been used by others (Coronnello et al., 2012; Enright et al., 2003). To compute such distributions, we calculate all possible binding energies between every existing k-mer and every miRNA. Only k-mers found in the 3'UTR of the genome were used since the clear majority of binding happens in this region. Limiting the k-mers to those found in the 3'UTR region also reduces the computational time necessary for the computation of the distributions. Counting of all existing k-mers in the human 3'UTR sequences was done using Jellyfish (Marcais & Kingsford, 2011). It has been shown that in most binding events a perfect match of the target sequence with the miRNA seed sequence is necessary. This binding characteristic was captured by computing two different background distributions for a given miRNA: one for 7-mers seeds and another for 8-mers seeds. To capture seedless miRNA-target binding, a third distribution that models the binding of the entire mature miRNA was computed. In this case, the size of the k-mer is the same as the miRNA of interest (Figure 1). Binding energies were calculated using RNAhybrid (Rehmsmeier et al., 2004).

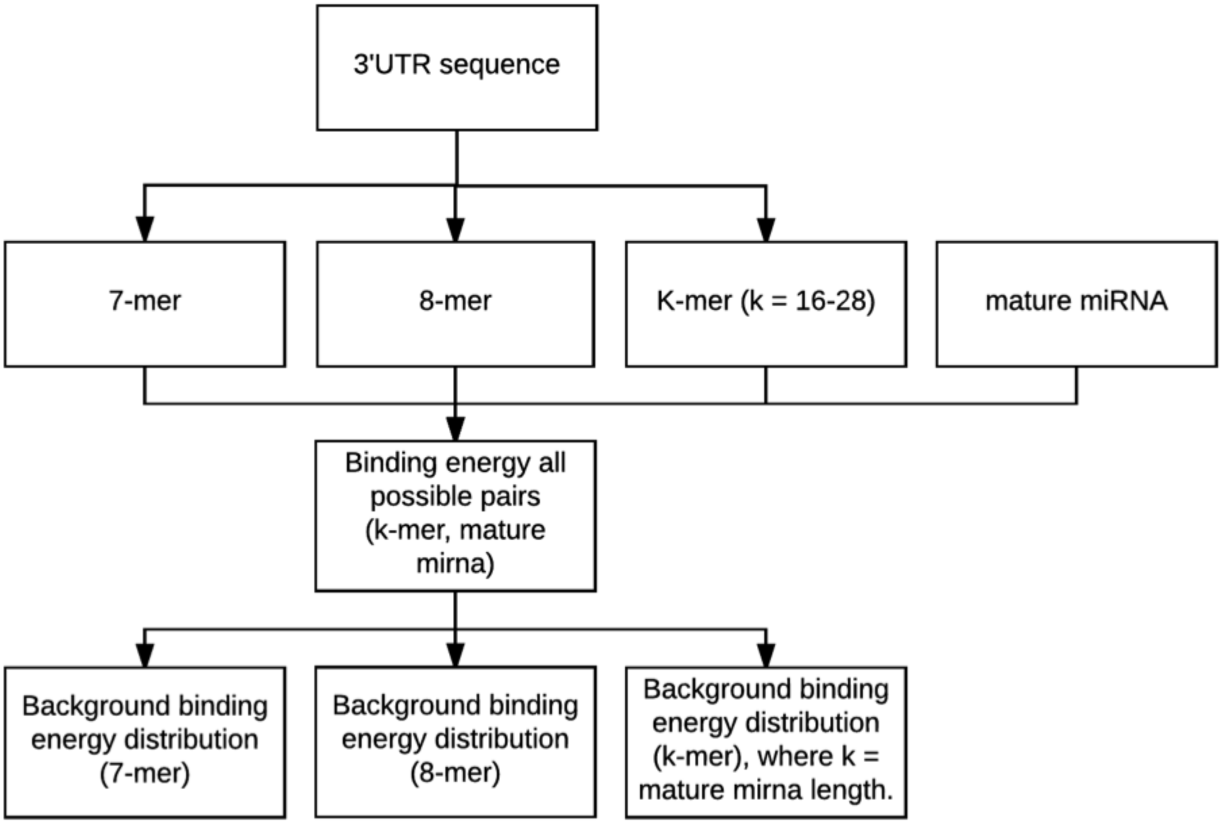

3'UTR sequence
7-mer
8-mer
K-mer (k = 16-28)
mature miRNA
Binding energy all possible pairs (k-mer, mature mirna)
Background binding energy distribution (7-mer)
Background binding energy distribution (8-mer)
Background binding energy distribution (k-mer), where k = mature mirna length.

Figure 1. Algorithm for the computation of binding energy background distribution

2.2.2 Computation of the log-ratio background distributions

The second step is the calculation of the background binding energy distribution that assesses how likely a SNP will affect binding energy. The computation of these distributions start from randomly sampling 50,000 known SNPs that are in the 3'UTR of the human genome. It is important to note that the selection of 50,000 known SNPs was determined after experimenting with SNP datasets of various sizes (10000, 20000, 30000, 40000, 50000, 70000 and 100000). The 50,000 SNP dataset was selected based on convergence and computational time (in future runs). For every SNP pair two sets of sequences are created: one that contains all possible sequences with the reference allele and another set with all possible sequences with the mutant allele. The binding energy between every sequence in both sets and a miRNA is calculated. Reference and mutated k-mers with the lowest binding energy are kept, and their p-values are calculated using the previously calculated background distributions. Finally, a log ratio of the binding energy p-values is calculated (Figure 2). This approach is repeated for 7-mer seeds, 8-mer seeds, and mature miRNAs independently.

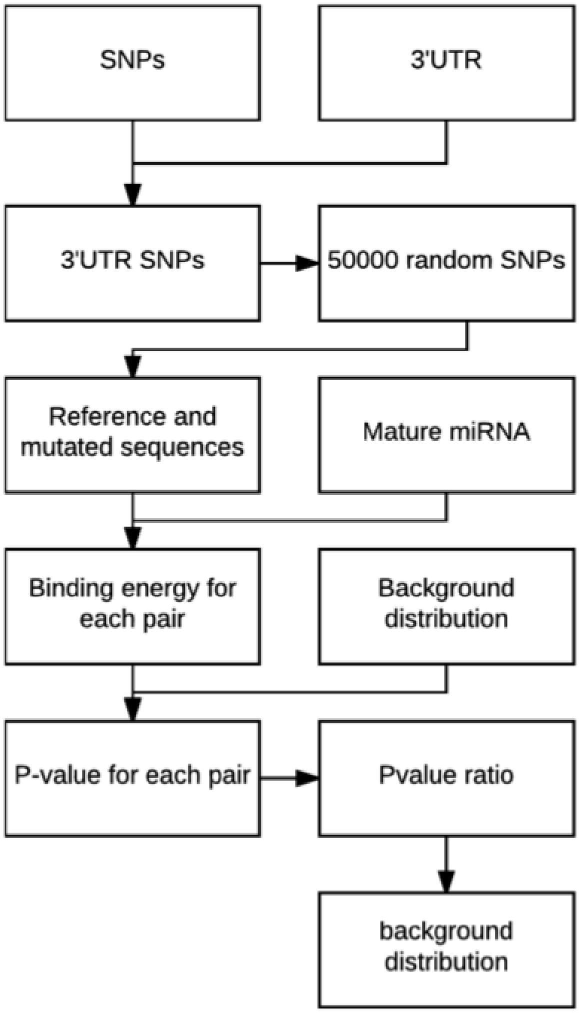


Figure 2. Algorithm for the computation of p-value log-ratio background distribution

2.2.3 Scoring the effect of a DSV in miRNA binding

A given DSV is scored for all possible known miRNAs. First, all possible 7-mers that contain the DSV being scored are created. These sequences are created based on the genomic region where the DSV is found. Therefore, two sets of possible sequences are created: one that contains the reference allele and another that contains the mutant allele. Each set contains all possible sequences, of the same length, that have the DSV of interest. Energy between all created sequences and the miRNA 7-mer seed is calculated. The mutant and reference sequences with the lowest binding energy are kept. Using the binding energy background distribution, a p-value is calculated for the mutant and reference sequence energies. Here, we assume that a miRNA will most likely bind at the position which will confer the most stable conformation, hence the lowest binding energy. Next, the log ratio of the two p-values is calculated. The p-value for the log ratio is then calculated using the background log ratio distribution. If this value is less than 0.05, then the SNP is considered to affect miRNA binding and the algorithm stops. If this not the case, then the procedure is repeated for 8-mer seed. If the p-value for the scoring of the 8-mer is greater than 0.05, then the process is repeated using the entire miRNA length (Figure 3).

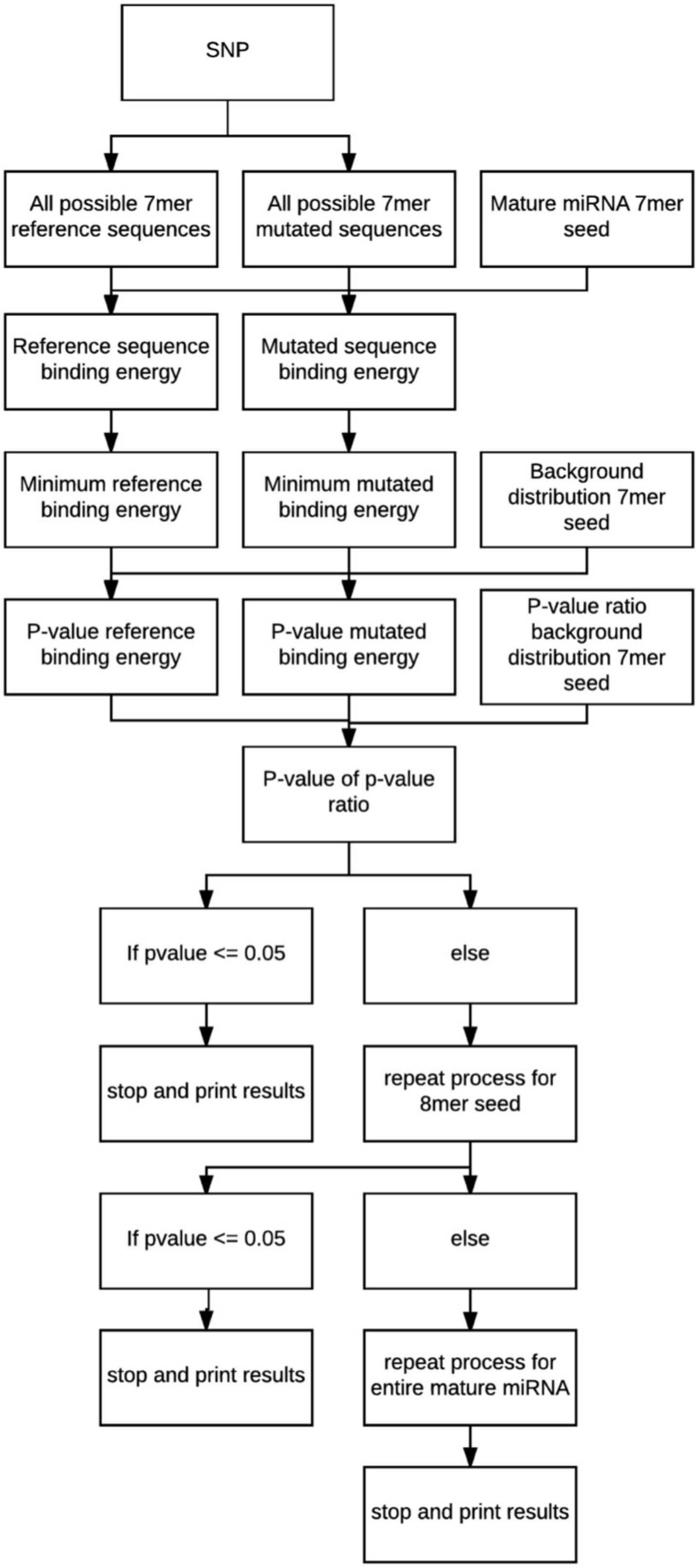

SNP
All possible 7mer reference sequences
All possible 7mer mutated sequences
Mature miRNA 7mer seed
Reference sequence binding energy
Mutated sequence binding energy
Minimum reference binding energy
Minimum mutated binding energy
Background distribution 7mer seed
P-value reference binding energy
P-value mutated binding energy
P-value ratio background distribution 7mer seed
P-value of p-value ratio
If pvalue <= 0.05
else
stop and print results
repeat process for 8mer seed
If pvalue <= 0.05
else
stop and print results
repeat process for entire mature miRNA
stop and print results

Figure 3. is-mirSNP scoring algorithm

2.2.4 Validation dataset

The first step in the evaluation of the performance of our algorithm was the creation of a *bona fide* validation dataset. The validation dataset only contains SNPs that were experimentally proven to affect miRNA binding (Adams, Furneaux, & White, 2007; Cheng et al., 2013; Elek et al., 2015; Feng et al., 2012; Goda, Murase, Kasezawa, Goda, & Yamakawa-Kobayashi, 2015; Landi et al., 2012; S. Y. Lee et al., 2015; Lin et al., 2012; H. Liu et al., 2016; Z. Liu et al., 2011; Menard et al., 2016; Nicoloso et al., 2010; Sethupathy et al., 2007; Stegeman et al., 2015; Tang et al., 2015; K. Wang et al., 2012; Wynendaele et al., 2010; Xiong et al., 2011; Xu et al., 2013; L. Yang et al., 2012; Yousef, 2015; J. Zhang et al., 2014; S. Zhang et al., 2013; Y. Zhang et al., 2016; Zu et al., 2013). The process of creating this set of miR-SNPs consisted of manually curating papers and selecting those that had direct experimental evidence of affecting miRNA binding. Computational prediction or indirect evidence of miRNA binding was not accepted, and these SNPs were not included in our validation dataset.

2.2.5 Performance comparison

Our validation dataset was used to assess the performance of our algorithm as well as other existing miRSNP prediction tools. We compared our performance with the following tools: MirSNP (C. Liu et al., 2012), PolymiRTS 3.0 (Bhattacharya, Ziebarth, & Cui, 2014), mirsnpscore (Thomas et al., 2011) and miRNASNP2 (Gong et al., 2015). The mrSNP tool could not be evaluated as its webserver has been offline for the entire time that this paper has been in preparation (Deveci et al., 2014). It is also worth noting that according to the mrSNP paper, it has a slightly worse performance than mirSNP, therefore the exclusion of this tool from our validation set does not introduce any bias or invalidate our results (Deveci et al., 2014).

2.3 RESULTS

2.3.1 Background distributions

We hypothesized that miRNA-mRNA binding energy for *bona fide* binding sites is statistically significant when compared to a random distribution of binding energies. To test this hypothesis, the mRNA-miRNA binding energies for known miRNA binding sites was calculated. The true binding sites were obtained from miRTarBase (Chou et al., 2016; Hsu et al., 2011) , and only sites validated using direct experimentation were used. Figure 2 shows the binding energy distribution with all validated targets for 4 distinct miRNAs: let-7f-5p, miR-98-5p, miR-103a-3p, and miR-146a-5p. Note that miRNA validated targets are in the left-side of the distribution and show low binding-energy. The median p-value for the binding energy of 5,369 validated targets across 482 human miRNAs is 0.0021.

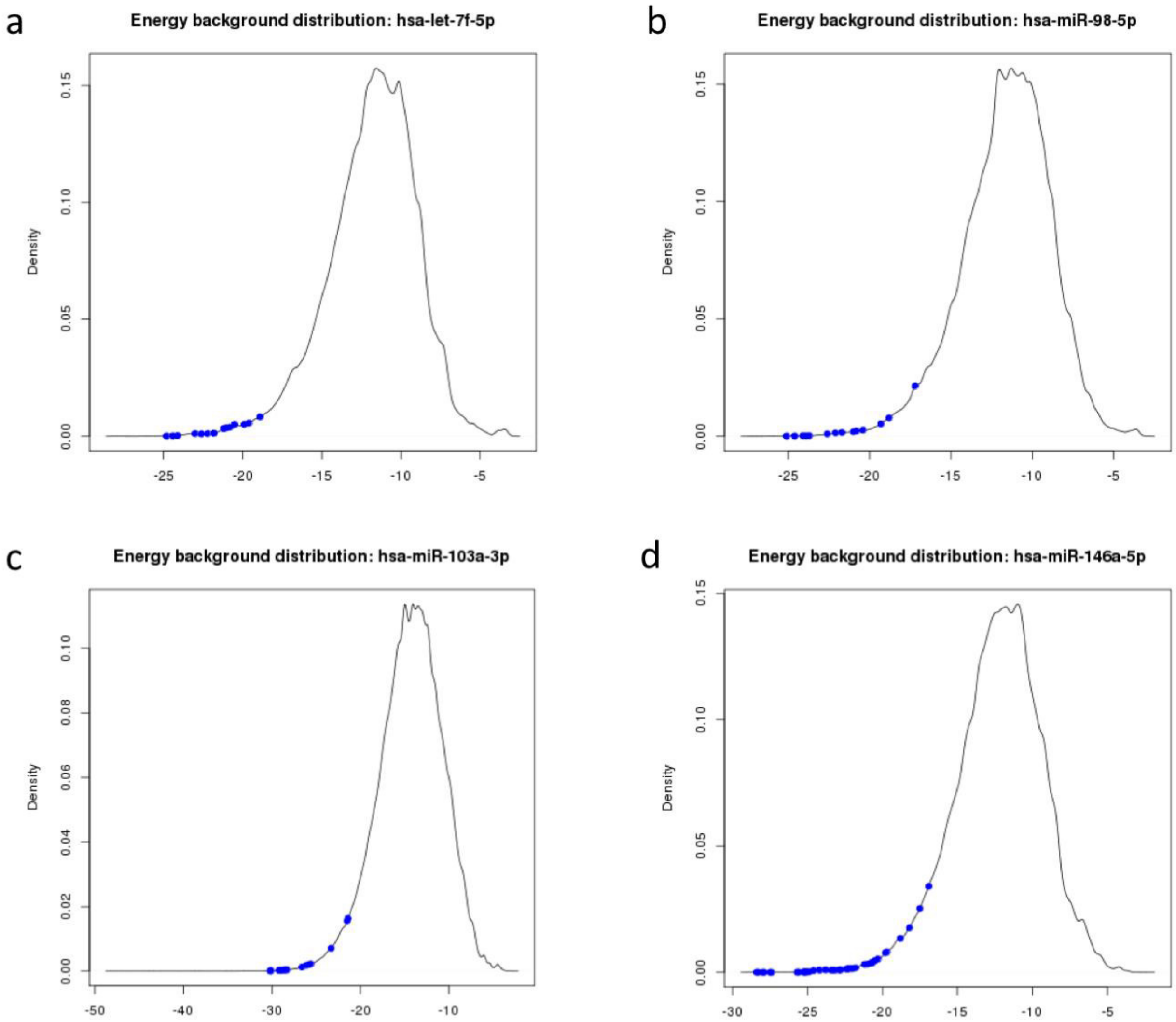


Figure 4. Background energy distributions with known miRNA targets (blue dots)

2.3.2 Performance and comparison to existing tools

The creation or destruction of a miRNA binding site by a SNP should alter the miRNA-mRNA binding energy. This idea is supported by our previous results that showed the correlation between energy and the presence of true binding. Therefore, we implemented and tested our miRNA SNP prediction algorithm, called is-mirSNP, using SNPs that are known to affect miRNA binding. We manually curated from the literature a set of 27 experimentally validated pairs of miRNA-SNPs known to affect miRNA binding sites. In order to reduce the number of false-positives, only miRNA-SNP pairs shown to alter binding sites through direct experimentation (luciferase assays, gene-reporter assays) were used. Our algorithm was capable of accurately predict the effect of 23/27 miRNA-SNPs pairs (85%) (Table 1). Our algorithm performed better than miRSNP, PolymiRTS 3.0, mirsnpscore and miRNASNP2 that were capable of accurately predict only 18 (66.6%), 6 (22.2%), 10 (37.03%) and 10 (37.03%) miRNA-SNP pairs, respectively (Table 1). The miRNA-SNP pairs that our algorithm failed to correctly predict, showed minor or no changes in the binding energy between the alleles. SNP rs7911488 showed a difference in binding energy between alleles that was not great enough to be considered significant by our algorithm. Most the SNPs that were not correctly classified by is-miRSNP were also missed by all other algorithms. The only exception was SNP rs2240688 which was correctly classified by miRSNP, mirsnpscore and miRNASNP2.

Table 1 Detailed is-rSNP results for DSVs in the validation set. MFE1/2: Minimum Free Energy for allele-1/2; p-value: MFE p-value; LR: log-ratio of p-value(1)/p-value(2); p-value(LR): the p-value of the log-ratio (LR).

