Viraj Rapolu
BIOL 395

Allelic Imbalance Analysis in Crohn's Disease Patients

Introduction

Crohn's Disease is an Inflammatory Bowel Disease (IBD) that affects up to 3 million Americans (Dahlhamer et al. 167). Crohn's Disease is heterogeneous, so it is expressed by a multitude of phenotypes. Due to the variety in phenotypes, progression and response to treatment is highly variable. We believe the solution to understanding the disease lies in the molecular characteristics of the colonic and intestinal tissue as opposed to clinical phenotypes. By understanding the molecular basis behind the disease, scientists can better determine progression, treatment response, and more.

After performing Genome Wide Association Studies (GWAS), 242 loci associated with IBD have been found (Mirkov et al. 224). Individual variants in those loci were found largely in noncoding regions. This presents a unique challenge, as it is difficult to determine how each variant affects gene expression. Variants in coding regions can directly affect gene expression by changing the amino acid sequence, whereas variants in noncoding regions can indirectly affect gene expression by influencing expression levels through a variety of methods.

This analysis is looking at allelic imbalance, which is a phenomenon where two different alleles in the genome are expressed at different levels (Wagner et al). This is determined by sequencing a functional assay to generate a readout. We can quantify how much of the signal from the assay comes from one allele compared to the other. If there is a signaling imbalance, we can conclude that there is allelic imbalance.

From an IBD sample, if a site has allelic imbalance, a natural hypothesis is that somehow the variant is affecting the level of gene expression. A heterozygous single nucleotide polymorphism (SNP) that does not alter function at that site should have an equal ratio of the two alleles, whereas a SNP that alters function at the site will not.

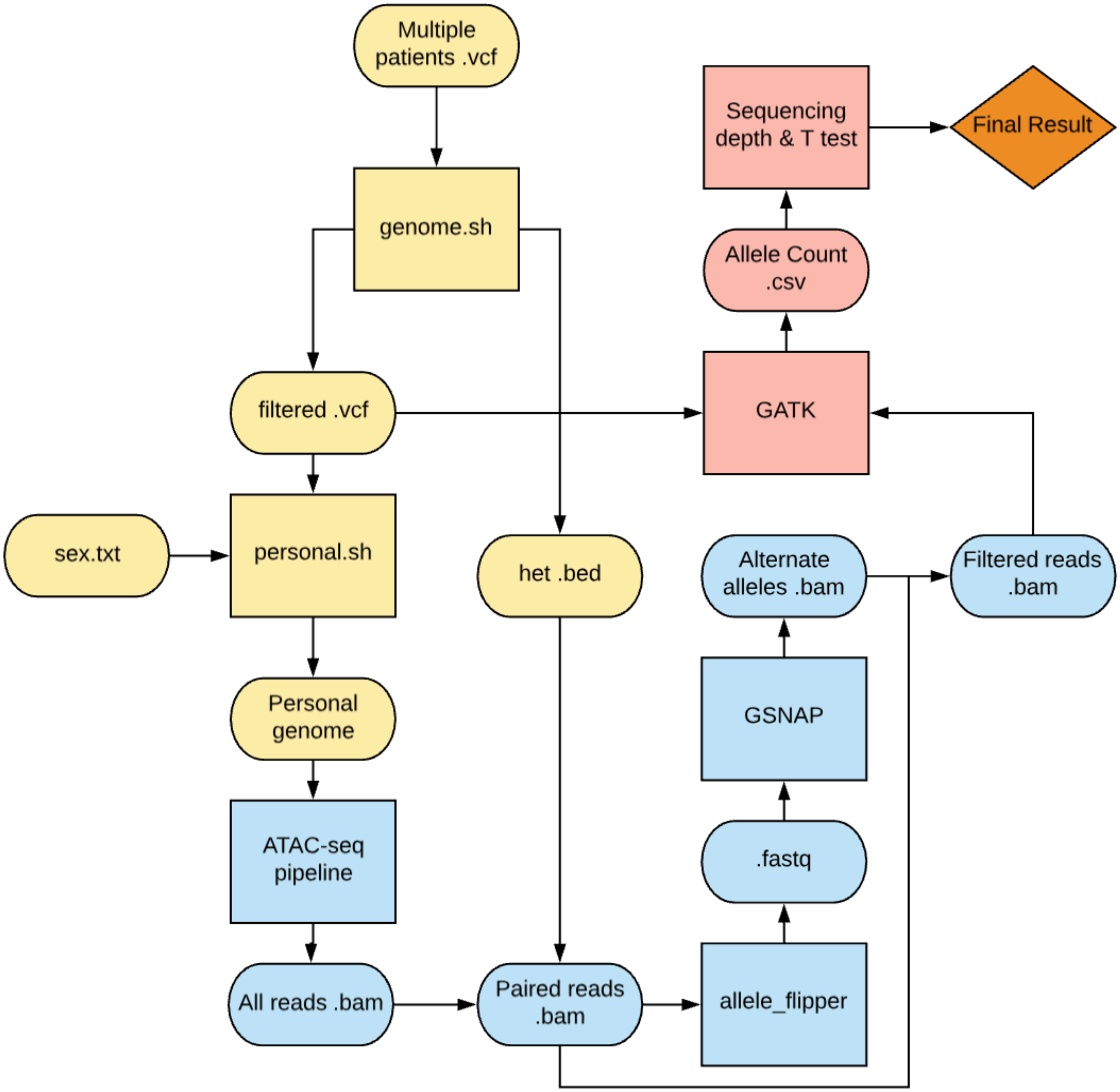
To gather information about noncoding regions, we use the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) method, which maps chromatin accessibility genome-wide. ATAC-seq data identifies regions of open chromatin that are regulatory elements, and in doing so generates sequencing reads at these regions of open chromatin (Buenrostro et al). If a variant changes the chromatin accessibility, we are inferring that it changes the activity of that regulatory element when the allele is present. For all allelic imbalance sites, it is telling us how variation may be affecting the overall gene regulatory program.

The purpose of this project was to develop a robust, accurate computational pipeline that reduces any potential mapping bias in allelic imbalance analysis using ATAC-seq data in non-coding regions. Mapping bias can arise when aligning a sequencing read to a genome. If the read or genome is imprecise, mapping can occur at a different location (Liu et al). This pipeline works to create an unbiased, personal genome, in addition to filtering for reads at heterozygous variant sites represented by the reference and alternate alleles. Data was later applied to this pipeline to determine allelic imbalance in Crohn's Disease patients.

Viraj Rapolu
BIOL 395

Methods

Figure 1. Overview of the allelic imbalance pipeline



Colors represent steps used to create key intermediate files and the final results, which are described in more detail below. Steps with rounded edges are files, but sharp edges are processes.

Viraj Rapolu
BIOL 395

Create Personal Genome

Typically, reference genomes are used to align reads. A reference genome is an assembly derived from the DNA from numerous individuals. A personal genome is a reference genome for a specific individual that has their variants at all heterozygous sites. The primary step for the entire pipeline is to generate a personal genome. From a .vcf file that contains genotype data from numerous patients, an individual .vcf file must be extracted. We ran the genome.sh script, which had an output of a .vcf file, homozygous and heterozygous .bed files, and also created the directory structure for the personal genome.

After obtaining the genotype information for the individual, the personal genome must be synthesized. We used the personal.sh script, which uses GSNAP, an alignment program. This script required a table of patient information, which we labeled sex.txt. This table simply included a patient identification number and gender on the same line, separated by a tab. Another input for personal.sh is the .vcf file obtained from the previous step. In addition, personal.sh uses the directory structure created from the genome.sh script. The output from personal.sh is a set of .fasta files for each chromosome in addition to index files to be used by GSNAP for aligning reads. The combination of these .fasta files and .bed files containing homozygous and heterozygous sites with their respective alleles makes up the personal genome.

Generating & Filtering Read Data

The Furey Lab ATAC-seq pipeline is used to align reads to .fasta files from the personal genome. This pipeline assigns position values for every read that successfully aligns to a unique location on the personal genome. These reads and their position values are stored in an indexed .bam file.

After generating this .bam file, reads in it must be filtered to contain reads that overlap at heterozygous SNPs. First, this .bam file is cross-referenced with the heterozygous .bed file generated during the first step of the personal genome to create a paired overlap .bam file that only contains reads at heterozygous positions. This paired overlap .bam file is ran through an allele_flipper script which references the reads to the heterozygous .bed file, switches the read to contain the alternate allele, and writes the updated read to outputs of two .fastq files. These .fastq files are aligned using GSNAP to create a new .bam file that contains alignments for reads with alternate alleles. This new .bam file is compared to the .bam file that contains all heterozygous reads. Reads that align to the same position in both of these files are stored in a third "filtered reads" .bam file.