RS ID	Gene	miRNA	Prediction	Allele1	MFE1	P-value(1)	Allele2	MFE2	P-value(2)	LR	P-value(LR)
rs5186	AGTR1	has-miR-155-5p	7mer Seed	A	-8.8	0.027	C	-4.5	0.608	3.106	0.000
rs7911488	BCL2	has-miR-1307-3p	miRNA	A	-23.9	0.024	G	-24.5	0.032	0.278	0.097
rs2239680	BIRC5	has-miR-335-5p	7mer Seed	T	-9.6	0.010	C	-5.5	0.310	3.439	0.001
rs1434536	BMPR1B	has-miR-125b-5p	miRNA	C	-21.5	0.010	T	-17.5	0.066	1.908	0.014
rs7213430	BRIP1	has-miR-101-3p	8mer Seed	G	-12.1	0.001	A	-9.8	0.018	2.858	0.029
rs2240688	CD133	has-miR-135b-3p	miRNA	T	-16.1	0.261	G	-16.1	0.261	0.000	0.584
rs9341070	ESR1	has-miR-206	7mer Seed	C	-8.3	0.037	T	-6.1	0.312	2.142	0.034
rs1063320	HLA-G	has-miR-148a-3p	7mer Seed	C	-9.2	0.064	G	-11.9	0.005	2.543	0.006
rs2229295	HNF1B	has-miR-214-5p	7mer Seed	G	-9.9	0.052	T	-13.2	0.003	2.904	0.016
rs2229295	HNF1B	has-miR-550a-5p	8mer Seed	G	-13.3	0.012	T	-15.2	0.002	1.939	0.039
rs56109847	HTR3E	has-miR-510-5p	7mer Seed	G	-10.9	0.449	A	-5.4	0.704	4.521	0.000
rs709805	KIAA0182	has-miR-324-3p	miRNA	G	-30.1	0.001	A	-25.7	0.007	2.178	0.011
rs1058205	KLK3	has-miR-3162-5p	7mer Seed	C	-11.4	0.014	T	-12.4	0.003	1.504	0.045
rs3660	KRT81	has-mir-17-5p	8mer Seed	C	-11.6	0.011	G	-8.9	0.083	2.043	0.010
rs3660	KRT81	has-mir-20b-5p	8mer Seed	C	-11.6	0.011	G	-8.9	0.083	2.043	0.010
rs4245739	MDM4	has-miR-191-5p	7mer Seed	C	-11.2	0.003	A	-5.1	0.688	5.458	0.000
rs12537	MTMR3	has-miR-181a-5p	8mer Seed	C	-5.7	0.222	T	-8.7	0.009	3.184	0.017
rs3134615	MYCL1	has-miR-1827	7mer Seed	C	-14.6	0.001	A	-10.4	0.049	3.736	0.005
rs2735383	NBS1	has-miR-629-5p	miRNA	C	-14.9	0.129	G	-17.1	0.297	0.837	0.144
rs6573	RAP1A	has-miR-196a-5p	7mer Seed	C	-7	0.220	A	-11	0.004	3.913	0.002
rs465646	REV3L	has-miR-25-3p	7mer Seed	G	-9.4	0.019	A	-7.4	0.159	2.129	0.035
rs465646	REV3L	has-miR-32-5p	7mer Seed	G	-9.4	0.019	A	-7.4	0.159	2.129	0.035
rs16917496	SET8	has-miR-502-5p	miRNA	C	-16.2	0.259	T	-16.2	0.420	0.481	0.233
rs334348	TGFBR1	has-miR-628-5p	miRNA	A	-20.4	0.001	G	-24.6	0.000	3.107	0.003

RS ID	Gene	miRNA	Prediction	Allele1	MFE1	P-value(1)	Allele2	MFE2	P-value(2)	LR	P-value(LR)
rs8126	TNFAIP2	miR-184	7mer Seed	C	-13.6	0.002	T	-11.4	0.015	2.118	0.037
rs1010	VAMP8	miR-370-3p	miRNA	T	-25.6	0.030	C	-30.6	0.001	3.101	0.003
rs9457	WFS1	miR-185-5p	8mer Seed	G	-10.4	0.015	C	-12.7	0.002	2.134	0.026

Table 2. Performance comparison of is-miRSNP and other 4 prediction tools. Yes or No values indicates if DSV-miRNA pair was recovered or not by the corresponding algorithm.

RS ID	**miRNA**	**is-mirSNP**	**miRSNP**	**PolymiRTS**	**mirsnpscore**	**miRNASNP2**	**References**
rs5186	miR-155-5p	Yes	Yes	Yes	No	No	(Sethupathy et al., 2007)
rs7911488	miR-1307-3p	No	No	No	No	No	(Tang et al., 2015)
rs2239680	miR-335-5p	Yes	Yes	No	Yes	Yes	(Zu et al., 2013)
rs1434536	miR-125b-5p	Yes	Yes	No	Yes	No	(Feng et al., 2012)
rs7213430	miR-101-3p	Yes	Yes	No	Yes	No	(H. Liu et al., 2016)
rs2240688	miR-135b-3p	No	Yes	No	Yes	Yes	(Cheng et al., 2013)
rs9341070	miR-206	Yes	No	No	No	No	(Adams et al., 2007)
rs1063320	miR-148a-3p	Yes	Yes	No	Yes	Yes	(Menard et al., 2016)
rs2229295	miR-214-5p	Yes	Yes	No	No	Yes	(Goda et al., 2015)
rs2229295	miR-550a-5p	Yes	Yes	No	No	Yes	(Goda et al., 2015)
rs56109847	miR-510-5p	Yes	Yes	No	No	No	(Y. Zhang et al., 2016)
rs709805	miR-324-3p	Yes	No	No	No	No	(Landi et al., 2012)
rs1058205	miR-3162-5p	Yes	Yes	Yes	Yes	Yes	(Yousef, 2015)
rs3660	mir-17-5p	Yes	Yes	Yes	Yes	Yes	(S. Y. Lee et al., 2015)
rs3660	mir-20b-5p	Yes	Yes	Yes	Yes	Yes	(S. Y. Lee et al., 2015)
rs4245739	miR-191-5p	Yes	Yes	No	Yes	Yes	(Wynendaele et al., 2010)
rs12537	miR-181a-5p	Yes	Yes	No	No	Yes	(Lin et al., 2012)
rs3134615	miR-1827	Yes	Yes	No	Yes	No	(Xiong et al., 2011)
rs2735383	miR-629-5p	No	No	No	No	No	(L. Yang et al., 2012)
rs6573	miR-196a-5p	Yes	No	No	Yes	No	(K. Wang et al., 2012)
rs465646	miR-25-3p	Yes	Yes	No	No	No	(S. Zhang et al., 2013)
rs465646	miR-32-5p	Yes	Yes	No	No	No	(S. Zhang et al., 2013)
rs16917496	miR-502-5p	No	No	No	No	No	(Xu et al., 2013)

RS ID	miRNA	is-mirSNP	miRSNP	PolymiRTS	mirsnpscore	miRNASNP2	References
rs334348	miR-628-5p	Yes	No	No	No	No	(Nicoloso et al., 2010)
rs8126	miR-184	Yes	Yes	Yes	No	Yes	(Z. Liu et al., 2011; J. Zhang et al., 2014)
rs1010	miR-370-3p	Yes	No	No	No	No	(Stegeman et al., 2015)
rs9457	miR-185-5p	Yes	No	Yes	No	No	(Elek et al., 2015)

.

2.3.3 Discussion

We developed a new algorithm, is-miRSNP, that can accurately predict the effect of a SNP on a miRNA binding site. Our tool addresses several issues that are not covered by existing tools. First, is-mirSNP can be easily deployed as a standalone software or be integrated into existing pipelines. It also can predict the effect of existing and novel SNPs. We also addressed the issue of variant prioritization by associating the difference and strength of miRNA binding sites with p-values, concepts which most researchers and scientists are familiar with.

The performance of is-mirSNP was compared with other tools, using a set of 27 experimentally validated miRNA-SNP pairs. To reduce bias and remove false-positives, we manually curated the literature and selected only miRNA-SNP pairs that were experimentally validated by direct experimentation. Is-miRSNP could correctly recover the highest number of validated miRNA-SNP pairs (85%). The second best-performing tool mirSNP, could recover 66.6% of all validated miRNA-SNP pairs, which agrees with the 70% recovery rate reported by another study (Deveci et al., 2014). Only one SNP-miRNA (rs2240688 and miR-135b-3p) pair was not predicted by our algorithm, and was accurately classified by miRSNP, mirsnpscore and miRNASNP2. Our tool only accounts for changes in binding energy between alleles. Therefore, SNPs that alters other binding features (i.e. miRNA-mRNA structure, conservation), but not energy will not be accurately classified by is-miRSNP. If there is a difference in binding energy between alleles but the difference is not statistically significant, then the SNP-miRNA pair won't be considered to alter binding. This is the case of SNP-miRNA pairs: rs7911488, miR-1307-3p; rs2735383, miR-629-5p. According to Yang et. al., SNP rs2735383 alters the binding of the pre-mir form of miR-629-5p, but the mature form was never tested for binding. Is-miRSNP and the other tools described here were designed to evaluate the binding of mature miRNAs, therefore it is expected that none of the methods will accurately predict the effect of SNP-non-mature miRNA pairs.

Unfortunately, we could not assess the rate of false positives predicted by these tools due to a lack of data. It is widely known that the medical literature is biased towards positive results and rarely reports failed experiments (Dickersin, Chan, Chalmers, Sacks, & Smith, 1987). Next, we will add to our algorithm the capability of predicting the effect of insertion-deletions (INDEL) and pre-miRNAs.

In summary, we have presented a new tool called is-miRSNP which shows better performance than existing methods, and addresses most of the weakness present in such tools.

3.0 PIPELINE FOR PRIORITIZATION AND FUNCTIONAL EVALUATION OF DNA SEQUENCE VARIANTS

3.1 BACKGROUND

Transcriptome studies became popular with the invention of microarray technologies, which allowed scientists to interrogate the entire genome at a low cost. Despite their affordability, microarrays have three main limitations: probe design requires *a priori* sequence information or whole genome information, cross-hybridization between similar sequences, and poor quantification of lowly and highly expressed genes (Hrdlickova, Toloue, & Tian, 2017; Kukurba & Montgomery, 2015).

RNA sequencing (RNA-Seq) through high-throughput NGS has revolutionized our understanding of the transcriptome and is now the method of choice for the study of gene expression. RNA-Seq is a much more powerful technology and does not suffer from the same limitations as microarray, it has less background noise and a greater dynamic detection range. RNA-Seq provides a quantitative view of gene expression, alternative splicing and allele-specific expression. It also directly reveals the sequence identity essential for the discovery of new transcripts and isoforms (Hrdlickova et al., 2017; Kukurba & Montgomery, 2015).

Knowledge about gene expression plays an important role in deciphering the underlying biology of diseases, however understanding the genetic variation causing expression changes is also paramount. The advent of microarrays and NGS technology has also revolutionized the field of GWAS, making it more affordable and common, thus being responsible for the sharp increase in the number of published and publicly-available GWAS datasets observed in recent years. The main goal of GWAS has been to identify genetic risk factors that can be used to uncover genetic clues about cellular and molecular pathways possibly leading to new prevention strategies and treatments (Bush & Moore, 2012). An important concept in GWAS data analysis and results interpretation is the one of linkage disequilibrium (LD). LD is a measure of the linkage between two markers. It expresses how the allele of one SNP is correlated or inherited with an allele of another SNP within a population. LD values are most commonly reported in terms of the statistical measure of correlation r^2 (value between 0 and 1). LD is calculated between two SNPs, and high r^2 values mean that two SNPs are linked. If two SNPs are in LD (high r^2 value), it means one allele of the first SNP is usually observed with one allele of the second SNP (Bush & Moore, 2012).

In recent years, we have seen an explosion in the amount of genomic data available. This excess of data has also created the need for new bioinformatics pipelines and methods capable of analyzing, interpreting and integrating genomic data. This chapter describes a new pipeline capable of analyzing RNA-Seq data, integrating its results with available GWAS using an LD-based approach, and functionally annotating DSVs associated with the phenotype being studied.

3.2 METHODS

This section describes, in detail, the pipeline for prioritization and functional evaluation of DSVs (**Figure 5**). The pipeline has four main sections: 1) analysis of RNA-Seq datasets, 2) identification of variants based on RNA-Seq data, 3) GWAS data analysis, and 4) functional annotation of DSVs.

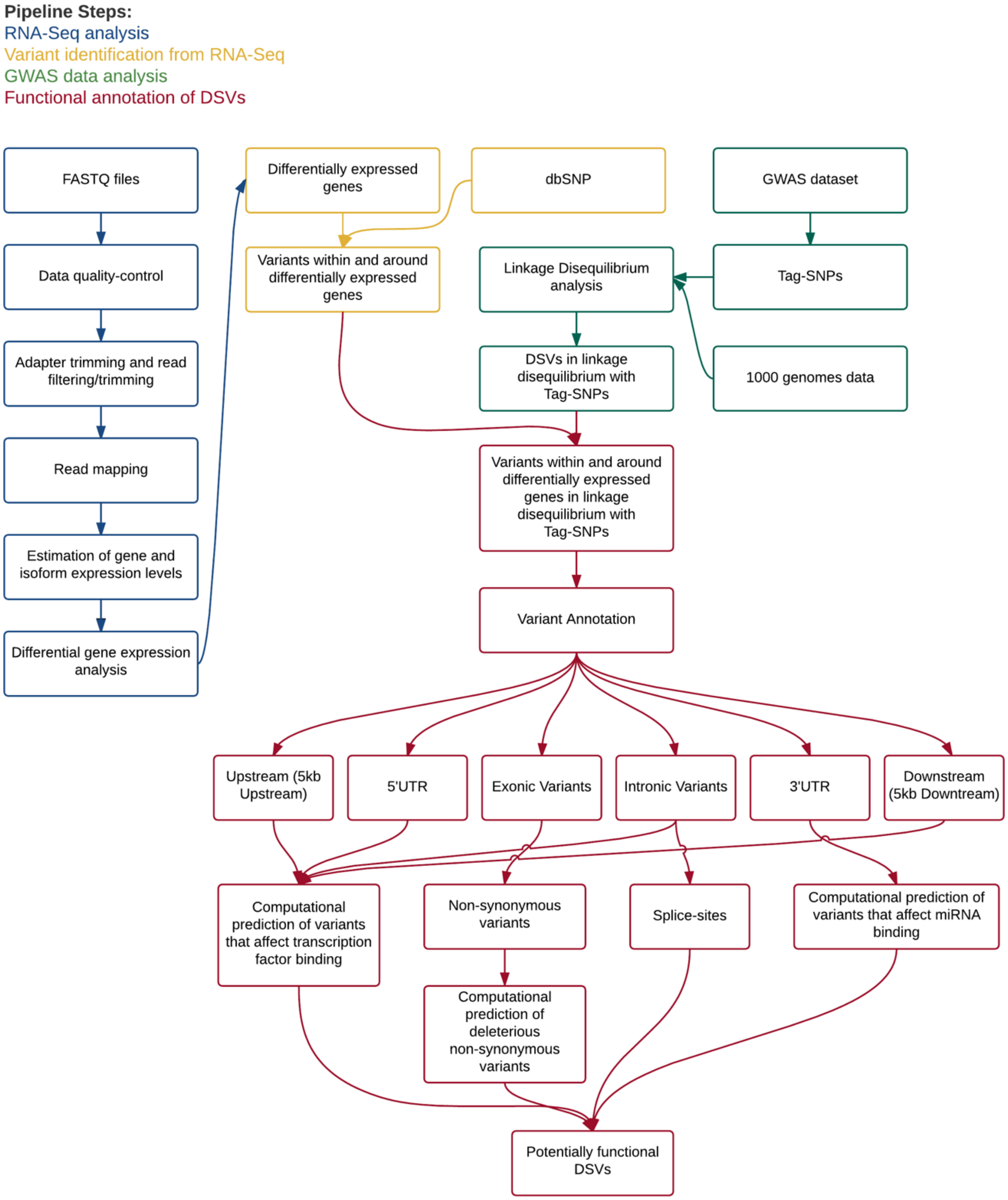

Pipeline Steps:
RNA-Seq analysis
Variant identification from RNA-Seq
GWAS data analysis
Functional annotation of DSVs
FASTQ files
Data quality-control
Adapter trimming and read filtering/trimming
Read mapping
Estimation of gene and isoform expression levels
Differential gene expression analysis
Differentially expressed genes
Variants within and around differentially expressed genes
dbSNP
GWAS dataset
Tag-SNPs
Linkage Disequilibrium analysis
1000 genomes data
DSVs in linkage disequilibrium with Tag-SNPs
Variants within and around differentially expressed genes in linkage disequilibrium with Tag-SNPs
Variant Annotation
Upstream (5kb Upstream)
5'UTR
Exonic Variants
Intronic Variants
3'UTR
Downstream (5kb Downtream)
Computational prediction of variants that affect transcription factor binding
Non-synonymous variants
Splice-sites
Computational prediction of variants that affect miRNA binding
Computational prediction of deleterious non-synonymous variants
Potentially functional DSVs

Figure 5. Pipeline for functional annotation of DSVs obtained from RNA-Seq and GWAS datasets

3.2.1 RNA-Seq analysis

Analysis of RNA-Seq data from raw FASTQ files consists of 5 main steps: 1) data quality-control, 2) adapter trimming, and read filtering and trimming; 3) read mapping; 4) estimation of gene and isoform expression; 5) differential gene expression analysis.

3.2.1.1 Data quality-control

The first step of the RNA-Seq analysis pipeline consists of assessing the quality of the raw data. This is an important step in the analysis pipeline as poor-quality data will likely lead to erroneous results. Several metrics should be calculated to assess RNA-Seq data quality, including: base sequence quality, base sequence content, reads GC content, number of missing calls, sequence length distribution, sequence duplication levels. In this pipeline the FastQC software (Andrews, 2012), is used to calculate and display data quality metrics.

3.2.1.2 Adapter trimming and read filtering and trimming

Adapter trimming, and read filtering and trimming are steps in the pipeline that directly depend on the library preparation method, and results from previous step. Adapter sequences can be present or not depending if they were added and/or removed during library preparation stage. If present they need to be removed, as they will affect read mapping and expression estimation. For the same reasons, low-quality reads should also be removed. Often, NGS reads will have low-quality bases at the 3’, 5’ or both ends. In such cases, read trimming is preferred to read filtering, since it maximizes data utilization as it preserves good quality-bases in the middle of the sequences. For the pipeline described here, Cutadapt (Martin, 2011) is used for adapter trimming, and read filtering and trimming.

3.2.1.3 Read mapping

The third step in the RNA-Seq analysis pipeline is read mapping. Reads need to be mapped to the genome before gene expression can be estimated. Several algorithms were recently developed to accomplish this task. We used Bowtie (Langmead, Trapnell, Pop, & Salzberg, 2009), an ultrafast, memory-efficient algorithm that uses Burrows-Wheeler indexing. Bowtie can accurately map an RNA-Seq sample to the human transcriptome in only a couple of hours. After read mapping is completed, the generated BAM files need to be evaluated for quality-control purposes. Metrics like number and percentage of mapped reads should be calculated to determine whether the data was appropriately mapped.

3.2.1.4 Estimation of gene and isoform expression

After read mapping, gene and isoform expression are ready to be estimated. This is a difficult problem since mapped reads do not always map to unique positions in the genome, making it hard to distinguish which gene and/or isoform the target sequence originated from. In this pipeline, we utilize the RSEM algorithm to estimate gene and isoform abundance. RSEM implements a quantification method for expression that uses the expectation maximization to accurately estimate read mapping uncertainty (Li & Dewey, 2011).

3.2.1.5 Differential gene expression analysis

The most common goals of RNA-Seq experiments are: 1) to understand the differences in the transcriptome between two cell types, 2) to identify transcriptional differences between health and disease states, and 3) investigate gene expression changes caused by interventions and drugs. Therefore, an important step in an RNA-Seq pipeline is to find differentially expressed genes and isoforms between two or more groups. Statistical tests designed for microarray data cannot be applied to RNA-Seq data. This is mainly because the distribution of RNA-Seq data follows a negative binomial distribution, and cannot be properly analyzed by methods that assume normality. Therefore, to perform differential expression analysis, a method that takes into consideration the negative binomial distribution needs to be used. Recently, several algorithms like EBSeq (Leng et al., 2013), DESeq (Love, Huber, & Anders, 2014), Limma-Voom (Ritchie et al., 2015) and EdgeR (McCarthy, Chen, & Smyth, 2012) have been developed. In the present pipeline, we utilize EdgeR, which implements a negative binomial regression model that can accurately identify differentially expressed genes. EdgeR is a fast algorithm that can handle large datasets, and it is flexible enough to take into consideration various experimental designs.