Determining Allelic Imbalance Levels

To determine the levels of allelic imbalance, we used the ASEReadCounter tool, a program developed by Genome Analysis Toolkit (GATK). The inputs for the ASEReadCounter were the .bam file containing filtered reads from the previous step and .index, .fasta, and .dict files that were generated in the personal genome. The output was a .csv file with reference and alternate allele counts at their respective positions.

Statistical analysis was performed on this data. Heterozygous sites that did not have at least 5 reads for each of its two alleles were not considered. Then, we performed a binomial distribution on the remaining data and selected for sites with a significance value of $p = 0.05$. Sites that met this threshold were less likely to display allelic imbalance by chance.

Viraj Rapolu
BIOL 395

Results

Mitigating Mapping Bias

Mapping bias creates an inaccurate alignment of reads to a genome. Since reads used in this analysis are no longer than 50 base pairs, there can be more than one location they can be aligned to, despite our alignment algorithms allowing for only a single mismatch of nucleotides. In the scope of this study, an inaccurate alignment may result in a heterozygous site being associated with allelic imbalance, and can result in future resources being wasted to investigate the variants. However, this pipeline mitigates mapping bias with two methods: creating a personal genome as opposed to a reference genome for reads to be aligned to, and generating reads with reference and alternate alleles at heterozygous sites.

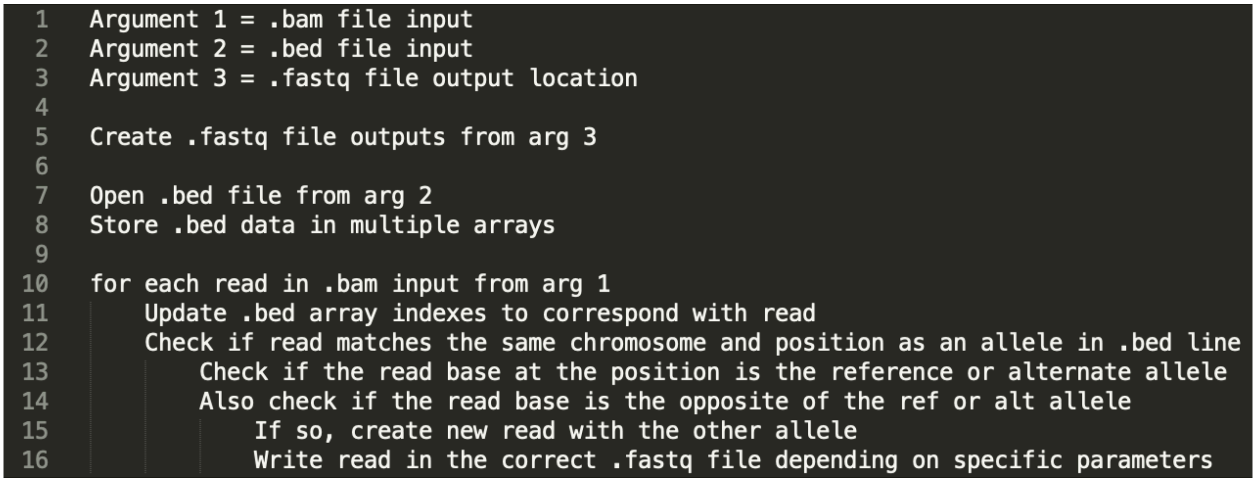
The personal genome contains all read data from the individual sampled. For example, an individual may contain both alleles at a heterozygous site, one from each parent. The reference genome may be homozygous at that site, and only contain one of the alleles. In this case, identical reads would align differently to the personal genome than they would to the reference genome. In some cases, reads may not align at all to the reference genome at that site, due to the specificity of the alignment algorithm.

Another way the pipeline decreases mapping bias is through the read generation process. As opposed to using all reads, the first step in the process is to filter for heterozygous reads, as allelic imbalance occurs at heterozygous sites. Furthermore, we looked at reads that have both the reference allele and alternate allele. If only one set of these reads was considered, it may align to a different location. Reads that do not align to the same position with the reference allele present and alternate allele present are filtered out. This creates another step that decreases the likelihood that a read without allelic imbalance is presented in the final dataset.

Viraj Rapolu
BIOL 395

Allele Flipper

A key component on forming this analysis is to ensure that in the alignment of reads to the custom genome, no bias has been introduced that would unfairly favor reads being aligned from one allele present to another. In order to do this, a large part of the project was centered on a test to ensure that a read at a heterozygous site would be represented both with its reference and alternate alleles. This way, each variant has a possibility of aligning to the personal genome. The general premise of the script was to match an allele from the .bed file to a read in the .bam file, ensuring that both were located on the same chromosome and the targeted allele's position matched the base position in the read. Below is pseudocode for the allele_flipper script written to perform this task.



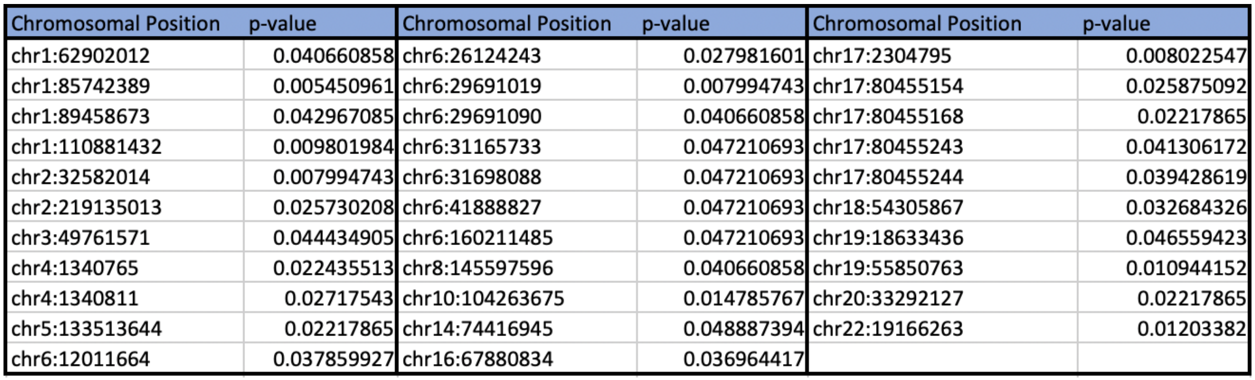
```
Argument 1 = .bam file input
Argument 2 = .bed file input
Argument 3 = .fastq file output location

Create .fastq file outputs from arg 3

Open .bed file from arg 2
Store .bed data in multiple arrays

for each read in .bam input from arg 1
    Update .bed array indexes to correspond with read
    Check if read matches the same chromosome and position as an allele in .bed line
        Check if the read base at the position is the reference or alternate allele
        Also check if the read base is the opposite of the ref or alt allele
            If so, create new read with the other allele
            Write read in the correct .fastq file depending on specific parameters
```

A critical component in the design of this script was efficiency. Each input file contains approximately 1 million lines of data. In the .bam file, each line contains information regarding the chromosome, position, and reference/alternate allele. In the .sam file, each line contains a quality score, read, sequence ID, chromosome, and position. At first, I was comparing each line of .bed data to each line of .bam data to determine if the positions and chromosomes aligned. While this worked, the implementation was too slow for the scope of data, and needed to be implemented in a faster, more memory-efficient method. Instead, I implemented lines 7 and 8 to create arrays that hold the .bed data. In lines 12-14, I compare data in the read to these array data, which is much faster due to arrays being the one of the quickest data structures to traverse.

Another area where efficiency is implemented is at line 11. Previously, the script was scanning through every line in the .bed file for each read, meaning that each read had to filter through approximately 1 million lines of data. Once again, the script was running too slowly, so changes had to be made. Since the data is ordered by position in the .bed file, the line of the last allele used was held, and the script simply scanned from that line until the end of the chromosome as opposed to the entire file. This dramatically increased the efficiency of the script, and made it viable to work in the pipeline on large sets of data.