3.2.2 Variant identification from RNA-Seq

Once RNA-Seq analysis is completed, the next step in the pipeline is to identify variants related to the genes of interest. In the proposed pipeline, these are the set of differentially expressed genes. It is worth noting that if a user does not have control samples, or is not interested in comparing multiple groups and/or treatments, differentially expressed genes can be replaced by any set of genes (i.e. genes inside biologically meaningful pathways, or genes with high or low expression). Variants are identified by intersecting the genomic coordinates of the differentially expressed genes and their flanking regions (5kb upstream and downstream of the gene) with a genetic variant database like dbSNP (Sherry et al., 2001).

3.2.3 GWAS data analysis

The previous step will result in a variant set that is only a fraction of the total number of variants in the human genome. However, it is likely that the number of variants will still be too large, and will include many DSVs that have no biological function. The sheer number of DSVs also make manual curation and interpretation of variants virtually impossible. To further reduce the number of variants, our pipeline utilizes results obtained from GWAS datasets. First, tag-SNPs or variants that are strongly associated with the condition being studied are obtained from available GWAS data. Next, LD is calculated between GWAS variants and variants located within 5kb to either side of differentially expressed genes. LD analysis is computed using the software PLINK (Purcell et al., 2007). The population data and haplotypes needed for LD calculations are obtained from whole-genome sequencing projects such as the 1000 Genomes Project (Genomes Project et al., 2010). The set of variants in strong LD ($R^2 \geq 0.8$) with GWAS variants are considered strongly associated with the condition being studied, and are next functionally annotated.

3.2.4 Functional Annotation of DSVs

The first step in predicting DSV function is annotation. It provides detailed information about the variants such as: 1) genomic region (3'UTR, 5'UTR, upstream, downstream, exon, intron); 2) gene; 3) functional characterization of exonic (synonymous, non-synonymous, stop-gain) and splice-site variants. In this stage of the pipeline all variants in strong LD with Tag-SNPs are annotated, and grouped by genomic region (i.e. upstream, 5'UTR, exon, intron, 3'UTR and downstream). DSVs are annotated using ANNOVAR (K. Wang, Li, & Hakonarson, 2010; H. Yang & Wang, 2015).

3.2.4.1 Functional annotation of exonic and splicing DSVs

The effects of non-synonymous DSVs are predicted using the CAROL algorithm (Lopes et al., 2012). CAROL combines the scores of the two most commonly used tools for predicting the effect of non-synonymous variants, namely PolyPhen (Adzhubei et al., 2010) and SIFT (Kumar, Henikoff, & Ng, 2009). Both algorithms use conservation and evolutionary patterns to predict whether a variant is deleterious or not. SIFT determines the effect of a variant by calculating the probability of an amino-acid change at all positions across a set of multiple aligned homologous sequences. PolyPhen predicts whether a variant is potentially harmful by calculating the posterior probability that the DSV is deleterious. PolyPhen also considers the changes in the protein 3D structure in its predictions. Often the predictions by PolyPhen and SIFT will not agree and/or will be hard to interpret. We solved this problem in our pipeline by using CAROL, which utilizes a weighted Z method to derive a SIFT-PolyPhen combined scored. CAROL not only solves the ambiguity of prediction between SIFT and PolyPhen, but it also has increased predictive power and accuracy when compared to these tools alone (Lopes et al., 2012). SIFT and PolyPhen scores as well as variants in splice-sites are obtained from ANNOVAR.

3.2.4.2 Functional annotation of 3’UTR DSVs

The impact of 3’UTR DSVs on miRNA binding sites is predicted using the is-miRSNP algorithm (described in detail on Chapter 2).

3.2.4.3 Functional annotation of intronic, promoters and 5’UTR DSVs

Intronic, promoter and 5’UTR DSVs are tested for their potential effect on TF binding sites. The effects of these variants on TF binding are predicted by using the BayesPI-BAR algorithm (J. Wang & Batmanov, 2015). The BayesPI-BAR algorithm utilizes a Bayesian approach to estimate the changes in the binding affinity of TF caused by a genetic variant. The method incorporates TF chemical potentials and putative direct protein-DNA interactions not considered by other methods (J. Wang & Batmanov, 2015).

4.0 ANALYSIS OF INFLAMMATORY BOWEL DISEASE DATASET

4.1 BACKGROUND

Idiopathic inflammatory bowel disease (IBD) is a group of chronic, relapsing and remitting gastrointestinal tract disorders that include two common subtypes, Crohn’s disease (CD) and ulcerative colitis (UC). Research over the last few decades has led to a general consensus that IBD pathogenesis involves complex interactions between host genetic factors, environmental triggers, and the gut microbiota (Ananthakrishnan, 2015; Fonseca-Camarillo & Yamamoto-Furusho, 2015; Podolsky, 2002). The genetic contribution to IBD was investigated by a GWAS and GWAS follow-up study involving more than 75,000 cases and controls (Jostins et al., 2012). This study, together with more recent genetic mapping studies in IBD, have identified 241 IBD-associated loci spanning 1,540 genes (de Lange et al., 2017; Jostins et al., 2012). Interestingly, for most of these IBD-associated loci, the SNP(s) with the strongest association signal(s) lie outside known genes in non-coding regions of the genome, where they are presumed to play a role in regulation of gene expression. This is a theme that has been observed in GWAS of many complex human diseases (Hindorff et al., 2009).

IBD are chronic, immune-mediated diseases characterized by persistent inflammation of the bowel. Chronic inflammation is thought to be due to an exacerbated immune response against commensal enteric organisms in a genetically predisposed host (Fonseca-Camarillo & Yamamoto-Furusho, 2015; Sartor, 2006). Production of proinflammatory cytokines and chemokines is increased in IBD, which ultimately leads to the abnormal activation of immune cells. Naïve CD4+ T cells have a crucial role in the initiation of the immune response by activating, expanding and differentiating into effector cells. Two important effector cell types, derived from naive CD4+ T cells, are Th1 and Th2. Each of these T effector cell types has a particular role in inflammation and secretes specific cytokines (Z. J. Liu, Yadav, Su, Wang, & Fei, 2009). Th1 cells are thought to be involved in CD, while Th2 cells are mostly observed in UC (Geremia, Biancheri, Allan, Corazza, & Di Sabatino, 2014; Z. J. Liu et al., 2009). High levels of IL17-A, a cytokine produced by a third type of effector T-cell named Th17, has been detected in both UC and CD samples (Z. J. Liu et al., 2009). Immunohistochemistry of these samples revealed enrichment for CD-68+ cells that express IL-17. Furthermore, recent data suggest that Th17 cells may play an important function in the immune response against extracellular pathogens. Th17 cells might be responsible for clearing pathogens that are not effectively dealt with by Th1 and Th2. Moreover, studies have shown that Th17 are important for the immune response modulation in various other auto-immune diseases like psoriasis and rheumatoid arthritis (Adamik et al., 2013; Z. J. Liu et al., 2009).

The maturation of naïve $CD4^{+}$ T cells into Th17 cells requires a special environment. Studies have shown that IL-1β and IL-23 are required for the differentiation and expansion of Th17 cells (Z. J. Liu et al., 2009). Another important molecule, prostaglandin E2 (PGE2), also seems to be also involved in the regulation of Th17 cells. PGE2 exhibits both anti-inflammatory and pro-inflammatory properties. In the presence of IL-23, IL-1β and PGE2, Th17 cells seem to be insensitive to the anti-inflammatory, yet still sensitive to the pro-inflammatory properties of PGE2 (Barrie et al., 2011). IL-17A and IL-17F are cytokines produced by Th17 cells and are directly linked to the pathogenesis of IBD (Adamik et al., 2013; Barrie et al., 2011). The presence of IL-23 and IL-1B preferentially induce the expression of IL-17F. The addition of PGE2 to IL-23 and IL-1B treatments induce the expression of IL-17A by CD4+ T cells via the PGE2 EP4 receptor. Therefore, the presence of PGE2 shifts the balance of IL-17F mediated immune response to a predominantly IL-17A response. In certain mouse models, IL17-A seems to protect against IBD onset, whereas IL-17F has the opposite effect (Adamik et al., 2013).

Despite the evidence that Th17 cells play an important role in IBD, their mechanism of action is not well understood, and only a few works have studied the effects of PGE2 on Th17-enriched cell populations (Adamik et al., 2013; Barrie et al., 2011; Z. J. Liu et al., 2009). In addition to Th17 modulation, IL-23, IL-1β and PGE2 have been shown to be directly involved in IBD, as each of these molecules have receptor encoded by genes (IL23R, IL1R1 nad PTGR4) implicated in IBD GWAS studies (Jostins et al., 2012; Lees, Barrett, Parkes, & Satsangi, 2011; Uniken Venema, Voskuil, Dijkstra, Weersma, & Festen, 2017).

The purpose of this work is to study the mechanisms by which PGE2 affects gene expression of activated, IL-23 and IL-1β-treated, effector memory T cells. Gene expression analyses were also linked to GWAS data in order to identify DNA sequence variants (DSVs) that can potentially modulate Th17 function and might be of importance to IBD pathology.

4.2 METHODS

4.2.1 Cell isolation, purification and stimulation

Peripheral blood mononuclear cells (PBMCs) were isolated by density gradient centrifugation on Ficoll-based Lymphocyte Separation Medium (MP Biochemicals) from eight anonymous healthy adult human leukopaks obtained from the Pittsburgh Central Blood Bank as approved by the University of Pittsburgh IRB. $CD4^{+}CD45RO^{+}CD197^{-}$ T cells were enriched by negative selection from PBMCs using the human CD4+ Effector Memory T Cell Isolation Kit (Miltenyi Biotec). Cell purity was assessed by flow cytometry via staining for CD4 and CD45RO surface expression. Isolated T cells were cultured at 1×10^{-6} cells/mL in X-VIVO-20 medium (Lonza BioWhittaker) supplemented with T cell activation beads (Miltenyi Biotec) coated with anti-CD2/-CD3/-CD28 antibodies (1 bead per 5 cells). The cells were treated with the T cell activation beads (ActivBeads) alone or in combination with either IL-23+IL-1β (50 ng/mL each, both from R&D Systems); PGE2 (1 μM, Sigma-Aldrich Chemical); or IL-23+IL-1β+PGE2. The cell cultures were incubated at 37°C, 5% CO2, for 6 hours because preliminary time course studies using cells from three donors showed that IL17A mRNA expression increased within 3 hours, peaked at 6 hours, and subsequently declined over the next 18 hours following treatment with T cell activation beads alone. The cells were then harvested and lysed with Qiazol reagent, and total RNA was extracted using the miRNeasy Mini Kit (Qiagen).

4.2.2 RNA sequencing

A total of 32 RNA samples (from four cell culture conditions for each of eight human cell donors) were converted to cDNA using the Ovation® RNA Amplification System V2 (NuGEN) following the manufacturer's recommended protocol. The cDNA produced by this kit was then sheared to an average size of 250 bp using a Covaris E220 acoustic wave system following manufacturer's recommended protocol. The sheared DNA was prepared for Illumina Sequencing using the TruSeq v2 kit for library preparation with barcode adaptors AD001-016, AD018-022, AD023, AD025, and AD027. The libraries were pooled in groups of eight and amplified, quantified by qPCR on a Life Technologies Gene Amp® 9700 RT- PCR device, and diluted to a concentration of 11 pM for deposition and clustering on an Illumina Flowcell using the Illumina cBot. The DNA was striped across 2 wells each of the flowcell. The DNA was then sequenced on an Illumina HiSeq 2000 using the Illumina SBS kit v3, 200 cycle, paired end chemistry. The output was de-multiplexed and base calls were converted to fastq format using CASAVA v1.8.2.

4.2.3 Analysis of gene expression

The RSEM algorithm (Li & Dewey, 2011) was used with the default parameters to map the sequencing reads on the human genome (version hg19) and estimate gene and isoform expression. Differentially expressed genes and isoforms were identified using edgeR (McCarthy et al., 2012). Since the analyzed samples were derived from eight individuals and four culture conditions each (T cell activation beads (ActivBeads) alone or in combination with either IL-23+IL-1β, PGE2 or IL-23+IL-1β+PGE2), we used edgeR with the "blocking" feature to account for the same donor samples. The threshold for differential gene expression was set to FDR<0.05

in all comparisons. Additional statistical analyses were done with in-house made scripts in R. Pathway analysis of differentially expressed genes was done using IPA (Ingenuity® Systems, www.ingenuity.com).

4.2.4 Evaluation of cis effects of IBD-associated SNPs within differentially expressed gene regions

Gene models and exon boundaries were obtained from the UCSC Genome Browser (hg19) (Karolchik et al., 2004). SNPs localized within IBD loci of the differentially expressed genes in IBD-associated genomic regions were retrieved from dbSNP version 147 (Sherry et al., 2001) and annotated using ANNOVAR (K. Wang et al., 2010; H. Yang & Wang, 2015).

For completeness, we also ranked the non-synonymous SNPs using CAROL (Lopes et al., 2012) a program that calculates how likely a particular variant impacts protein structure and function. CAROL score is based on Polyphen (Adzhubei et al., 2010) and Sift (Kumar et al., 2009) which were obtained from ANNOVAR. SNPs inside introns and regulatory regions (upstream, downstream, 5'UTR) were evaluated using the BayesPI-BAR algorithm (J. Wang & Batmanov, 2015) for the transcription factors that had at least 5 reads in each of the samples and treatments. SNPs in the 3' untranslated regions (3'UTR) were evaluated for their potential influence on miRNA binding using MirSNP (C. Liu et al., 2012). Also, SNPs that alter splice sites were reported.

For each one of the IBD-associated regions, Jostins et al. (Jostins et al., 2012) and DeLange et al. (de Lange et al., 2017) identified one SNP that could best explain the GWAS signal. LD analysis of these tag-SNPs, and SNPs previously identified as functionally important by our analyses was performed using PLINK (Purcell et al., 2007). LD analysis was done using haplotypes for all individuals of European ancestry listed in the 1000 Genomes Project phase 3.

4.3 RESULTS

4.3.1 The effects of PGE2 dominate the differential gene expression observed in activated effector memory T cells treated with IL-23+IL-1β, PGE2 or all three mediators

In the all-groups comparison (edgeR, ANOVA-like test) we identified a total of 3,541 genes as differentially expressed between at least one of the groups (FDR < 0.05) (**Table 3**). Unsupervised hierarchical bi-clustering showed two major groups of samples: PGE2-induced cells and non-PGE2-induced cells (**Figure 6**). In addition, one or more isoforms of 1,122 genes were differentially expressed in an all group comparison (total number of differentially expressed isoforms = 1,349; **Table 4**)

Table 3. Summary table of differentially expressed genes

COMPARISON	NUMBER OF SIGNIFICANT GENES (FDR < 0.05)	NUMBER OF SIGNIFICANT GENES IN IBD-ASSOCIATED REGIONS (FDR < 0.05)	NUMBER OF NON-SIGNIFICANT GENES
all groups' comparison	3541	326	17106
Control+IL23.IL1B vs IL23.IL1B.PGE2+PGE2	3562	336	17085
IL23.IL1B.PGE2 vs IL23.IL1B	2472	241	18175
IL23.IL1B vs Control	159	17	20488

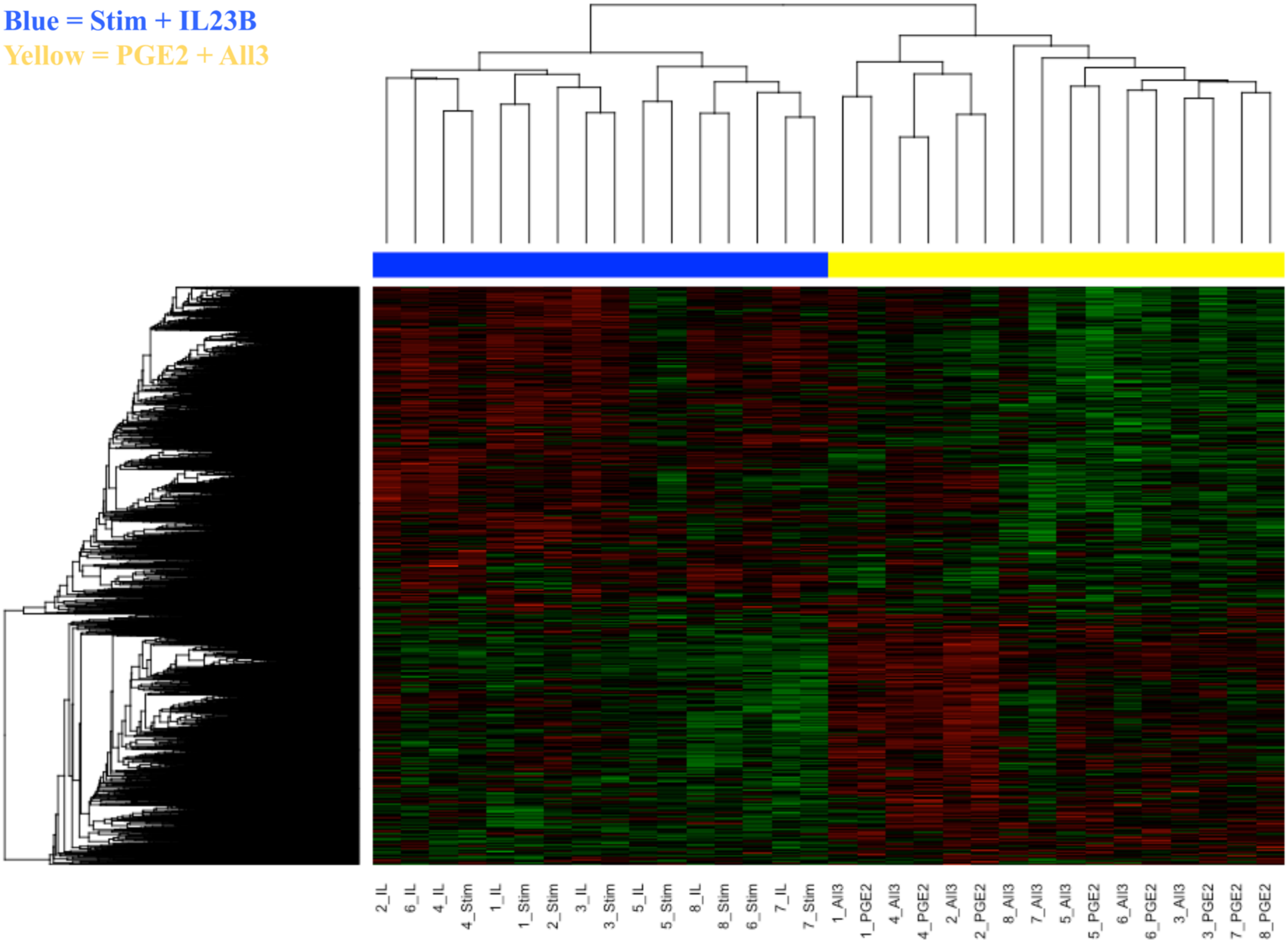


Figure 6. Hierarchical bi-cluster of differentially expressed genes across all stimuli (ANOVA-like test). High expression values are in red and low expression values are in green.

Table 4. Summary table of differentially expressed isoforms

COMPARISON	NUMBER OF SIGNIFICANT ISOFORMS (FDR < 0.05)	NUMBER OF SIGNIFICANT ISOFORMS IN IBD-ASSOCIATED REGIONS (FDR < 0.05)	NUMBER OF NON-SIGNIFICANT ISOFORMS
all groups' comparison	1349	184	55306
Control+IL23.IL1B vs IL23.IL1B.PGE2+PGE2	1859	236	54796
IL23.IL1B.PGE2 vs IL23.IL1B	637	85	56018
IL23.IL1B vs Control	65	7	56590

4.3.2 Effect of PGE2 on the transcriptome of activated effector memory T cells

We found 3,562 differentially expressed genes in effector memory T cells treated with PGE2 *vs* without PGE2 (**Table 3**). In general, samples from the same donor are clustered together within each of the two largest groups (**Figure 7**), despite using the patient_ID as group variable (blocking) during the analysis. Exceptions are samples #2 and #6 in the non-PGE2 treated cells and #7 and #8 in the PGE2 treated cells. A significant number of differentially expressed genes (336 of the 3,562 genes) belong to the group of 1,540 genes that are located within IBD loci (Fisher's exact test, *p*-value=e-6).

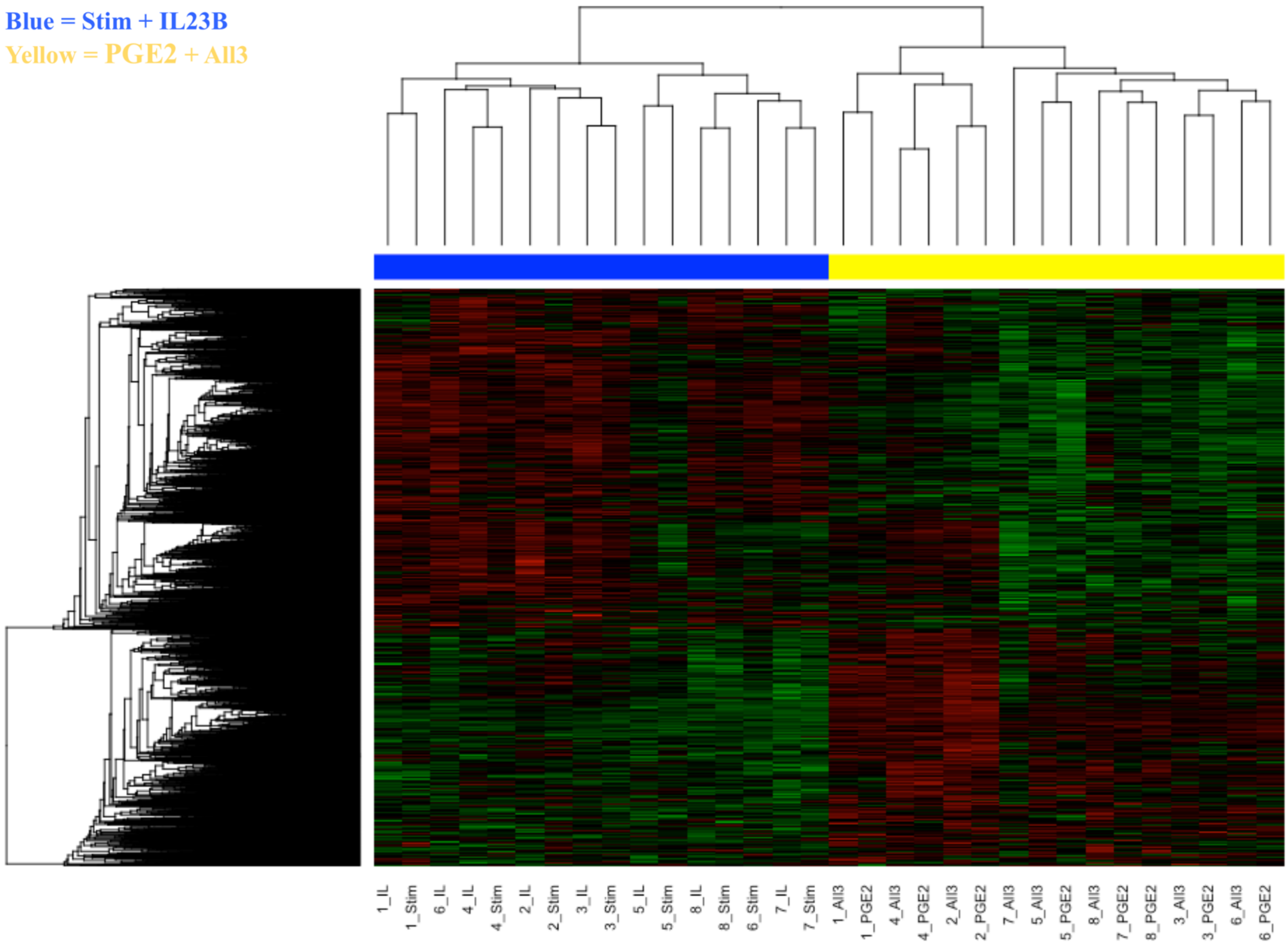


Figure 7. Hierarchical bicluster of DE genes between ActivBeads+IL23+IL1β *vs* ActivBeads+IL23+IL1β+PGE2 stimuli across all samples. High expression values are in red and low expression values are in green.

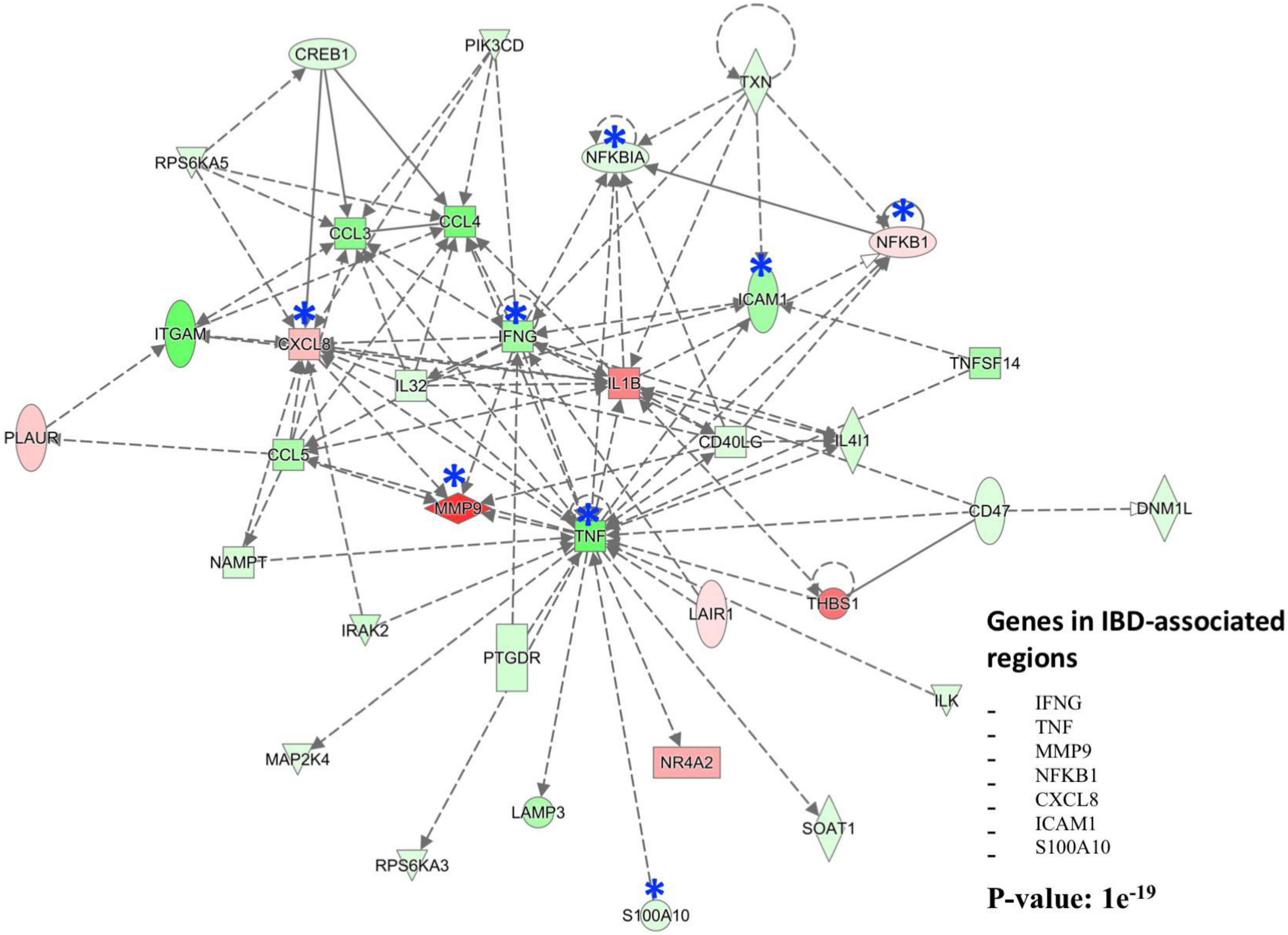
4.3.3 Effects of IL-23 and IL-1β on the transcriptome of activated effector memory T cells

We also looked at the gene expression differences induced by IL23+IL1β, and we found their effects to be relatively small. A total of 159 genes were differentially expressed as a result of IL23+IL1β stimulation (**Table 3**), and 17 of them belong to IBD regions.

4.3.4 Pathway analysis reveals important relationships to IBD

Ingenuity Pathway analysis revealed several networks with statistically overrepresented differentially expressed genes in these categories. In particular, we found that the 2,472 differentially expressed genes between samples activated with ActivBeads+IL23+IL1β+PGE2 *vs* ActivBeads+IL23+IL1β are overrepresented in 25 networks. Similarly, the differentially expressed genes in samples activated with ActivBeads+IL23+IL1β *vs* ActivBeads alone were overrepresented in two networks (p-value $\leq$ 0.001). **Figure 8** presents four networks from the ActivBeads+IL23+IL1β *vs* ActivBeads+IL23+IL1β+PGE2 comparison that are particularly interesting because they include several up- or down-regulated genes inside IBD-associated regions, suggesting relevance to IBD. These networks are related to the function and development of the immune response. Genes such as *IL2, IL7R, CREM and FOS* that are immune system response genes and located within IBD-associated regions were found in the networks.

a) Cellular Development, Cellular Growth and Proliferation, Cellular Movement

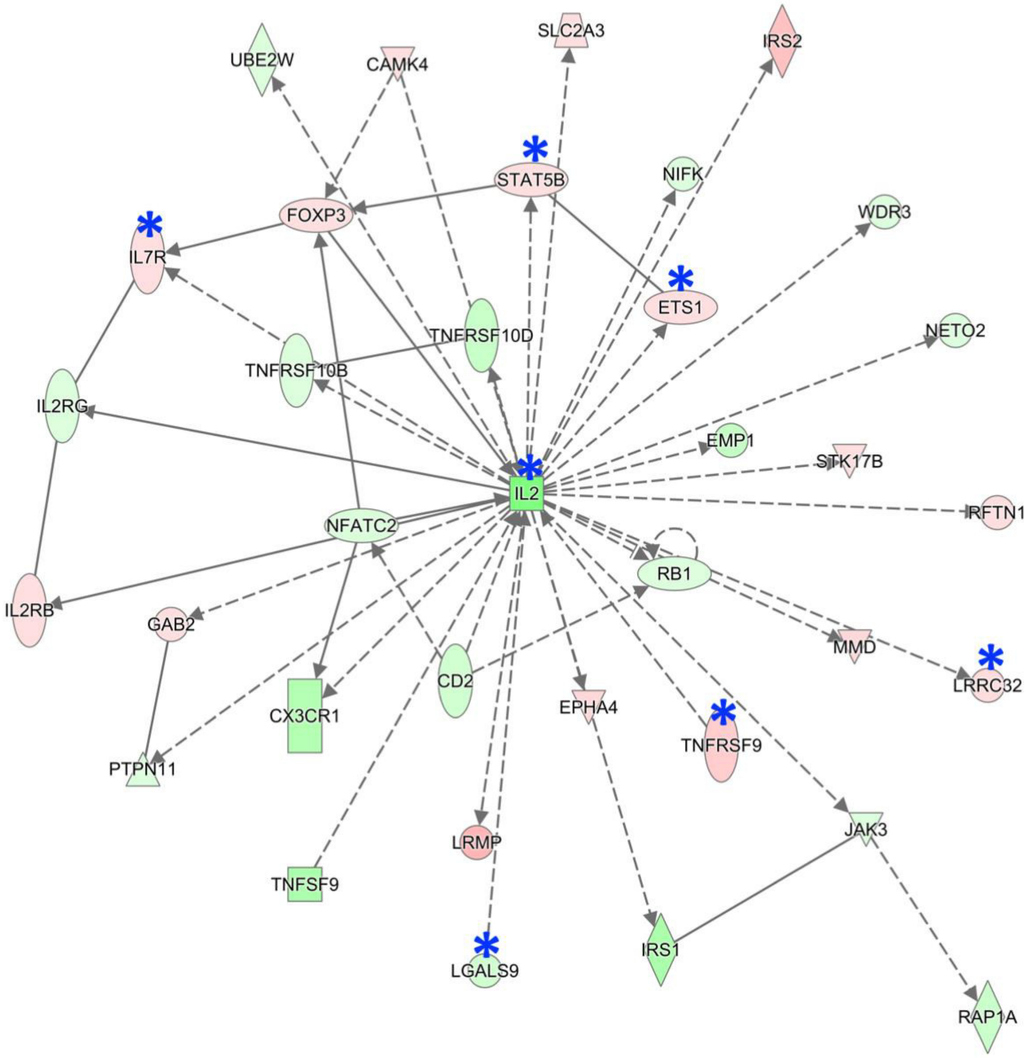


Genes in IBD-associated regions

- IFNG
- TNF
- MMP9
- NFKB1
- CXCL8
- ICAM1
- S100A10

P-value: $1e^{-19}$

b) Cellular Development, Hematological System Development and Function, Hematopoiesis

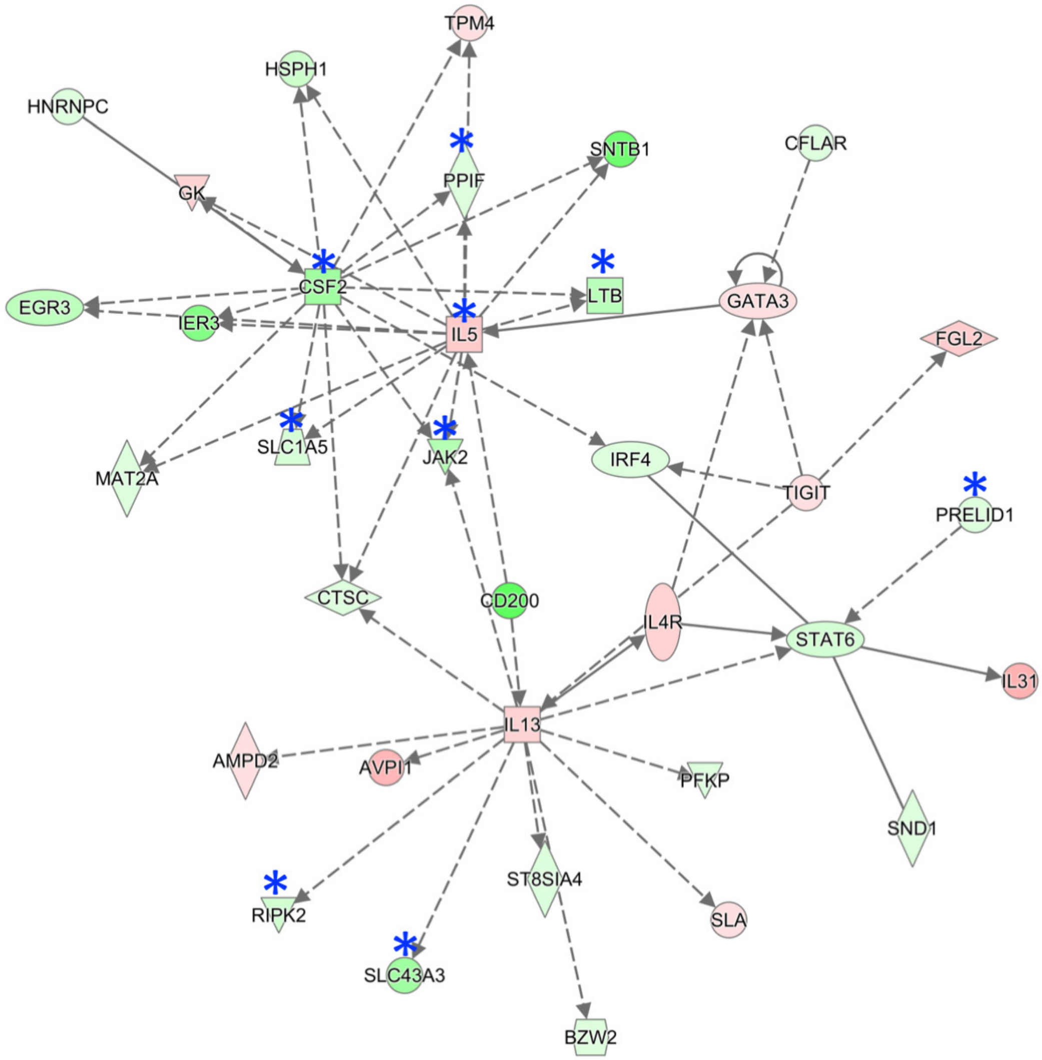


Genes in IBD-associated regions

- IL7R
- STAT5B
- IL2
- LGALS9
- LRRC32
- TNFRSF9

P-value: $1e^{-19}$

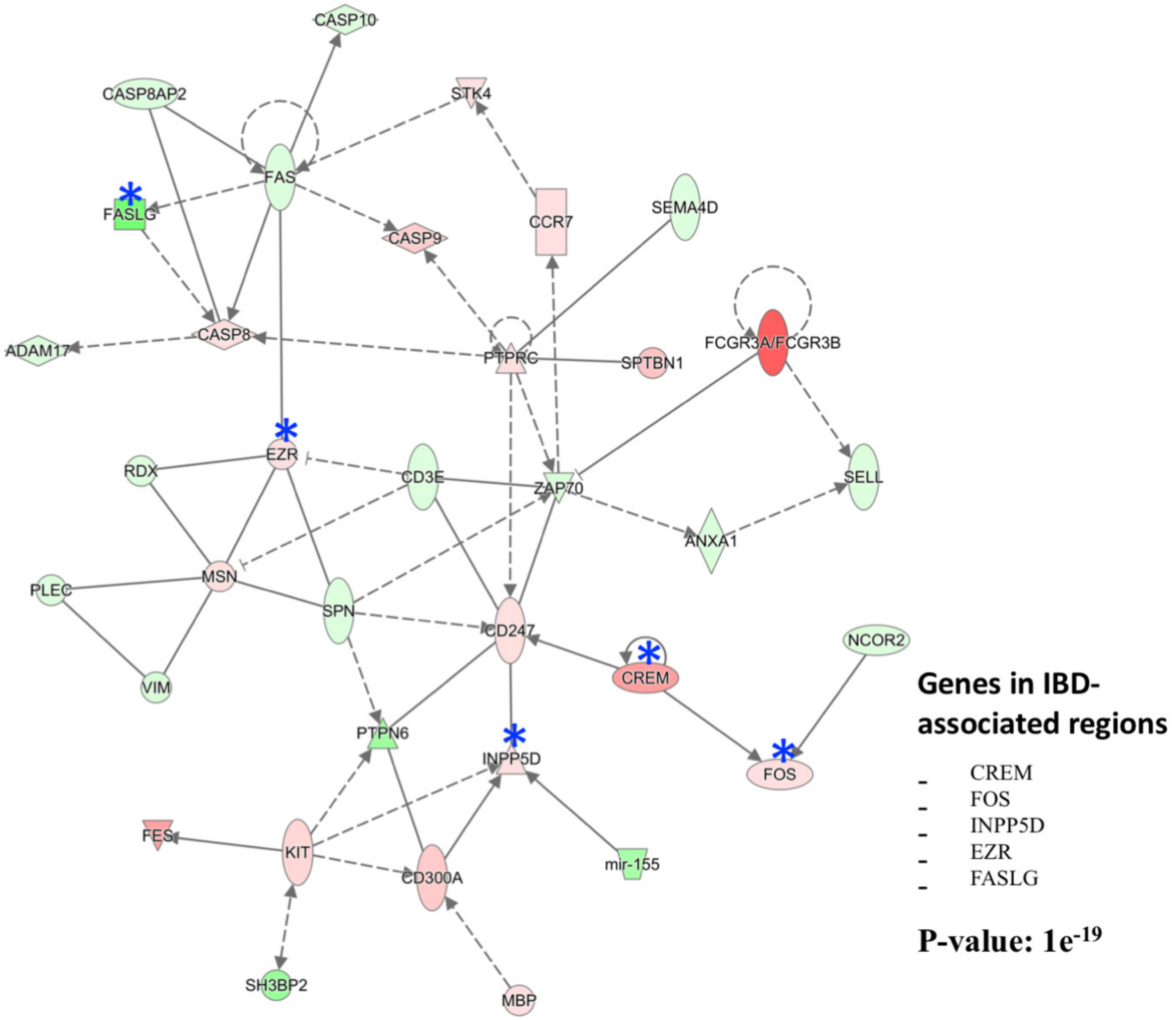
c) Cell-mediated Immune Response, Cellular Development, Cellular Function and Maintenance



Genes in IBD-associated regions

- IL5
- CSF2
- PPIF
- JAK2
- SLC1A5
- SLC43A3
- RIPK2
- PRELID1
- LTB

P-value: 1e^{-19}



d) Cell Death and Survival, Cellular Development, Cellular Growth and Proliferation

Figure 8. Networks of differentially expressed genes ActivBeads+IL23+IL1β *vs* ActivBeads+IL23+IL1β+PGE2. Blue asterisks indicated genes located inside IBD-associated regions. P-value (calculated by the Fisher's exact test) is indicative of how overrepresented the differentially expressed genes are in the given network**.**

4.3.5 SNPs with potential functional role in IBD

Jostins *et al.* (Jostins et al., 2012) and DeLange *et al.* (de Lange et al., 2017) identified 241 IBD-associated genomic regions that contain 1,540 genes. 326 of these genes are in the group of

3,541 differentially expressed between one or more of the four treatments (*p*-value=e-5 Fisher's exact test). (**Table 3**). They include several genes related to immune responses such as TNF, IFN-γ, IL7R, IL2, IL21. When we compared gene expression differences between IL23+IL1β versus IL23+IL1β+PGE2 stimulation, we found a total of 2472 differentially expressed genes, 241 of which are in IBD-associated regions (*p*-value=1e^{-6} Fisher's exact test). We selected and further analyzed SNPs located within 5kb upstream and downstream of these 241 genes. We found a total of 582,510 SNPs in these regions, with most of them being intronic (478,013 or ~82%) (**Table 5**). We assessed the functional role of (i) non-synonymous SNPs in altering protein folding and function; (ii) SNPs in 3'-UTR in altering miRNA binding; and (iii) promoter, 5'-UTR and intronic SNPs in altering the binding sites of known transcription factors. A total of 23,809 non-synonymous exonic SNPs, belonging to 333 of these genes, were predicted to be deleterious. From all 23,809 non-synonymous SNPs predicted to be deleterious, 6 of them were inside DE genes and in strong LD with IBD-tag SNPs (**Table 6**).