Data from CD and non-IBD samples

Figure 2. Table of sites with allelic imbalance in a Crohn's Disease patient

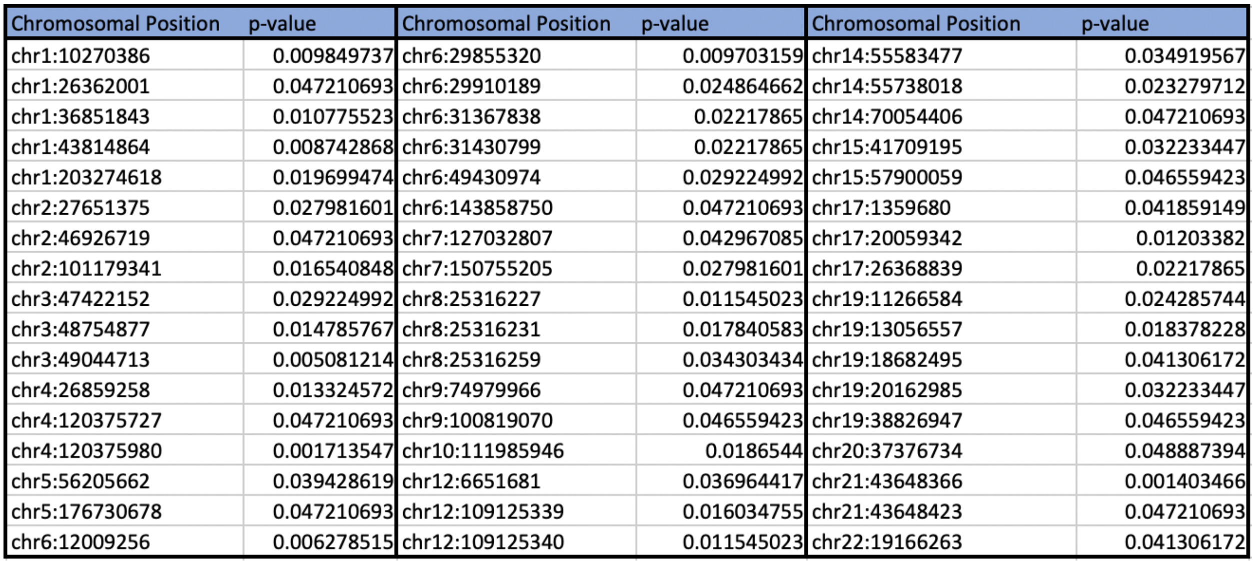


Chromosomal Position	p-value	Chromosomal Position	p-value	Chromosomal Position	p-value
chr1:62902012	0.040660858	chr6:26124243	0.027981601	chr17:2304795	0.008022547
chr1:85742389	0.005450961	chr6:29691019	0.007994743	chr17:80455154	0.025875092
chr1:89458673	0.042967085	chr6:29691090	0.040660858	chr17:80455168	0.02217865
chr1:110881432	0.009801984	chr6:31165733	0.047210693	chr17:80455243	0.041306172
chr2:32582014	0.007994743	chr6:31698088	0.047210693	chr17:80455244	0.039428619
chr2:219135013	0.025730208	chr6:41888827	0.047210693	chr18:54305867	0.032684326
chr3:49761571	0.044434905	chr6:160211485	0.047210693	chr19:18633436	0.046559423
chr4:1340765	0.022435513	chr8:145597596	0.040660858	chr19:55850763	0.010944152
chr4:1340811	0.02717543	chr10:104263675	0.014785767	chr20:33292127	0.02217865
chr5:133513644	0.02217865	chr14:74416945	0.048887394	chr22:19166263	0.01203382
chr6:12011664	0.037859927	chr16:67880834	0.036964417		

Positions of 32 sites with evidence of allelic imbalance. Sites were filtered by a binomial distribution using a p value < 0.05

Viraj Rapolu
BIOL 395

Figure 3. Table of sites with allelic imbalance in a non-IBD patient



Chromosomal Position	p-value	Chromosomal Position	p-value	Chromosomal Position	p-value
chr1:10270386	0.009849737	chr6:29855320	0.009703159	chr14:55583477	0.034919567
chr1:26362001	0.047210693	chr6:29910189	0.024864662	chr14:55738018	0.023279712
chr1:36851843	0.010775523	chr6:31367838	0.02217865	chr14:70054406	0.047210693
chr1:43814864	0.008742868	chr6:31430799	0.02217865	chr15:41709195	0.032233447
chr1:203274618	0.019699474	chr6:49430974	0.029224992	chr15:57900059	0.046559423
chr2:27651375	0.027981601	chr6:143858750	0.047210693	chr17:1359680	0.041859149
chr2:46926719	0.047210693	chr7:127032807	0.042967085	chr17:20059342	0.01203382
chr2:101179341	0.016540848	chr7:150755205	0.027981601	chr17:26368839	0.02217865
chr3:47422152	0.029224992	chr8:25316227	0.011545023	chr19:11266584	0.024285744
chr3:48754877	0.014785767	chr8:25316231	0.017840583	chr19:13056557	0.018378228
chr3:49044713	0.005081214	chr8:25316259	0.034303434	chr19:18682495	0.041306172
chr4:26859258	0.013324572	chr9:74979966	0.047210693	chr19:20162985	0.032233447
chr4:120375727	0.047210693	chr9:100819070	0.046559423	chr19:38826947	0.046559423
chr4:120375980	0.001713547	chr10:111985946	0.0186544	chr20:37376734	0.048887394
chr5:56205662	0.039428619	chr12:6651681	0.036964417	chr21:43648366	0.001403466
chr5:176730678	0.047210693	chr12:109125339	0.016034755	chr21:43648423	0.047210693
chr6:12009256	0.006278515	chr12:109125340	0.011545023	chr22:19166263	0.041306172