A total of 4,460 SNPs found in the 5'UTR, intronic, upstream and downstream regions of 241 genes in the IBD regions were predicted to disrupt TF binding and were in strong LD with IBD-tag SNPs. Some of those SNPs are predicted to disrupt binding sites for TFs located in the IBD regions, like BACH2 (promoters of ADAM30, JAK2 and introns of CREM, SMAD3, STAT3), CEBPB (promoters of IRGM, USP1), SMAD3 (promoter of CREM) among others. **Table** 7 presents 34 of these SNPs that are located upstream or downstream of differentially expressed genes. This list includes 13 SNPs in the promoter regions of important immune related genes like CREM, IL8R1, JAK2 and MAP3K8.

We identified 71 SNPs located in the 3'UTR regions of 43 DE genes in IBD loci that are in strong LD with IBD-tag SNPs and are predicted to change the binding of miRNAs. These are predicted to alter the binding of a total of 23 different miRNAs (**Table 8**). We found 18 SNPs in 3'UTR regions of DE genes that are in strong LD with IBD-tag SNPs (**Table 8**). Linkage disequilibrium and functional prediction showed that the tag SNP rs727088 located on the 3' UTR of the CD226 gene is predicted to alter miR-181a-5p and miR-502-5p binding **(Table 8).** Finally, splice site analysis identified 5 SNPs that changes splice sites and are in linkage disequilibrium with IBD tag-SNPs, however these are not in differentially expressed genes (**Table 9**).

Table 5. Summary of DE genes between IL23+IL1β vs. IL23+IL1β +PGE2 and the number of DSVs found within and around these genes.

FPKM: fragment per kilobase per million.

GENE	MEAN.FPKM values in different stimuli				# of SNPs with respect to gene's location					
	STIM (Ctrl)	IL23+IL1β	PGE2	IL23+IL1β +PGE2	upstream	5-UTR	exonic	intronic	3-UTR	downstream
ACSL6	6.3575	6.8375	3.30125	3.3475	46	46	342	2588	140	47
ACYP1	5.6625	4.17375	6.5625	6.5175	0	23	53	592	18	23
ADK	12.90125	13.98	18.71	20.785	0	10	5	348	0	0
ADO	7.55875	10.655	5.93625	5.4825	38	23	148	0	125	51
ANKRD33B	1.41375	1.9525	0.9325	1.35125	48	4	139	4327	394	39
ANXA6	9.8375	8.855	7.2675	6.73	43	38	362	2878	48	48
APOBEC3C	18.525	15.98	11.55875	11.2725	71	14	104	260	36	55
APOBEC3G	25.45	25.0825	14.9575	17.47375	44	19	216	579	24	50
ATF6B	13.7	11.74625	15.9425	14.53375	0	7	311	602	28	30
ATP1B1	2.05875	2.05625	3.05375	3.41	0	0	84	480	55	0
ATP6V1F	13.485	10.55	11.03	14.805	33	16	75	117	11	0
ATXN2L	45.165	46.67875	38.60625	36.10625	36	18	603	860	89	51
BACH2	35.60125	33.2425	58.6925	55.79	32	46	26	12616	0	0
BCL9L	7.77875	7.13	6.50375	5.64	4	40	854	419	128	0
BORCS5	2.05	1.9025	3.3925	3.21625	0	44	110	5346	77	52
C6orf48	78.2425	68.72	79.99375	96.265	11	81	51	239	17	45
C6orf99	0	0	0.14625	0.4575	42	47	22	1858	8	44
CACNA1I	0.4175	0.31	0.41375	0.52125	47	0	37	153	0	0
CAMSAP2	5.105	6.1075	5.12875	4.495	39	13	546	4809	86	44
CBLL1	22.15625	23.2075	19.3475	17.9325	61	42	218	642	119	43
CBX7	4.60875	4.01375	5.2375	5.055	37	4	120	908	167	64
CCDC86	9.34125	10.4525	6.45875	7.095	44	23	213	385	35	0

CCR2	8.36375	7.825	12.68875	12.65625	46	31	216	183	102	47
CCR5	9.0175	8.42125	12.8025	11.48875	0	0	225	0	0	0
CCR6	8.50875	9.2375	11.335	12.93375	44	45	150	1189	89	46
CD26	57.55	62.48625	36.325	38.92625	46	49	168	4338	63	40
CD40	0.7575	0.6575	0.6625	1.2225	43	23	128	536	40	47
CD48	98.10875	104.165	54.72	62.7475	56	12	161	1719	76	57
CD5	39.29125	36.045	75.25125	67.13625	47	10	266	1103	71	46
CD6	28.54125	25.80625	32.85375	32.1475	48	19	290	2251	72	49
CD74	68.0725	63.30375	74.11375	75.7825	39	13	141	540	44	19
CDC42SE2	127.65	135.19	156.70375	152.48875	29	43	37	5908	122	39
CDKN1B	44.0125	38.32875	65.74	61.6575	48	41	150	166	72	55
CLIC1	53.51625	53.15125	38.7725	45.115	22	53	86	230	11	0
CLTC	26.62375	29.5325	23.9675	24.345	0	0	378	2492	127	0
CNN2	8.67375	7.68875	10.2725	10.485	39	23	221	653	70	1
COMMD7	4.35875	4.34625	5.26125	5.58125	42	14	83	1811	38	48
CPEB4	9.7225	10.48625	6.4025	6.1475	47	57	231	2560	179	30
CREBL2	7.86375	7.99375	9.94125	9.68125	51	29	30	1289	143	50
CREM	44.7525	52.5425	202.63	206.66	53	100	216	3658	156	50
CRTC3	12.585	12.03	14.05375	14.4225	48	9	356	4228	0	0
CSF2	170.12625	211.3525	64.11	91.325	42	2	70	114	14	35
CUL2	9.2425	10.7275	8.38125	7.70625	38	28	247	3703	69	49
CXCL8	0.425	1.54375	0.625	3.7925	52	18	53	103	87	50
CXCR5	1.8875	2.005	1.1625	1.2425	47	34	165	468	80	0
CYTH2	2.33125	1.8975	2.29125	2.515	46	22	201	598	170	51
DCTPP1	17.2675	18.09125	12.13375	13.0675	30	12	97	222	24	44
DENND1B	4.89125	4.9075	4.11375	3.8525	40	17	390	10312	303	37
DGKE	16.78	18.73	27.02125	27.01625	0	11	274	1288	227	40
DNMT3A	12.48875	11.76375	7.22375	7.4675	0	12	468	2123	91	42

DOCK9	31.04875	31.72375	26.8475	23.3775	0	0	8	1667	0	0
DOK6	5.2875	5.64625	4.165	4.28	0	0	119	10102	304	48
DPAGT1	12.88625	11.63625	9.83625	9.27125	49	35	177	246	28	0
DUSP16	19.99	19.7	36.64	37.28375	31	35	326	3609	207	82
DUSP28	2.19625	1.77875	2.50625	2.63625	0	45	84	144	34	55
E2F3	7.355	7.37875	6.38125	5.6975	0	0	118	989	148	46
ECHDC1	11.79625	11.9525	9.3725	8.2625	41	18	133	2182	49	0
EGR2	197.86	203.23	120.2175	121.855	42	25	195	156	63	48
EIF3CL	73.66875	79.34	59.08875	63.20875	0	0	0	0	0	0
ERAP2	19.0375	21.23	13.64875	13.92625	53	48	503	1973	102	36
ETS1	124.93	129.51625	152.0525	152.275	0	22	16	614	0	0
EXOC6	6.3875	6.9475	6.75375	5.515	45	14	157	3606	0	0
EZH2	23.1625	27.0275	17.485	20.1125	0	0	226	3440	27	61
EZR	142.1775	139.6075	197.61375	199.25625	0	0	0	0	0	0
F5	25.05125	28.0925	22.38625	20.9725	50	10	1097	3190	111	39
FADS3	0.745	0.6575	1.80625	1.8075	31	4	203	886	21	57
FAM134C	6.4325	6.5225	6.8375	7.955	0	7	275	1174	98	41
FAM213B	4.1125	3.7425	2.70125	2.46625	34	2	138	169	113	0
FASLG	47.3525	47.40625	14.81625	13.61125	33	10	126	306	58	33
FBRS	16.3275	17.2125	13.19125	13.91625	47	0	372	534	71	46
FCGR3B	0.1825	0.0925	0.11	0.25125	49	73	176	331	77	42
FCMR	22.4275	17.155	30.52375	26.385	39	13	176	734	43	0
FERMT3	8.58	9.10875	6.37375	7.26125	51	11	375	853	26	0
FKRP	1.855	1.53	1.43875	2.08625	0	21	253	416	59	40
FNDC3A	22.315	24.75625	17.89125	17.79	42	34	489	9625	86	36
FNIP1	60.75625	69.25	56.19125	58.82875	47	12	460	6336	98	45
FOS	7.61125	8.2975	12.2825	12.98125	37	18	181	100	38	49
FOSL1	12.67625	13.8375	7.42	7.27625	37	10	121	323	42	0

FOSL2	20.99875	24.13125	55.6375	57.9825	0	0	157	782	113	53
FOXP1	92.74	99.51625	83.7925	86.35875	0	20	28	1234	0	0
FRYL	45.1475	47.81875	44.71625	42.94625	0	0	1148	4394	89	38
FTH1	146.36875	132.44125	644.92625	618.415	58	56	90	144	22	0
FURIN	16.84125	19.155	19.36625	25.745	57	32	61	347	0	0
FYN	96.0625	97.06	138.69625	134.32375	33	19	195	9002	60	42
GALC	9.43875	9.85875	15.05	16.03375	45	28	376	2749	90	43
GART	16.94375	18.62125	15.11	15.14375	0	33	466	1806	56	53
GFI1	19.1575	20.85375	32.08875	34.99625	24	17	196	443	73	50
GLS	85.67375	93.4625	72.92125	73.285	38	22	212	3337	205	54
GNPDA1	11.68875	12.5825	7.08625	7.9425	41	5	139	459	65	35
GPR18	13.93	16.2425	12.81	11.77625	0	0	180	0	0	0
GPR183	97.025	111.4075	154.83125	158.77125	0	0	148	0	0	0
GPR65	17.13	19.53625	32.765	31.01625	34	34	143	198	134	41
GTF3C2	8.85375	9.795	8.7425	7.94125	56	19	365	1300	33	45
HDAC7	13.81875	13.0675	8.23	9.4975	37	3	476	1919	58	0
HLA-B	1042.84625	1018.9725	902.7325	937.99125	108	7	338	332	51	92
HLA-DQB1	13.16	13.05625	15.4425	16.89125	216	12	320	1644	105	107
HLA-DRA	14.3675	13.50875	17.56	19.9175	48	12	101	231	20	56
HLA-DRB1	20.90125	18.51875	25.7075	25.9425	204	28	247	2565	92	268
HSPA1B	31.1175	29.25875	35.50625	36.9625	51	52	123	0	32	44
HYOU1	18.83	19.8	14.895	16.00125	59	8	470	806	88	49
ICAM1	18.5425	23.51	7.2875	10.28625	41	27	297	674	78	0
IDE	4.49875	5.1775	4.2775	4.02625	33	10	425	5126	105	51
IFIH1	34.91875	37.81875	24.55125	26.8375	38	23	539	2178	10	42
IFITM10	0.05	0.02875	0.06	0	65	8	60	760	156	43
IFNG	632.92	1164.26875	242.44875	496.45125	40	12	56	202	26	49
IKZF1	157.725	166.2275	141.93875	150.265	43	33	259	3715	282	60

IKZF3	73.95375	74.64375	55.75625	58.94875	27	30	208	3947	304	40
IL10	17.77	17.49875	8.9125	10.77875	38	7	74	199	67	46
IL10RB	18.57	17.5825	26.0875	24.2625	0	12	144	1424	54	36
IL13	29.64875	31.64625	41.6625	58.32375	51	3	80	126	66	49
IL18R1	9.4725	10.28875	15.7975	16.205	38	10	232	1867	91	45
IL1R2	1.74625	2.19	3.1825	3.8225	64	26	214	1921	16	36
IL2	317.105	533.93875	85.3525	167.25875	38	6	56	169	13	43
IL21	5.78625	8.26875	3.46375	3.54875	0	0	46	276	25	30
IL3	102.58125	121.96125	38.31875	49.21125	59	11	87	131	33	47
IL5	53.415	55.43	92.9825	93.4825	39	16	74	83	28	45
IL6ST	26.425	32.08375	21.76875	21.44125	0	28	395	2362	254	45
IL7R	278.1125	286.45625	488.495	446.68875	44	10	279	1011	181	43
IMPG2	0.42875	0.56125	0.25625	0.32625	39	14	587	3837	178	47
INPP5D	11.085	10.2475	13.105	12.645	67	45	462	7217	67	64
IP6K1	3.68875	3.06125	4.03875	4.3025	37	36	163	2440	130	0
IRF1	33.75	34.885	21.3625	24.0825	39	10	120	368	101	43
IRF8	7.995	9.97	2.33375	4.3175	49	4	239	1399	94	63
ITGAL	43.12125	42.7675	35.76375	33.97625	48	10	533	2192	75	0
ITGAV	1.71375	1.96625	3.65125	3.81375	0	0	457	2038	230	59
JAK2	6.28875	8.005	3.62	4.04	53	25	491	7865	72	49
KDELR2	23.2025	27.31125	18.8175	21.77125	33	21	145	1168	86	58
KIAA1109	26.9475	29.305	41.77625	39.04375	55	6	1744	7707	15	45
KIF3B	7.75625	8.555	9.09875	10.44125	29	8	315	2171	163	45
LGALS9	3.505	3.69875	1.81875	2.33375	52	14	205	994	31	60
LIF	3.82125	6.06625	2.56375	3.10625	36	13	114	144	150	46
LMNA	18.7425	17.72	11.4325	12.9325	0	40	480	2310	53	39
LPXN	73.2575	75.26875	94.035	93.22875	0	36	199	2148	24	41
LRRC32	6.1225	6.22375	10.00125	9.90875	40	12	403	416	110	64

LTA	180.585	191.72125	16.92875	25.4075	41	25	75	58	28	8
LTB	74.44375	61.1325	30.3525	32.86	43	0	122	194	10	39
MAML2	18.1675	23.7625	13.78375	15.21875	48	41	329	13643	0	0
MANBA	2.50625	2.4525	3.50125	3.59875	39	13	439	5624	49	57
MAP3K8	5.9325	6.81375	10.11	12.5525	63	29	203	1208	40	46
MAP4	43.5075	46.39375	35.295	37.99	40	21	107	6644	100	0
MAPKAPK5	7.29	8.85875	5.83375	5.99375	0	0	164	2058	14	41
MIER1	23.23625	26.93125	22.55	20.04125	0	26	237	2514	240	43
MMP9	0.0125	0.0025	0.4525	0.35125	43	6	426	363	17	48
MOB4	59.91375	72.77125	45.3025	50.10125	0	18	13	0	0	0
NAA25	13.07	15.48	10.805	11.02875	0	0	155	767	105	48
NBL1	1.0425	0.995	2.14	2.13625	0	60	29	0	0	0
NBN	11.79125	13.87	10.1675	9.87375	37	32	444	2224	86	47
NDFIP1	18.0325	20.84875	13.1175	13.46625	44	18	81	1949	116	49
NFKB1	91.39	107.44125	104.76875	119.1925	34	15	397	4935	43	42
NFKBIZ	30.7775	37.38375	36.13875	45.67	0	6	295	660	0	0
NOTCH1	4.1475	4.35	2.25625	2.80875	0	0	1547	3035	102	62
NOTCH2	12.69875	14.26	11.01125	11.92875	28	37	1070	5908	164	33
P2RY11	15.145	12.6125	8.10875	9.31875	0	16	5	0	0	0
PCBD2	2.725	2.66375	3.65	4.1425	26	1	42	2625	92	28
PDE4A	0.89	1.0825	4.2775	3.805	45	49	543	2155	179	54
PFKFB3	23.36375	28.605	50.3625	60.015	53	49	294	5342	169	74
PHACTR2	10.91125	10.8775	20.81625	17.5475	49	23	302	9283	217	0
PHTF1	4.6875	6.36875	7.50875	8.54125	43	20	365	2452	34	45
PIM3	62.5025	66.19875	28.40375	37.4125	16	23	183	131	76	51
PLAU	0.4375	0.86875	0.0975	0.11875	0	0	194	0	33	0
PLB1	0.54375	0.41	0.8125	1.19125	51	11	803	7799	0	0
PPIF	14.73	16.1125	10.6225	11.7	42	6	93	298	84	36

PRELID1	19.5225	20.8625	16.74375	16.48125	0	18	89	95	0	0
PRELID3B	40.66375	47.0375	37.325	39.11	0	0	83	0	0	0
PRKAR2A	2.42875	2.0725	2.745	3.07125	0	13	174	3631	44	41
PRKCD	6.9725	7.895	5.92	6.11375	34	19	302	1619	20	47
PSMB8	45.265	46.595	33.1225	35.29	0	8	158	170	19	23
PTGDR2	1.10875	1.0025	0.34875	0.3325	45	7	172	91	72	0
PTGER4	37.88375	44.81625	29.96875	32.84375	36	24	207	375	59	45
PTPN1	16.40875	17.24125	20.2225	22.11	50	26	159	3157	56	0
PTPRC	560.6825	578.32	861.07875	763.0225	46	18	605	4969	108	37
PWP2	3.84375	4.25375	3.02	3.00875	0	19	573	1449	40	47
RAD50	8.1875	8.9125	8.11	7.405	46	27	689	3170	0	0
RANGAP1	7.06125	6.55875	8.11875	9.1975	50	40	311	2003	102	43
RASGRP1	56.8575	62.57625	83.2575	80.73625	53	4	288	3248	94	0
RAVER1	20.0575	20.61125	17.3925	15.65	0	13	390	811	50	0
REL	59.2125	67.2825	70.33375	77.5175	0	23	254	1790	237	19
REV3L	11.78875	14.0525	11.61375	10.705	0	33	1332	7509	60	49
RIPK2	13.17	13.9025	9.665	9.8425	0	46	209	1281	45	38
RNASEH2C	5.21625	4.13	5.38125	5.35875	39	15	114	103	110	0
RNASET2	12.20125	11.275	16.155	14.61125	62	22	117	1449	5	49
RNF145	19.87	21.28625	19.74625	24.31125	49	44	235	2210	68	48
RORC	3.44	4.485	3.78625	5.72	58	38	266	1084	115	0
RPAP2	4.58375	5.33625	4.98375	3.98875	0	6	246	3200	57	50
RPAP3	6.5575	8.53375	6.65125	5.85	49	9	290	1771	74	52
RSL1D1	104.19125	121.75625	94.21375	92.0625	76	18	307	830	241	62
S100A10	161.8775	157.3075	117.51125	115.1775	26	24	43	453	19	42
S100A11	289.02375	277.97625	227.85625	213.49625	39	13	58	137	10	37
S1PR5	0.4125	0.35625	1.14625	1.11375	50	13	213	162	45	39
SDCCAG3	4.405	5.08	3.68625	3.67875	0	9	321	582	58	60

SELL	81.14875	83.26875	66.795	64.5175	45	17	183	956	61	39
SELP	0.8225	0.8025	0.555	0.48	57	7	459	1926	20	32
SENP1	16.74625	19.01125	16.04875	14.67375	0	0	48	465	97	37
SERBP1	72.44875	81.74875	57.8975	63.175	61	22	173	801	235	58
SH2B3	18.74125	18.72875	13.4325	12.44375	39	20	320	1691	153	0
SLAMF1	100.5325	110.96	83.32125	93.7975	57	31	122	872	0	0
SLAMF7	13.22125	15.43625	7.62375	8.77875	40	22	190	748	74	54
SLAMF8	0.58375	0.63375	0.22125	0.13375	0	0	2	0	0	0
SLC1A5	39.70125	41.11375	23.665	28.035	33	60	271	575	30	51
SLC22A4	1.03875	0.695	1.28	1.4725	57	15	258	661	0	0
SLC35D1	6.875	7.04875	8.35375	8.17	31	28	174	2035	209	37
SLC39A8	20.51875	23.82625	32.69875	32.6175	36	37	188	3804	176	83
SLC43A3	48.5875	77.33625	19.08625	32.26125	48	36	114	451	0	0
SLC5A6	7.8075	9.59375	7.64625	7.02625	0	11	345	630	38	58
SNAPC4	3.1775	3.1325	2.2625	2.1275	68	2	976	1491	15	58
SON	107.31375	122.93125	106.5775	110.74	0	18	1265	1139	132	0
SP100	53.21125	59.87125	45.36375	41.49875	53	21	251	2639	19	0
SP110	31.6775	30.78125	25.77375	23.79	0	27	407	2515	43	50
SP140	22.865	21.51625	18.225	17.71375	0	9	435	4125	49	42
SP140L	21.7875	20.90375	18.10875	17.19	46	17	325	3416	49	57
SPRY4	2.88375	2.56125	0.6575	0.54	0	58	203	382	206	62
STAT1	143.56625	145.75	67.1625	73.9075	51	28	229	2148	83	41
STAT4	60.9225	68.895	95.7825	102.24875	32	21	248	5229	22	42
STAT5B	50.7025	54.0725	65.30625	69.7075	31	4	273	3137	127	48
TAB2	203.6275	218.28875	195.6325	181.75875	0	0	0	890	0	0
TAGAP	86.15	85.29	60.3725	54.71875	30	35	358	369	65	46
TAP1	78.345	76.22	41.11	49.89	0	11	364	350	0	0
TAP2	26.22625	28.7575	15.9575	15.8125	35	7	351	588	198	31

TFAM	8.98125	9.7075	6.8325	7.84875	36	34	101	467	174	40
TGFBR3	7.06375	6.97	15.32375	14.27125	39	8	0	367	0	0
THEMIS	25.62	27.4025	18.9425	17.53	51	17	363	9097	71	55
TMEM50B	15.1825	14.29625	20.1425	18.9375	40	9	70	1318	79	53
TMEM9	1.59625	0.8375	1.17875	1.94625	0	0	56	831	54	62
TNF	1052.3125	1286.32625	247.55375	342.72875	9	13	96	74	45	31
TNFAIP3	56.635	66.42	90.05	91.4075	0	2	359	549	79	44
TNFRSF14	10.58125	9.91	7.5825	7.4975	0	0	188	460	147	0
TNFRSF9	15.545	16.78125	30.69125	32.84875	36	11	134	1057	220	43
TNFSF8	262.02875	285.77875	198.4675	197.96875	57	15	118	1417	160	38
TPPP	1.22125	1.27875	2.47	2.39625	47	4	144	1760	359	61
UBE2D3	147.98625	162.2975	127.3825	123.565	0	83	37	1461	0	0
USP12	19.4925	23.105	17.095	17.74125	0	15	110	2534	126	45
USP36	34.18125	36.06	49.16	45.1	0	1	723	2250	113	48
VCL	19.7675	19.55625	13.12875	12.61	46	11	467	4777	81	0
VMP1	187.17375	210.935	169.69875	170.51	0	8	131	5033	28	0
WSB1	59.98625	65.2775	96.5075	86.6825	0	35	228	842	61	41
ZDHHC11	3.8925	3.99125	7.17375	6.905	0	0	112	2091	82	38
ZFP36L1	208.62625	202.07375	138.36625	126.01625	51	54	173	373	71	38
ZFP36L2	18.19125	18.06625	34.6575	30.67	0	29	257	58	0	0
ZMIZ1	17.6675	17.57125	13.3325	13.3225	21	38	488	12034	216	58
ZNF831	1.74125	1.63125	2.62	2.31625	42	0	920	2610	195	43
ZSWIM8	12.83	10.88625	11.80625	12.9675	60	19	715	529	5	0

Table 6. Potentially deleterious exonic DSVs in DE genes and in LD with Tag-SNPs

RS ID	GENE	TAG SNP	LD R2	POLYPHEN SCORE	W POLYPHEN	SIFT SCORE	W SIFT	CAROL PREDICTION	CAROL SCORE
rs1805078	GALC	rs8005161	0.594	0.989	4.510	0.973	3.612	Deleterious	0.999
rs11556887	SP110	rs6716753	0.270	0.999	6.908	0.94	2.813	Deleterious	1.000
rs1131992	SDCCAG3	rs10781499	0.236	0.986	4.269	0.859	1.959	Deleterious	0.993
rs3812577	SDCCAG3	rs10781499	0.221	0.986	4.269	0.871	2.048	Deleterious	0.993
rs11230562	CD6	rs11230563	0.194	0.998	6.215	0.96	3.219	Deleterious	1.000
rs45604939	FNDC3A	rs2026029	0.180	0.994	5.116	0.996	5.521	Deleterious	1.000

Table 7. Promoter region DSVs in DE genes that potentially alter TF Binding sites, and are in strong LD with tag-SNPs

RS ID	GENE	REGION	TAG SNP	LD R2	TF
rs306587	MAP3K8	upstream	rs1042058	1	RXRA, XBP1, POU2F2, ELF1, SP100, GFI1, RREB1, SP4, EGR1, KLF7
rs630923	CXCR5	upstream	rs630923	1	SRF, TGIF1, ZBTB7A, PRDM1, PLAG1, EGR1, HOXB3, RXRA, VDR, MYB, CTCF, SP100, HMBOX1
rs28667727	GPR65	upstream	rs8005161	1	MAF, SMAD4, SMAD3, GCM1, HIC1, TCF12, RFX5, VDR, MYB, EGR1, SP4, SP1
rs34953890	PDE4A	upstream	rs11879191	1	GFI1, MYC, PRDM1, RXRA, PLAG1, ESR2, VDR, NANOG, FOXP3, CEBPG, ZBTB3
rs1494571	IL7R	downstream	rs3194051	1	ZNF143, SP4, ZBTB7A, SP1, EGR1, KLF7, YY1, MAF, ZBTB33, POU2F2, GFI1
rs2769346	RNASET2	upstream	rs1819333	1	MAF, ZEB1, MYC, POU2F2, FOXJ2, MTF1, FOXP1, EGR1, ZFX, GCM1, ZBTB3, HIC1
rs2149083	RNASET2	upstream	rs1819333	0.99619	VDR, MTF1, ESR2, FOXJ2, ZBTB3, MAF, FOXP1, SP1, FOXK1, ZFX, MYC, HIC1, ZEB1, POU2F2
rs2149084	RNASET2	upstream	rs1819333	0.99619	MYC, MAF, ZEB1, POU2F2, GFI1, MTF1, ZBTB3, HIC1, PRDM1, ELF1, FOXK1
rs11567685	IL7R	upstream	rs3194051	0.995244	MAF, POU2F2, ZNF143, POU6F1, HIC1, AHR, RFX5, SP4, ELF1, FOXP1, ZEB1, FOXJ2, ZFX, SP1
rs8011558	ZFP36L1	upstream	rs194749	0.994497	NANOG, BBX, RXRA, HLF, ARID5A, ARID5B, SRF, POU2F2, SMAD4, YY1, CEBPB
rs4934730	CREM	upstream	rs34779708	0.98712	ATF3, XBP1, BATF, ATF2, RFX5, RREB1, TGIF1, FOXP3, HOMEZ
rs35388511	CUL2	downstream	rs34779708	0.982761	POU2F2, RREB1, ZBTB7A, CEBPB, ATF6, EGR1, FOXJ2, FOXP1, FOXK1, ATF3, MYC, RXRA, YY1, BHLHE40
rs34815241	CUL2	downstream	rs34779708	0.982761	MTF1, RXRA, ZBTB3, CEBPB, AHR, MXI1, MYC, BHLHE40, GFI1, ZNF143, ESR2, ZBTB33, XBP1, ZEB1, VDR
rs16935880	CREM	upstream	rs34779708	0.982761	RXRA, AHR, ZEB1, GFI1, RREB1, SMAD3, ATF6, RFX5, ZBTB3, ATF3
rs9338188	MAP3K8	upstream	rs1042058	0.979507	VDR, PRDM1, SPIB, ELF1, SRF, ESR2, RFX5, GFI1, SMAD3
rs2384275	CUL2	downstream	rs34779708	0.978464	BBX, MTF1, XBP1, E4F1, GFI1, MAF, ATF3, ZBTB3, TGIF1, SMAD3, RXRA
rs3829110	SNAPC4	downstream	rs10781499	0.974351	RXRA, POU6F1, SPIB, ZFX, HOMEZ, MAF, PRDM1, BACH1, ESR2, HIC1, TGIF1, YY1, FOXK1
rs10781500	SNAPC4	downstream	rs10781499	0.974351	POU2F2, CEBPB, HIC1, YY1, ARID5A, SP100, HLF, DBP
rs3829111	SNAPC4	downstream	rs10781499	0.974351	TGIF1, MYC, EOMES, PBX3, YY1, MXI1, PLAG1, ZFX, RXRA, ARNT, MTF1, ATF2, HIC1, MAF
rs3812558	SNAPC4	downstream	rs10781499	0.969998	CEBPB, BACH1, ZBTB33, POU2F2, POU6F1, PRDM1, SP4, SRF, YY1, FOXJ2, HIC1, RXRA

rs7139746	FNDC3A	downstream	rs2026029	0.947661	GFI1, FOXJ2, RXRA, ZFX, BACH1, BACH2, MAF, SPIB, FOXK1
rs13015714	IL18R1	upstream	rs6708413	0.945823	POU2F2, FOXP3, ELF1, FOXP1, PRDM1, FOXK1, SPIB, FOXJ2, CEBPB, RXRA, HLF, RREB1, GFI1
rs2027030	KIF3B	downstream	rs6142618	0.933081	FOXP1, POU2F2, ARID5A, FOXJ2, EGR1, HOMEZ, SP4, MAF, POU6F1, BBX, HOXB3, FOXK1, LHX4
rs948788	DOK6	downstream	rs727088	0.898146	ELF1, ATF3, PRDM1, YY1, ZNF143, SPIB, RXRA, POU2F2, VDR, MAF, RFX5
rs2365358	DOK6	downstream	rs727088	0.898146	SP4, CRX, ZFX, PRDM1, MYC, EGR1, ZEB1, SP1, HIC1, KLF7, SMAD4, VDR, EOMES, FOXP1, FOXJ2
rs75203923	DOK6	downstream	rs727088	0.898146	FOXP1, ZEB1, FOXJ2, FOXK1, VDR, EGR1, BHLHE40, RXRA, SP1, MYC, SP4, ARNT, GCM1, HIC1
rs3812560	SNAPC4	downstream	rs10781499	0.85618	VDR, ESR2, MAF, SMAD4, RXRA, ELF1, ZBTB7A, CTCF, RFX5, ZNF143, MYB, SP1
rs2069776	IL2	downstream	rs7657746	0.853055	XBP1, MAF, ATF3, CEBPG, ZBTB3, BBX, GCM1, RXRA, HOMEZ, HLTF, AHR, YY1, HIC1
rs3812575	SNAPC4	upstream	rs10781499	0.840385	SRF, RXRA, MYC, PRDM1, HLTF, RREB1, ESR2, SP4, KLF7, HIC1, RFX5
rs1887428	JAK2	upstream	rs10758669	0.82493	FOXP3, EGR1, RFX5, ZEB1, VDR, AHR, RREB1, SP1, ATF3, ZNF143, BACH2
rs4957083	TPPP	downstream	rs4957048	0.818356	FOXJ2, SP4, SP1, KLF7, CTCF, EGR1, POU2F2, MYC, ZBTB7A, EOMES, ZEB1
rs72705102	TPPP	downstream	rs4957048	0.812482	VDR, ZEB1, SMAD3, E4F1, RXRA, RREB1
rs1465788	ZFP36L1	upstream	rs194749	0.808403	LHX4, RXRA, HIC1, NANOG, SMAD3, MYB, TGIF1, SRF, ZBTB7A, MYC, RREB1, MXI1, RFX5, GFI1, HLTF
rs28665408	DOK6	downstream	rs727088	0.803789	GFI1, PRDM1, CEBPB, CEBPG, YY1, HOMEZ, FOXK1, RFX5, FOXJ2, FOXP1, POU2F2, CTCF

Table 8. SNPs that potentially alter miRNA Binding sites and are in strong LD with tag-SNPs

RS ID	GENE	REGION	TAG SNP	LD R2	MIRNA	DE GENE
rs7603250	IL18RAP	UTR3	rs6708413	1	miR-148a-3p, miR-181a-5p, miR-25-3p, miR-3138	YES
rs1564823	CPEB4	UTR3	rs17695092	1	miR-206, miR-3138	YES
rs9611591	TOB2	UTR3	rs727563	1	miR-629-5p	NO
rs11648503	ZFP90	UTR3	rs1728785	1	miR-155-5p, miR-25-3p, miR-324-3p, miR-370-3p, miR-628-5p	YES
rs14316	TM9SF4	UTR3	rs6142618	1	miR-135b-3p, miR-1827, miR-196a-5p, miR-3138, miR-629-5p	NO
rs835576	NOTCH2	UTR3	rs3897478	1	miR-181a-5p, miR-206	NO
rs699779	NOTCH2	UTR3	rs3897478	1	miR-125a-5p, miR-125b-5p, miR-370-3p, miR-502-5p	NO
rs17696407	CPEB4	UTR3	rs17695092	1	miR-335-5p	YES
rs17780256	SLC39A11	UTR3	rs17780256	1	miR-1307-3p, miR-510-5p	YES
rs72812861	CPEB4	UTR3	rs17695092	1	miR-335-5p	YES
rs727088	CD226	UTR3	rs727088	1	miR-181a-5p, miR-502-5p	YES
rs7559479	IL18RAP	UTR3	rs6708413	1	miR-370-3p	YES
rs1976074	CPEB4	UTR3	rs17695092	1	miR-370-3p, miR-502-5p	YES
rs449454	NDFIP1	UTR3	rs6863411	0.995622	miR-148a-3p	YES
rs3087783	ZFP90	UTR3	rs1728785	0.993747	miR-155-5p, miR-1827, miR-196a-5p, miR-3162-5p	YES
rs60474474	PTPN2	UTR3	rs1893217	0.992134	miR-125a-5p, miR-135b-3p, miR-148a-3p, miR-155-5p, miR-25-3p, miR-324-3p, miR-628-5p	NO
rs13900	CCL2	UTR3	rs3091315	0.989913	miR-135b-3p, miR-3162-5p, miR-335-5p	NO
rs7515633	TNFRSF14	UTR3	rs6667605	0.988174	miR-148a-3p, miR-184, miR-3138	YES
rs40837	IL27	UTR3	rs26528	0.987515	miR-135b-3p, miR-185-5p, miR-3162-5p, miR-335-5p, miR-510-5p, miR-629-5p	NO
rs17591857	CREM	UTR3	rs34779708	0.987161	miR-370-3p	YES
rs17591781	CREM	UTR3	rs34779708	0.987161	miR-155-5p, miR-25-3p	YES
rs12042319	DOCK7	UTR3	rs1748195	0.986678	miR-181a-5p, miR-196a-5p, miR-628-5p	NO
rs10910090	TNFRSF14	UTR3	rs6667605	0.984193	miR-181a-5p, miR-370-3p, miR-502-5p	YES
rs41312668	SLC30A7	UTR3	rs11583043	0.980255	miR-135b-3p, miR-196a-5p	NO
rs17100939	SLC30A7	UTR3	rs11583043	0.97522	miR-184	NO

rs7118	ZFP90	UTR3	rs1728785	0.969546	miR-148a-3p, miR-155-5p, miR-206, miR-25-3p, miR-370-3p, miR-628-5p	YES
rs1878036	TSPAN14	UTR3	rs7097656	0.961809	miR-155-5p	YES
rs6061216	PLAGL2	UTR3	rs6142618	0.94972	miR-155-5p, miR-25-3p	YES
rs583121	LSM14A	UTR3	rs587259	0.947327	miR-184, miR-191-5p	NO
rs1056441	LIME1	UTR3	rs6062504	0.945447	miR-135b-3p, miR-1827, miR-191-5p, miR-196a-5p, miR-3138, miR-502-5p	NO
rs9950174	DOK6	UTR3	rs727088	0.940146	miR-155-5p, miR-3138	YES
rs202639	PHF5A	UTR3	rs727563	0.937457	miR-206	NO
rs2032933	RMI2	UTR3	rs529866	0.936696	miR-1307-3p, miR-191-5p	NO
rs1568681	IL18R1	UTR3	rs6708413	0.935242	miR-185-5p, miR-3162-5p	YES
rs39602	LNPEP	UTR3	rs1363907	0.929415	miR-135b-3p, miR-196a-5p, miR-3138, miR-3162-5p, miR-629-5p	NO
rs1045100	ATG16L1	UTR3	rs6752107	0.913202	miR-206	NO
rs174546	FADS1	UTR3	rs4246215	0.911455	miR-25-3p, miR-324-3p, miR-502-5p	NO
rs503279	FUT2	UTR3	rs516246	0.892695	miR-25-3p, miR-370-3p	NO
rs570794	FUT2	UTR3	rs516246	0.892695	miR-1307-3p, miR-148a-3p, miR-155-5p, miR-184, miR-25-3p, miR-370-3p	NO
rs571689	FUT2	UTR3	rs516246	0.892695	miR-196a-5p, miR-3162-5p, miR-510-5p, miR-629-5p	NO
rs507855	FUT2	UTR3	rs516246	0.889292	miR-125b-5p, miR-335-5p, miR-629-5p	NO
rs507766	FUT2	UTR3	rs516246	0.889292	miR-125a-5p, miR-125b-5p	NO
rs632111	FUT2	UTR3	rs516246	0.888496	miR-125b-5p, miR-510-5p	NO
rs504963	FUT2	UTR3	rs516246	0.888496	miR-184, miR-185-5p, miR-191-5p	NO
rs1130161	HLA-DQA1	UTR3	rs477515	0.887342	miR-185-5p	NO
rs17367849	POLR3H	UTR3	rs727563	0.886313	miR-1307-3p	YES
rs4625	DAG1	UTR3	rs3197999	0.883128	miR-191-5p	NO
rs603985	FUT2	UTR3	rs516246	0.879409	miR-206, miR-3162-5p	NO
rs485073	FUT2	UTR3	rs516246	0.879409	miR-206	NO
rs1857335	TNFSF15	UTR3	rs6478106	0.876764	miR-125a-5p, miR-125b-5p, miR-191-5p, miR-196a-5p, miR-324-3p, miR-370-3p, miR-502-5p	NO
rs6997	TCTA	UTR3	rs3197999	0.871181	miR-135b-3p, miR-1827, miR-196a-5p	NO
rs9538	ZC3H7B	UTR3	rs727563	0.868642	miR-125a-5p, miR-125b-5p, miR-185-5p, miR-191-5p, miR-25-3p, miR-502-5p, miR-510-5p, miR-628-5p	YES
rs3208703	KIF21B	UTR3	rs7554511	0.852169	miR-181a-5p, miR-3138, miR-324-3p, miR-370-3p, miR-628-5p	NO
rs736106	DUSP16	UTR3	rs11612508	0.847733	miR-155-5p, miR-184, miR-3138	YES
rs8139993	DESI1	UTR3	rs727563	0.847436	miR-629-5p	YES
rs7592344	MARS2	UTR3	rs1440088	0.845669	miR-185-5p	NO
rs1054609	ZPBP2	UTR3	rs12946510	0.844528	miR-125b-5p	NO
rs10457487	RSPO3	UTR3	rs9491697	0.834573	miR-135b-3p, miR-196a-5p	NO

rs699780	NOTCH2	UTR3	rs3897478	0.833398	miR-148a-3p, miR-1827, miR-502-5p, miR-510-5p	NO
rs5758364	PHF5A	UTR3	rs727563	0.832553	miR-181a-5p, miR-502-5p	NO
rs9611577	TEF	UTR3	rs727563	0.823785	miR-1827, miR-191-5p, miR-196a-5p, miR-3138, miR-324-3p, miR-370-3p	NO
rs174544	FADS1	UTR3	rs174537	0.822688	miR-1307-3p, miR-206, miR-502-5p	NO
rs13001714	IL1RL1	UTR3	rs13001325	0.820097	miR-155-5p	YES
rs12712142	IL1RL1	UTR3	rs13001325	0.820097	miR-3138	YES
rs72707016	TPPP	UTR3	rs4957048	0.818356	miR-206	YES
rs28364691	TPPP	UTR3	rs4957048	0.818356	miR-181a-5p, miR-324-3p, miR-370-3p, miR-502-5p, miR-510-5p	YES
rs7558	TPPP	UTR3	rs4957048	0.818356	miR-1307-3p, miR-510-5p	YES
rs3762951	TPPP	UTR3	rs4957048	0.818356	miR-125a-5p, miR-125b-5p, miR-148a-3p, miR-206, miR-324-3p, miR-370-3p, miR-502-5p, miR-510-5p	YES
rs1260631	LSM14A	UTR3	rs587259	0.818352	miR-125a-5p, miR-125b-5p	NO
rs1790974	DOK6	UTR3	rs727088	0.815421	miR-206, miR-3138	YES
rs72707007	TPPP	UTR3	rs4957048	0.812482	miR-125a-5p	YES

Table 9. DSVs that alter potentially alter splice-sites and are in LD with tag-SNPs

RS ID	GENE	FUNCTION	TAG SNP	LD R2
rs8373	ZFP91	splicing	rs11229555	1
rs80212515	MUC19	splicing	rs11564258	1
rs11078928	GSDMB	splicing	rs12946510	0.825071
rs80212515	MUC19	splicing	rs148319899	0.81926
rs2004640	IRF5	splicing	rs4728142	0.654786

4.3.6 Effects of PGE2 on expression of isoforms in genes inside IBD-associated regions

An important aspect of gene regulation is the differential expression of different gene isoforms (splice variants) in response to different stimuli. For example, genes expressed in many cancers seem to have shorter 3'-UTRs (Mayr & Bartel, 2009), possibly avoiding miRNA silencing. We used RSEM to detect isoforms in our samples and edgeR to identify the differentially expressed isoforms. Overall, we found 1,122 genes with 1,349 isoforms expressed at different levels in at least one of the group comparisons (edgeR multigroup comparison / FDR < 0.05) (**Table 4**). In agreement with the gene model-based results described above, PGE2 seems to be the main factor contributing to differential expression of isoforms, since the comparison **PGE2-treated vs non-PGE2** treated cells give the highest number of differentially expressed isoforms (**Table 4**). CREM, for example, is a gene with significant over-expression of isoforms NM_182720 (Refseq variant 7), NM_182717 (Refseq variant 4), NM_182719 (Refseq variant 6), NM_182718 (Refseq variant 5), NM_182723 (Refseq variant 10) in PGE2-stimulated cells. CREM plays a key role in the regulation of immune responses, including the Th17 immune response (Hedrich, Crispin, et al., 2012; Hedrich, Rauen, Kis-Toth, Kyttaris, & Tsokos, 2012; Koga et al., 2014; Lippe et al., 2012; Rauen, Hedrich, Juang, Tenbrock, & Tsokos, 2011; Rauen, Hedrich, Tenbrock, & Tsokos, 2013). The expression of different CREM isoforms is likely to be the product of activation of distinct promoters. Therefore, we assessed whether any SNPs inside the different CREM promoters can alter transcription factor binding sites. Two SNPs located in the various CREM promoters were predicted to alter one or more transcription factor binding sites and were in LD with IBD tag SNPs (**Table 7**). Other immune response genes with differential expression of isoforms include TNFAIP3, IL4R, CCR4 and CCL4L2.