Positions of 51 sites with evidence of allelic imbalance. Sites were filtered by a binomial distribution using a p value < 0.05

Samples were taken from a patient with Crohn's Disease and a patient with non-IBD. Because the sample size for each category is one, these data will not give us much insight into discovering sites with allelic imbalance that play an important role in Crohn's Disease. However, after collecting data from more Crohn's Disease patients, we can identify sites with allelic imbalance that appear in numerous patients, and investigate those to learn more about the molecular characteristics of the disease. The non-IBD data gives us a baseline to compare Crohn's Disease sites to, so that sites prevalent in both CD patients and non-IBD patients are not investigated.

Discussion

This pipeline will allow us to conduct allelic imbalance analysis on more patients. Currently, we only have the final result for a few patients, but once more data is available, we will put it through the pipeline it to further analyze sites with allelic imbalance. Running a larger dataset will give us more information on which loci are have higher levels of allelic imbalance. By collecting data on which loci have significant levels of allelic imbalance, we learn more about the molecular characteristics of the disease, which could lead to long-term clinical progress.

This pipeline is currently designed for ATAC-seq data. ATAC-seq data is taken from regions of open chromatin, not transcribed regions. These regions are upstream of transcribed genes, and changes here may affect gene expression levels. Due to this, information gained will correspond to gene regulatory information. Another direction this pipeline could be applied to is RNA-seq data. Since RNA-seq looks at transcribed regions of DNA, changes at these sections will correspond to changes in genes themselves. The only change in this pipeline that would be required to work with RNA-seq data is in generating and filtering read data. Instead of using the ATAC-seq pipeline, we would use the RNA-seq pipeline that has already been developed by the Furey Lab. Since the file formats will remain constant throughout the rest of this allelic imbalance analysis pipeline, nothing else needs to be changed.

In the future, I would like to tackle a similar problem, but from another angle. As opposed to focusing on regulatory regions that affect levels of gene expression, I would like to look into protein or mRNA counts to see levels of gene expression. By seeing which genes are at higher or lower expression levels in Crohn’s Disease patients relative to colon tissue in non-IBD patients, we can then look at the allelic imbalance data to see if the loci correspond to upstream regions of the genes with altered expression. The end goal will still be the same, as this may help better understand Crohn’s Disease from a molecular perspective.

Viraj Rapolu
BIOL 395

References

Buenrostro, Jason D., et al. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” Current Protocols in Molecular Biology, 5 Jan. 2015, doi:10.1002/0471142727.mb2129s109.

Dahlhamer, James M., et al. “Prevalence of Inflammatory Bowel Disease Among Adults Aged ≥18 Years — United States, 2015.” MMWR. Morbidity and Mortality Weekly Report, vol. 65, no. 42, 28 Oct. 2016, pp. 1166–1169., doi:10.15585/mmwr.mm6542a3.

Liu, Xinan, et al. “IMapSplice: Alleviating Reference Bias through Personalized RNA-Seq Alignment.” Plos One, vol. 13, no. 8, 10 Aug. 2018, doi:10.1371/journal.pone.0201554.

Mirkov, Maša Umićević, et al. “Genetics of Inflammatory Bowel Disease: beyond NOD2.” The Lancet Gastroenterology & Hepatology, vol. 2, no. 3, 2 Mar. 2017, pp. 224–234., doi:10.1016/s2468-1253(16)30111-x.

Wagner, James R., et al. “Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human.” PLoS Computational Biology, vol. 6, no. 7, 8 Jul. 2010, doi:10.1371/journal.pcbi.1000849.