4.4 DISCUSSION

Our results show that PGE2 induces differential expression of more genes in activated effector memory T cells than interleukins IL1β and IL23 combined, and that its presence/absence drives major differences in gene expression with or without co-stimulation by IL23 and IL1β. Differential expression analysis reveals that PGE2 changes the expression levels of many genes relevant in IBD. In addition to genes differentially expressed in response to PGE2, we also found that many genes express alternatively spliced isoforms, such as CREM, which is known to play a role in modulating key cytokine gene transcription. Since PGE2 is a key mediator in IBD, this finding points to a potential mechanism for IBD facilitation.

We also performed extensive analysis to identify genomic variants that can potentially affect the function of genes that are regulated by PGE2 and potentially relevant in IBD. Our goal was to identify a tractable number of DSVs that are potentially functional. Therefore, we took an innovative data-driven approach which combines RNA-Seq data, publicly available GWAS results, and state-of-the-art computational tools.

Only a small number of non-synonymous exonic variants that were in moderated linkage-disequilibrium with IBD tag-SNPs were identified. These findings corroborate the observation that DSVs in non-coding, presumably regulatory regions of the genome comprise the majority of DSVs associated with most complex diseases (Hindorff et al., 2009).

Therefore, we focused our analysis on the identification of regulatory functional variants. In the 3'UTR of DE genes, we found 18 DSVs that potentially alter miRNA binding sites and are in strong LD with IBD-tag SNPs. Two of these DSVs (rs7559479, rs7118) have already been associated with other complex diseases or modulation of immune response. The presence of SNP rs7559479, located in the 3'UTR of the IL18RAP gene, was statistically associated with higher expression of IL18. This is in agreement with our analysis which shows that rs7559479 disrupts binding of miRNA has-miR-370-3p, thus altering expression of IL18 (Martinez-Hervas et al., 2015). A recent study has shown that SNP rs7118 in the 3'UTR of ZFP90 is statistically associated with UC and that it potentially disrupts microRNA binding sites (Arnold, Ellwanger, Hartsperger, Pfeufer, & Stumpflen, 2012).

We also identified DSVs in upstream, downstream, 5'UTR and intronic regions that potentially alter transcription factor binding sites. Several of the identified DSVs (rs63093, rs1494571, rs11567685 and rs13015714) have been shown to be associated with other complex diseases. SNP rs63093 has been shown to create a transcription factor binding site for MEF2C, thus reducing CXCR5 promoter activity (Mitkin, Muratova, Schwartz, & Kuprash, 2016). Our analysis did not consider TF MEF2C, as this transcription factor is not expressed in our samples. However, this does not invalidate our results as DSVs can alter binding for multiple TFs at the same time. DSV rs1494571 has been shown to be statistically associated with modulation of vaccine-induced immunity to measles (Haralambieva et al., 2011) and with lymphocyte development in non-hodgkin lymphoma (Schuetz et al., 2013). Studies have shown an association between snp rs511567685 and multiple sclerosis in Iranian and Jordanian patients (Haj, Nikravesh, Kakhki, & Rakhshi, 2015; Ibayyan et al., 2014), whereas rs13015714 is associated with IBD and celiac disease (Koskinen et al., 2009; Latiano et al., 2013). The activity of IL18R1 was shown to have a dose-dependent correlation with rs13015714, which could be an effect of TF binding alteration (Koskinen et al., 2009).

5.0 CONCLUSIONS, LIMITATIONS AND FUTURE WORK

5.1 CONCLUSIONS

Functional annotation of DSVs is crucial to the understanding of the etiology of complex disease. Computational methods and pipelines capable of analyzing and predicting the function of DSVs from large omics datasets are in great need. This dissertation has presented a novel algorithm that can accurately predict the effect of 3'UTR DSVs in miRNA binding. A new pipeline for the functional prioritization of DSVs obtained from omics data is also described and tested.

The is-mirSNP algorithm uses a novel approach for predicting the effect of 3'UTR DSVs on miRNA binding. The energy-binding approach used by is-miRSNP allows for the empirical calculation of background distributions that are the foundation of a statistically sound approach to the problem. The results obtained from is-mirSNP are easy to interpret, and greatly facilitate the task of variant prioritization. Experiments performed using an unbiased, manually curated validation set of experimentally validated DSVs-miRNA showed that is-mirSNP outperform all other existing methods The pipeline for functional annotation of DSVs presented in this dissertation utilizes a LD-based approach for the identification of variants related to the problem being studied. The pipeline can analyze RNA-seq data from raw data, and uses state-of-the-art computational tools for assigning function to DSVs. The analysis pipeline here presented can be.

applied to any set of RNA-Seq and GWAS datasets, requiring only the modification of parameters unique to the design of the data being analyzed. All algorithms and methods used in the pipeline are of publicly available, and can, therefore, be utilized by any user. The pipeline performance was tested in an IBD-related dataset. RNA-Seq analyses identified biologically relevant genes and isoforms that are differentially expressed in the presence of PGE2. Functional annotation of DSVs assigned function to variants already known to be related to IBD, and new DSVs related to genes and processes potentially relevant in IBD were identified and functionally annotated.

In conclusion, this dissertation presented novel computational methods that can accurately assign function to DSVs related to diseases. These algorithms use novel approaches and solve existing informatics problems. The impact of the work described here lies in the future use and application of these computational tools to new and existing biological datasets.

5.2 LIMITATIONS

The results presented in this dissertation suggest that is-miRSNP is an accurate method for predicting DSVs that affect miRNA binding sites. The following limitations need to be taken into consideration when interpreting the results here presented:

a. The validation dataset used to estimate the accuracy of is-miRSNP contains only 23 variants. The small size of the validation dataset is a direct consequence of the lack of studies that experimentally validated the effect of DSVs on miRNA binding.

b. The false-positive rate of the is-miRSNP algorithm could not be calculated due to the lack of true-negative DSV-miRNA data. This is the consequence of a well-known problem in the biomedical literature: papers are biased towards positive results and rarely report failed experiments (Dickersin et al., 1987)

The proposed pipeline for prioritization and functional evaluation of DSVs could identify variants that are potentially biologically meaningful in IBD, but the following limitations need to be considered:

a. Pipeline requires data from a biological meaningful GWAS study, which might not be available.

b. Functionally annotated DSVs were not experimentally validated.

5.3 FUTURE WORK

The work presented in this dissertation directly leads to the following future directions:

5.3.1 Expand is-miRSNP to predict effect of INDELS

The insertion or the deletion of bases in DNA (INDEL) are a common type of genetic variation although not as common as single nucleotide variations (Gudbjartsson et al., 2015). INDELs have been implicated in complex disease, and can have similar effects as those shown by single-point mutations, as they can potentially affect protein coding sequences and alter miRNA and TF binding sites. Therefore, the functional annotation of INDELs is important in the understanding of disease etiology. Currently, the is-miRSNP algorithm can only predict the effect of 3’UTR single point variation, and the expansion of the algorithm so that it can accurately predict the effect of INDELs in 3’UTR would greatly enhance its applications. Modifications in is-miRSNP so that it can support the prediction of INDELs would mainly involve the adjustment of the length of mutated and reference sequences so that it can utilize the same background distributions currently used by SNPs. Correction of the sequence length for binding energy calculation is crucial as binding energy will change proportionately with the size of the sequence.

5.3.2 Investigate is-miRSNP background distributions

The is-miRSNP algorithm requires the pre-computation of empirical background distributions, and the computation of such distributions are the most intense computational step of the algorithm. The background binding distributions and the distributions of the log-ratio of the p-value ratios need to be calculated for every new miRNA added to the predictions. Therefore, the algorithm performance and usability would improve if empirical distributions could be modelled and generalized. Modelling and generalization of the distributions would require further investigation of the similarities and differences in the behavior of the binding of the different miRNAs.

5.3.3 Expand is-miRSNP prediction to other organisms

Pre-clinical and translational medical research is often done in animal models, therefore an algorithm capable of predict the effect of 3’UTR DSV on miRNA binding would greatly benefit the scientific community. Organism-specific background distribution for other organisms needs to be computed before the algorithm is up to the task. In addition, datasets of experimentally validated variants that affect miRNA binding will need to be evaluated to assess the algorithm performance in different organisms.

5.3.4 Expand pipeline for functional prioritization of DSVs to integrate different omics data types

The pipeline for functional annotation of DSVs would benefit from the integration of other omics data types. The ENCODE project has shown that RNA production and processing is quantitatively correlated with transcription factor binding at gene promoters and with the state of the chromatin (Consortium, 2012) Therefore, integration of ChiP-Seq and ATAC-Seq results to the pipeline would provide a more accurate picture of the current transcriptional state of the samples, that would ultimately lead to better functional predictions of identified DSVs.

5.3.5 Add variant calling step to RNA-Seq analysis

Recent work has shown that variants can be accurately called from RNA-Seq data (Piskol, Ramaswami, & Li, 2013; Sun, Bhagwate, Prodduturi, Yang, & Kocher, 2016; C. Wang et al., 2014). Adding a variant-calling step to the pipeline for functional prioritization of DSVs would leverage the usage of RNA-Seq data, and help in the integration of RNA-Seq results with existing GWAS datasets. A variant calling step would specially benefit the case where disease-normal paired samples (i.e. tumor-normal pairs, or disease-control samples) are being analyzed as variants present in normal samples could be immediately removed from analysis and variants present only in disease could be assigned higher importance.

BIBLIOGRAPHY

Adamik, J., Henkel, M., Ray, A., Auron, P. E., Duerr, R., & Barrie, A. (2013). The IL17A and IL17F loci have divergent histone modifications and are differentially regulated by prostaglandin E2 in Th17 cells. *Cytokine, 64*(1), 404-412. doi:10.1016/j.cyto.2013.05.010

Adams, B. D., Furneaux, H., & White, B. A. (2007). The micro-ribonucleic acid (miRNA) miR-206 targets the human estrogen receptor-alpha (ERalpha) and represses ERalpha messenger RNA and protein expression in breast cancer cell lines. *Mol Endocrinol, 21*(5), 1132-1147. doi:10.1210/me.2007-0022

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Ambros, V. (2004). The functions of animal microRNAs. *Nature, 431*(7006), 350-355. doi:10.1038/nature02871

Ananthakrishnan, A. N. (2015). Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol, 12*(4), 205-217. doi:10.1038/nrgastro.2015.34

Andrews, S. (2012). FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Arnold, M., Ellwanger, D. C., Hartsperger, M. L., Pfeufer, A., & Stumpflen, V. (2012). Cis-acting polymorphisms affect complex traits through modifications of microRNA regulation pathways. *PLoS One, 7*(5), e36694. doi:10.1371/journal.pone.0036694

Baralle, D., & Baralle, M. (2005). Splicing in action: assessing disease causing sequence changes. *J Med Genet, 42*(10), 737-748. doi:10.1136/jmg.2004.029538

Barrie, A., Khare, A., Henkel, M., Zhang, Y., Barmada, M. M., Duerr, R., & Ray, A. (2011). Prostaglandin E2 and IL-23 plus IL-1beta differentially regulate the Th1/Th17 immune response of human CD161(+) CD4(+) memory T cells. *Clin Transl Sci, 4*(4), 268-273. doi:10.1111/j.1752-8062.2011.00300.x

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell, 116*(2), 281-297.

Bhattacharya, A., Ziebarth, J. D., & Cui, Y. (2014). PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. *Nucleic Acids Res, 42*(Database issue), D86-91. doi:10.1093/nar/gkt1028

Brookes, A. J. (1999). The essence of SNPs. *Gene, 234*(2), 177-186.

Buckland, P. R. (2006). The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochim Biophys Acta, 1762*(1), 17-28. doi:10.1016/j.bbadis.2005.10.004

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Comput Biol, 8*(12), e1002822. doi:10.1371/journal.pcbi.1002822

Cheng, M., Yang, L., Yang, R., Yang, X., Deng, J., Yu, B., . . . Lu, J. (2013). A microRNA-135a/b binding polymorphism in CD133 confers decreased risk and favorable prognosis of lung cancer in Chinese by reducing CD133 expression. *Carcinogenesis, 34*(10), 2292-2299. doi:10.1093/carcin/bgt181

Chou, C. H., Chang, N. W., Shrestha, S., Hsu, S. D., Lin, Y. L., Lee, W. H., . . . Huang, H. D. (2016). miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res, 44*(D1), D239-247. doi:10.1093/nar/gkv1258

Clancy, S. (2008). RNA splicing: introns, exons and spliceosome. *Nature Education, 1*(1).

Clark, P. M., Loher, P., Quann, K., Brody, J., Londin, E. R., & Rigoutsos, I. (2014). Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. *Sci Rep, 4*, 5947. doi:10.1038/srep05947

Cobb, J., Busst, C., Petrou, S., Harrap, S., & Ellis, J. (2008). Searching for functional genetic variants in non-coding DNA. *Clin Exp Pharmacol Physiol, 35*(4), 372-375. doi:10.1111/j.1440-1681.2008.04880.x

Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:10.1038/nature11247

Coronnello, C., Hartmaier, R., Arora, A., Huleihel, L., Pandit, K. V., Bais, A. S., . . . Benos, P. V. (2012). Novel modeling of combinatorial miRNA targeting identifies SNP with potential role in bone density. *PLoS Comput Biol, 8*(12), e1002830. doi:10.1371/journal.pcbi.1002830

Danan, C., Manickavel, S., & Hafner, M. (2016). PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites. *Methods Mol Biol, 1358*, 153-173. doi:10.1007/978-1-4939-3067-8_10

Darnell, R. B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA, 1*(2), 266-286. doi:10.1002/wrna.31

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., . . . Barrett, J. C. (2017). Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*. doi:10.1038/ng.3760

Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation. *Cell, 166*(3), 538-554. doi:10.1016/j.cell.2016.07.012

Deveci, M., Catalyurek, U. V., & Toland, A. E. (2014). mrSNP: software to detect SNP effects on microRNA binding. *BMC Bioinformatics, 15*, 73. doi:10.1186/1471-2105-15-73

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith, H., Jr. (1987). Publication bias and clinical trials. *Control Clin Trials, 8*(4), 343-353.

Diehl, A. G., & Boyle, A. P. (2016). Deciphering ENCODE. *Trends Genet, 32*(4), 238-249. doi:10.1016/j.tig.2016.02.002

Elek, Z., Nemeth, N., Nagy, G., Nemeth, H., Somogyi, A., Hosszufalusi, N., . . . Ronai, Z. (2015). Micro-RNA Binding Site Polymorphisms in the WFS1 Gene Are Risk Factors of Diabetes Mellitus. *PLoS One, 10*(10), e0139519. doi:10.1371/journal.pone.0139519

Ellwanger, D. C., Buttner, F. A., Mewes, H. W., & Stumpflen, V. (2011). The sufficient minimal set of miRNA seed types. *Bioinformatics, 27*(10), 1346-1350. doi:10.1093/bioinformatics/btr149

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biol, 5*(1), R1. doi:10.1186/gb-2003-5-1-r1

Feng, N., Xu, B., Tao, J., Li, P., Cheng, G., Min, Z., . . . Hua, L. (2012). A miR-125b binding site polymorphism in bone morphogenetic protein membrane receptor type IB gene and prostate cancer risk in China. *Mol Biol Rep, 39*(1), 369-373. doi:10.1007/s11033-011-0747-9

Fonseca-Camarillo, G., & Yamamoto-Furusho, J. K. (2015). Immunoregulatory Pathways Involved in Inflammatory Bowel Disease. *Inflamm Bowel Dis*. doi:10.1097/MIB.0000000000000477

Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res, 19*(1), 92-105. doi:10.1101/gr.082701.108

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi:10.1038/nature09534

Geremia, A., Biancheri, P., Allan, P., Corazza, G. R., & Di Sabatino, A. (2014). Innate and adaptive immunity in inflammatory bowel disease. *Autoimmun Rev, 13*(1), 3-10. doi:10.1016/j.autrev.2013.06.004

Goda, N., Murase, H., Kasezawa, N., Goda, T., & Yamakawa-Kobayashi, K. (2015). Polymorphism in microRNA-binding site in HNF1B influences the susceptibility of type 2 diabetes mellitus: a population based case-control study. *BMC Med Genet, 16*, 75. doi:10.1186/s12881-015-0219-5

Gong, J., Liu, C., Liu, W., Wu, Y., Ma, Z., Chen, H., & Guo, A. Y. (2015). An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools. *Database (Oxford), 2015*, bav029. doi:10.1093/database/bav029

Gudbjartsson, D. F., Sulem, P., Helgason, H., Gylfason, A., Gudjonsson, S. A., Zink, F., . . . Stefansson, K. (2015). Sequence variants from whole genome sequencing a large group of Icelanders. *Sci Data, 2*, 150011. doi:10.1038/sdata.2015.11

Ha, M., & Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol, 15*(8), 509-524. doi:10.1038/nrm3838

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., . . . Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell, 141*(1), 129-141. doi:10.1016/j.cell.2010.03.009

Haj, M. S., Nikravesh, A., Kakhki, M. P., & Rakhshi, N. (2015). Association study of four polymorphisms in the interleukin-7 receptor alpha gene with multiple sclerosis in Eastern Iran. *Iran J Basic Med Sci, 18*(6), 593-598.

Haralambieva, I. H., Ovsyannikova, I. G., Kennedy, R. B., Vierkant, R. A., Pankratz, V. S., Jacobson, R. M., & Poland, G. A. (2011). Associations between single nucleotide polymorphisms and haplotypes in cytokine and cytokine receptor genes and immunity to measles vaccination. *Vaccine, 29*(45), 7883-7895. doi:10.1016/j.vaccine.2011.08.083

Hedrich, C. M., Crispin, J. C., Rauen, T., Ioannidis, C., Apostolidis, S. A., Lo, M. S., . . . Tsokos, G. C. (2012). cAMP response element modulator alpha controls IL2 and IL17A expression during CD4 lineage commitment and subset distribution in lupus. *Proc Natl Acad Sci U S A, 109*(41), 16606-16611. doi:10.1073/pnas.1210129109

Hedrich, C. M., Rauen, T., Kis-Toth, K., Kyttaris, V. C., & Tsokos, G. C. (2012). cAMP-responsive element modulator alpha (CREMalpha) suppresses IL-17F protein expression in T lymphocytes from patients with systemic lupus erythematosus (SLE). *J Biol Chem, 287*(7), 4715-4725. doi:10.1074/jbc.M111.323261

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A, 106*(23), 9362-9367. doi:10.1073/pnas.0903103106

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA, 8*(1). doi:10.1002/wrna.1364

Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., . . . Huang, H. D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res, 39*(Database issue), D163-169. doi:10.1093/nar/gkq1107

Hudson, T. J. (2003). Wanted: regulatory SNPs. *Nat Genet, 33*(4), 439-440. doi:10.1038/ng0403-439

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nat Rev Genet, 6*(4), 287-298. doi:10.1038/nrg1578

Ibayyan, L., Zaza, R., Dahbour, S., El-Omar, A., Samhouri, B., El-Khateeb, M., & Ahram, M. (2014). The promoter SNP, but not the alternative splicing SNP, is linked to multiple sclerosis among Jordanian patients. *J Mol Neurosci, 52*(4), 467-472. doi:10.1007/s12031-013-0151-0

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., . . . Cho, J. H. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature, 491*(7422), 119-124. doi:10.1038/nature11582

Kannan, L., Ramos, M., Re, A., El-Hachem, N., Safikhani, Z., Gendoo, D. M., . . . Waldron, L. (2016). Public data and open source tools for multi-assay genomic investigation of disease. *Brief Bioinform, 17*(4), 603-615. doi:10.1093/bib/bbv080

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res, 32*(Database issue), D493-496. doi:10.1093/nar/gkh103

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., . . . Hardison, R. C. (2014). Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A, 111*(17), 6131-6138. doi:10.1073/pnas.1318948111

Knight, J. C. (2005). Regulatory polymorphisms underlying complex disease traits. *J Mol Med (Berl), 83*(2), 97-109. doi:10.1007/s00109-004-0603-7

Koga, T., Hedrich, C. M., Mizui, M., Yoshida, N., Otomo, K., Lieberman, L. A., . . . Tsokos, G. C. (2014). CaMK4-dependent activation of AKT/mTOR and CREM-alpha underlies autoimmunity-associated Th17 imbalance. *J Clin Invest, 124*(5), 2234-2245. doi:10.1172/JCI73411

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., . . . Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol, 17*(7), 909-915. doi:10.1038/nsmb.1838

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., . . . Ule, J. (2011). iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J Vis Exp*(50). doi:10.3791/2638

Koskinen, L. L., Einarsdottir, E., Dukes, E., Heap, G. A., Dubois, P., Korponay-Szabo, I. R., . . . Saavalainen, P. (2009). Association study of the IL18RAP locus in three European populations with coeliac disease. *Hum Mol Genet, 18*(6), 1148-1155. doi:10.1093/hmg/ddn438

Krek, A., Grun, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., . . . Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nat Genet, 37*(5), 495-500. doi:10.1038/ng1536

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harb Protoc, 2015*(11), 951-969. doi:10.1101/pdb.top084970

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi:10.1038/nprot.2009.86

Landi, D., Barale, R., Gemignani, F., & Landi, S. (2011). Prediction of the biological effect of polymorphisms within microRNA binding sites. *Methods Mol Biol, 676*, 197-210. doi:10.1007/978-1-60761-863-8_14

Landi, D., Gemignani, F., Pardini, B., Naccarati, A., Garritano, S., Vodicka, P., . . . Landi, S. (2012). Identification of candidate genes carrying polymorphisms associated with the risk of colorectal cancer by analyzing the colorectal mutome and microRNAome. *Cancer, 118*(19), 4670-4680. doi:10.1002/cncr.27435

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Latchman, D. S. (1997). Transcription factors: an overview. *Int J Biochem Cell Biol, 29*(12), 1305-1312.

Latiano, A., Palmieri, O., Pastorelli, L., Vecchi, M., Pizarro, T. T., Bossa, F., . . . Andriulli, A. (2013). Associations between genetic polymorphisms in IL-33, IL1R1 and risk for inflammatory bowel disease. *PLoS One, 8*(4), e62144. doi:10.1371/journal.pone.0062144

Lee, L. A., Arvai, K. J., & Jones, D. (2015). Annotation of Sequence Variants in Cancer Samples: Processes and Pitfalls for Routine Assays in the Clinical Laboratory. *J Mol Diagn, 17*(4), 339-351. doi:10.1016/j.jmoldx.2015.03.003

Lee, S. Y., Choi, J. E., Jeon, H. S., Hong, M. J., Choi, Y. Y., Kang, H. G., . . . Park, J. Y. (2015). A genetic variation in microRNA target site of KRT81 gene is associated with survival in early-stage non-small-cell lung cancer. *Ann Oncol, 26*(6), 1142-1148. doi:10.1093/annonc/mdv100

Lees, C. W., Barrett, J. C., Parkes, M., & Satsangi, J. (2011). New IBD genetics: common pathways with other diseases. *Gut, 60*(12), 1739-1753. doi:10.1136/gut.2009.199679

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., . . . Kendziorski, C. (2013). EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics, 29*(8), 1035-1043. doi:10.1093/bioinformatics/btt087

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell, 120*(1), 15-20. doi:10.1016/j.cell.2004.12.035

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi:10.1186/1471-2105-12-323

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., . . . Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature, 456*(7221), 464-469. doi:10.1038/nature07488

Lin, Y., Nie, Y., Zhao, J., Chen, X., Ye, M., Li, Y., . . . Li, Y. (2012). Genetic polymorphism at miR-181a binding site contributes to gastric cancer susceptibility. *Carcinogenesis, 33*(12), 2377-2383. doi:10.1093/carcin/bgs292

Lippe, R., Ohl, K., Varga, G., Rauen, T., Crispin, J. C., Juang, Y. T., . . . Tenbrock, K. (2012). CREMalpha overexpression decreases IL-2 production, induces a T(H)17 phenotype and accelerates autoimmunity. *J Mol Cell Biol, 4*(2), 121-123. doi:10.1093/jmcb/mjs004

Liu, C., Zhang, F., Li, T., Lu, M., Wang, L., Yue, W., & Zhang, D. (2012). MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics, 13*, 661. doi:10.1186/1471-2164-13-661

Liu, H., Gao, F., Dahlstrom, K. R., Li, G., Sturgis, E. M., Zevallos, J. P., . . . Liu, Z. (2016). A variant at a potentially functional microRNA-binding site in BRIP1 was associated with risk of squamous cell carcinoma of the head and neck. *Tumour Biol, 37*(6), 8057-8066. doi:10.1007/s13277-015-4682-6

Liu, Z., Wei, S., Ma, H., Zhao, M., Myers, J. N., Weber, R. S., . . . Wei, Q. (2011). A functional variant at the miR-184 binding site in TNFAIP2 and risk of squamous cell carcinoma of the head and neck. *Carcinogenesis, 32*(11), 1668-1674. doi:10.1093/carcin/bgr209

Liu, Z. J., Yadav, P. K., Su, J. L., Wang, J. S., & Fei, K. (2009). Potential role of Th17 cells in the pathogenesis of inflammatory bowel disease. *World J Gastroenterol, 15*(46), 5784-5788.

Lopes, M. C., Joyce, C., Ritchie, G. R., John, S. L., Cunningham, F., Asimit, J., & Zeggini, E. (2012). A combined functional annotation score for non-synonymous variants. *Hum Hered, 73*(1), 47-51. doi:10.1159/000334984

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Macarthur, D. G. (2012). Challenges in clinical genomics. *Genome Med, 4*(5), 43. doi:10.1186/gm342

Macintyre, G., Bailey, J., Haviv, I., & Kowalczyk, A. (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics, 26*(18), i524-530. doi:10.1093/bioinformatics/btq378

Maragkakis, M., Reczko, M., Simossis, V. A., Alexiou, P., Papadopoulos, G. L., Dalamagas, T., . . . Hatzigeorgiou, A. G. (2009). DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res, 37*(Web Server issue), W273-276. doi:10.1093/nar/gkp292

Marcais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics, 27*(6), 764-770. doi:10.1093/bioinformatics/btr011

Martin, M. (2011). Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet.journal, 17*(1). doi:http://dx.doi.org/10.14806/ej.17.1.200

Martinez-Hervas, S., Martinez-Barquero, V., Nunez Savall, E., Lendinez, V., Olivares, L., Benito, E., . . . Ascaso, J. F. (2015). [Plasma IL-18 levels are related to insulin and are modulated by IL-18 gene polymorphisms]. *Clin Investig Arterioscler, 27*(6), 265-271. doi:10.1016/j.arteri.2015.04.004

Mayr, C., & Bartel, D. P. (2009). Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell, 138*(4), 673-684. doi:10.1016/j.cell.2009.06.016

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res, 40*(10), 4288-4297. doi:10.1093/nar/gks042

Menard, C., Rezende, F. A., Miloudi, K., Wilson, A., Tetreault, N., Hardy, P., . . . Sapieha, P. (2016). MicroRNA signatures in vitreous humour and plasma of patients with exudative AMD. *Oncotarget, 7*(15), 19171-19184. doi:10.18632/oncotarget.8280

Miller, I. J., & Bieker, J. J. (1993). A novel, erythroid cell-specific murine transcription factor that binds to the CACCC element and is related to the Kruppel family of nuclear proteins. *Mol Cell Biol, 13*(5), 2776-2786.

Mishra, P. J., Humeniuk, R., Mishra, P. J., Longo-Sorbello, G. S., Banerjee, D., & Bertino, J. R. (2007). A miR-24 microRNA binding-site polymorphism in dihydrofolate reductase gene leads to methotrexate resistance. *Proc Natl Acad Sci U S A, 104*(33), 13513-13518. doi:10.1073/pnas.0706217104

Mishra, P. J., Mishra, P. J., Banerjee, D., & Bertino, J. R. (2008). MiRSNPs or MiR-polymorphisms, new players in microRNA mediated regulation of the cell: Introducing microRNA pharmacogenomics. *Cell Cycle, 7*(7), 853-858. doi:10.4161/cc.7.7.5666

Mitkin, N. A., Muratova, A. M., Schwartz, A. M., & Kuprash, D. V. (2016). The A Allele of the Single-Nucleotide Polymorphism rs630923 Creates a Binding Site for MEF2C Resulting in Reduced CXCR5 Promoter Activity in B-Cell Lymphoblastic Cell Lines. *Front Immunol, 7*, 515. doi:10.3389/fimmu.2016.00515

Mohamed, A. M., Thenoz, M., Solly, F., Balsat, M., Mortreux, F., & Wattel, E. (2014). How mRNA is misspliced in acute myelogenous leukemia (AML)? *Oncotarget, 5*(20), 9534-9545. doi:10.18632/oncotarget.2304

Nicoloso, M. S., Sun, H., Spizzo, R., Kim, H., Wickramasinghe, P., Shimizu, M., . . . Calin, G. A. (2010). Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility. *Cancer Res, 70*(7), 2789-2798. doi:10.1158/0008-5472.CAN-09-3541

Orkin, S. H., Kazazian, H. H., Jr., Antonarakis, S. E., Goff, S. C., Boehm, C. D., Sexton, J. P., . . . Giardina, P. J. (1982). Linkage of beta-thalassaemia mutations and beta-globin gene polymorphisms with DNA polymorphisms in human beta-globin gene cluster. *Nature, 296*(5858), 627-631.

Pampin, S., & Rodriguez-Rey, J. C. (2007). Functional analysis of regulatory single-nucleotide polymorphisms. *Curr Opin Lipidol, 18*(2), 194-198. doi:10.1097/MOL.0b013e3280145093

Pazin, M. J. (2015). Using the ENCODE Resource for Functional Annotation of Genetic Variants. *Cold Spring Harb Protoc, 2015*(6), 522-536. doi:10.1101/pdb.top084988

Piskol, R., Ramaswami, G., & Li, J. B. (2013). Reliable identification of genomic variants from RNA-seq data. *Am J Hum Genet, 93*(4), 641-651. doi:10.1016/j.ajhg.2013.08.008

Podolsky, D. K. (2002). Inflammatory bowel disease. *N Engl J Med, 347*(6), 417-429. doi:10.1056/NEJMra020831

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Qu, H., & Fang, X. (2013). A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics Proteomics Bioinformatics, 11*(3), 135-141. doi:10.1016/j.gpb.2013.05.001

Rauen, T., Hedrich, C. M., Juang, Y. T., Tenbrock, K., & Tsokos, G. C. (2011). cAMP-responsive element modulator (CREM)alpha protein induces interleukin 17A expression and mediates epigenetic alterations at the interleukin-17A gene locus in patients with systemic lupus erythematosus. *J Biol Chem, 286*(50), 43437-43446. doi:10.1074/jbc.M111.299313

Rauen, T., Hedrich, C. M., Tenbrock, K., & Tsokos, G. C. (2013). cAMP responsive element modulator: a critical regulator of cytokine production. *Trends Mol Med, 19*(4), 262-269. doi:10.1016/j.molmed.2013.02.001

Rehmsmeier, M., Steffen, P., Hochsmann, M., & Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA, 10*(10), 1507-1517. doi:10.1261/rna.5248604

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res, 43*(7), e47. doi:10.1093/nar/gkv007

Rusinov, V., Baev, V., Minkov, I. N., & Tabler, M. (2005). MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res, 33*(Web Server issue), W696-700. doi:10.1093/nar/gki364

Sartor, R. B. (2006). Mechanisms of disease: pathogenesis of Crohn's disease and ulcerative colitis. *Nat Clin Pract Gastroenterol Hepatol, 3*(7), 390-407. doi:10.1038/ncpgasthep0528

Scacheri, C. A., & Scacheri, P. C. (2015). Mutations in the noncoding genome. *Curr Opin Pediatr, 27*(6), 659-664. doi:10.1097/MOP.0000000000000283

Schuetz, J. M., Daley, D., Leach, S., Conde, L., Berry, B. R., Gallagher, R. P., . . . Brooks-Wilson, A. R. (2013). Non-Hodgkin lymphoma risk and variants in genes controlling lymphocyte development. *PLoS One, 8*(9), e75170. doi:10.1371/journal.pone.0075170

Sethupathy, P., Borel, C., Gagnebin, M., Grant, G. R., Deutsch, S., Elton, T. S., . . . Antonarakis, S. E. (2007). Human microRNA-155 on chromosome 21 differentially interacts with its polymorphic target in the AGTR1 3' untranslated region: a mechanism for functional single-nucleotide polymorphisms related to phenotypes. *Am J Hum Genet, 81*(2), 405-413. doi:10.1086/519979

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res, 29*(1), 308-311.

Singh, R. K., & Cooper, T. A. (2012). Pre-mRNA splicing in disease and therapeutics. *Trends Mol Med, 18*(8), 472-482. doi:10.1016/j.molmed.2012.06.006

Stegeman, S., Amankwah, E., Klein, K., O'Mara, T. A., Kim, D., Lin, H. Y., . . . Batra, J. (2015). A Large-Scale Analysis of Genetic Variants within Putative miRNA Binding Sites in Prostate Cancer. *Cancer Discov, 5*(4), 368-379. doi:10.1158/2159-8290.CD-14-1057

Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., & Kocher, J. A. (2016). Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief Bioinform*. doi:10.1093/bib/bbw069

Tak, Y. G., & Farnham, P. J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin, 8*, 57. doi:10.1186/s13072-015-0050-4

Tang, R., Qi, Q., Wu, R., Zhou, X., Wu, D., Zhou, H., . . . Wang, W. (2015). The polymorphic terminal-loop of pre-miR-1307 binding with MBNL1 contributes to colorectal carcinogenesis via interference with Dicer1 recruitment. *Carcinogenesis, 36*(8), 867-875. doi:10.1093/carcin/bgv066

Thomas, L. F., Saito, T., & Saetrom, P. (2011). Inferring causative variants in microRNA target sites. *Nucleic Acids Res, 39*(16), e109. doi:10.1093/nar/gkr414

Uniken Venema, W. T., Voskuil, M. D., Dijkstra, G., Weersma, R. K., & Festen, E. A. (2017). The genetic background of inflammatory bowel disease: from correlation to causality. *J Pathol, 241*(2), 146-158. doi:10.1002/path.4817

Wang, C., Davila, J. I., Baheti, S., Bhagwate, A. V., Wang, X., Kocher, J. P., . . . Asmann, Y. W. (2014). RVboost: RNA-seq variants prioritization using a boosting method. *Bioinformatics, 30*(23), 3414-3416. doi:10.1093/bioinformatics/btu577

Wang, J., & Batmanov, K. (2015). BayesPI-BAR: a new biophysical model for characterization of regulatory sequence variations. *Nucleic Acids Res, 43*(21), e147. doi:10.1093/nar/gkv733

Wang, K., Li, J., Guo, H., Xu, X., Xiong, G., Guan, X., . . . Bai, Y. (2012). MiR-196a binding-site SNP regulates RAP1A expression contributing to esophageal squamous cell carcinoma risk and metastasis. *Carcinogenesis, 33*(11), 2147-2154. doi:10.1093/carcin/bgs259

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res, 38*(16), e164. doi:10.1093/nar/gkq603

Wang, T., Xiao, G., Chu, Y., Zhang, M. Q., Corey, D. R., & Xie, Y. (2015). Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res, 43*(11), 5263-5274. doi:10.1093/nar/gkv439

Winter, J., Jung, S., Keller, S., Gregory, R. I., & Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol, 11*(3), 228-234. doi:10.1038/ncb0309-228

Wynendaele, J., Bohnke, A., Leucci, E., Nielsen, S. J., Lambertz, I., Hammer, S., . . . Bartel, F. (2010). An illegitimate microRNA target site within the 3' UTR of MDM4 affects ovarian cancer progression and chemosensitivity. *Cancer Res, 70*(23), 9641-9649. doi:10.1158/0008-5472.CAN-10-0527

Xiong, F., Wu, C., Chang, J., Yu, D., Xu, B., Yuan, P., . . . Lin, D. (2011). Genetic variation in an miRNA-1827 binding site in MYCL1 alters susceptibility to small-cell lung cancer. *Cancer Res, 71*(15), 5175-5181. doi:10.1158/0008-5472.CAN-10-4407

Xu, J., Yin, Z., Gao, W., Liu, L., Yin, Y., Liu, P., & Shu, Y. (2013). Genetic variation in a microRNA-502 minding site in SET8 gene confers clinical outcome of non-small cell lung cancer in a Chinese population. *PLoS One, 8*(10), e77024. doi:10.1371/journal.pone.0077024

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc, 10*(10), 1556-1566. doi:10.1038/nprot.2015.105

Yang, L., Li, Y., Cheng, M., Huang, D., Zheng, J., Liu, B., . . . Lu, J. (2012). A functional polymorphism at microRNA-629-binding site in the 3'-untranslated region of NBS1 gene confers an increased risk of lung cancer in Southern and Eastern Chinese population. *Carcinogenesis, 33*(2), 338-347. doi:10.1093/carcin/bgr272

Yousef, G. M. (2015). miRSNP-Based Approach Identifies a miRNA That Regulates Prostate-Specific Antigen in an Allele-Specific Manner. *Cancer Discov, 5*(4), 351-352. doi:10.1158/2159-8290.CD-15-0230

Zhang, J., Yu, H., Zhang, Y., Zhang, X., Zheng, G., Gao, Y., . . . Zhou, L. (2014). A functional TNFAIP2 3'-UTR rs8126 genetic polymorphism contributes to risk of esophageal squamous cell carcinoma. *PLoS One, 9*(11), e109318. doi:10.1371/journal.pone.0109318

Zhang, S., Chen, H., Zhao, X., Cao, J., Tong, J., Lu, J., . . . Lu, D. (2013). REV3L 3'UTR 460 T>C polymorphism in microRNA target sites contributes to lung cancer susceptibility. *Oncogene, 32*(2), 242-250. doi:10.1038/onc.2012.32

Zhang, Y., Li, Y., Hao, Z., Li, X., Bo, P., & Gong, W. (2016). Association of the Serotonin Receptor 3E Gene as a Functional Variant in the MicroRNA-510 Target Site with Diarrhea Predominant Irritable Bowel Syndrome in Chinese Women. *J Neurogastroenterol Motil, 22*(2), 272-281. doi:10.5056/jnm15138

Ziebarth, J. D., Bhattacharya, A., Chen, A., & Cui, Y. (2012). PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic Acids Res, 40*(Database issue), D216-221. doi:10.1093/nar/gkr1026

Zu, Y., Ban, J., Xia, Z., Wang, J., Cai, Y., Ping, W., & Sun, W. (2013). Genetic variation in a miR-335 binding site in BIRC5 alters susceptibility to lung cancer in Chinese Han populations. *Biochem Biophys Res Commun, 430*(2), 529-534. doi:10.1016/j.bbrc.2012.12.001