Abstract

The exact cause of Crohn's Disease (CD), one of the main types of inflammatory bowel disease (IBD), remains an open question. Previous Genome-Wide Association Studies (GWAS) have associated Crohn's disease with >240 genetic loci, but the mechanism behind these associations is still unknown. Many of the identified sites lie within the non-coding region of human genome, suggesting gene regulatory element as one of the contributors to Crohn's Disease. From this observation, we posed the question that motivated this research: How does genetic variation influence chromatin accessibility that may contribute to developing Crohn's disease? Allelic imbalance is a phenomenon where the two alleles of a given gene are expressed at different levels in a given cell, either because of epigenetic inactivation of one of the two alleles, or because of genetic variation in regulatory regions (Wagner, Pokholok, Gunderson, Pastinen, & Blanchette, 2010). We developed a computational pipeline to perform allelic imbalance analysis on both Crohn's disease patients and non-IBD control individuals to discover the correlation between chromatin accessibility with different copies of SNP alleles. We performed the analysis based on patients' genotype information and genome-wide, sequencing-based chromatin accessibility assay results. 239 sites of allelic imbalance were detected in chromatin accessibility data from the CD group, and 273 imbalance sites from the non-IBD group. By comparing the identified sites of imbalance with GWAS results and reviewing the function of genes adjacent to these imbalanced sites, we identified several sites potentially contributing to Crohn's disease. These identified sites/regions provide interesting targets for future studies to determine the cause of Crohn's disease.

Introduction

Although the exact causation of Crohn's disease is still unknown, strong genetic factors are suggested by the occurrence of this disease in families. Trying to understand how genetic factors contribute to Crohn's disease, GWAS studies have associated it with various genetic sites. Many of these sites lie within the non-coding region of human genome (Verstockt, Smith, & Lee, 2018). Since these regions does not participate in protein production, they are more likely to be contributing to the disease by regulating the expression level of other genes. From this observation, we posed the question that motivated this research: Is chromatin accessibility in Crohn's disease patients influenced by different alleles in single nucleotide polymorphisms (SNPs)?

Theoretically, if a SNP is heterozygous at a site where both alleles have the same effect on chromatin accessibility, the chromatin is assumed to have equal accessibility at that loci for both alleles. To test for chromatin accessibility, the Assay for Transposase-Accessible Chromatin(ATAC-seq) was used. It functions by inserting transposase Tn5 into the genome. Regions free of nucleosomes can efficiently add Tn5 and can be isolated and amplified by PCR more easily. DNA where Tn5 was added can be sequenced. Therefore, when mapping sequencing reads to a reference genome, regions with higher chromatin accessibility are expected to have more reads aligned to it. Thus, ATAC-seq reflects chromatin accessibility at a certain location by its corresponding read enrichment levels. If the loci with different alleles on the two chromosomes have similar chromatin accessibility, both have equal chance to be accessed by the Tn5 transposes during ATAC-seq. When the sequence reads were aligned, we would expect seeing roughly equal number of reads aligned to each of the alleles at this heterozygous site. On the contrary, if there is a significant imbalance between read numbers at

the two alleles, it suggests that the allele present at this SNP site affects the chromatin accessibility of the region.

However, allelic bias introduced during sequence alignment can be a major problem when looking at allele specific expression levels (Stevenson, Coolon, & Wittkopp, 2013). When aligning sequencing results, the standard human reference genome such as hg19/hg38 is usually used. However, the standard reference genome only contains one variation for each SNP site. This would punish reads containing the alternate alleles for their lower similarity, and they would become less likely to align to their corresponding position. By favoring one allele over the other, reads with the allele contained in the reference genome can be more easily aligned and cause false positives when looking for allelic imbalance. Another possible bias source is the overlapping ends for paired reads. alleles in the overlapping region may get counted twice when measuring for chromatin accessibility. For this study, we established an allelic imbalance detection pipeline that can reduce these biases and decrease the false positive rate during detection.

Furthermore, by duplicating the experiment process on multiple subjects, we explored whether we can discover allelic imbalance at sites specific to and prevalent among Crohn's disease patient population. Such observations would support the hypothesis that these sites are associated with Crohn's disease. More evidence is needed and there are several aspects that we wish to look at in future studies --- such as cross-referencing with GWAS results to confirm known associations with functions. Also, whether these SNPs influence the expression of target genes of accessible chromatin can be tested using experiments such as RNA-seq. If these tests can detect

expression peaks coherent with our result, the connection between these sites and the Crohn's disease can be further established.

Method

Data collection

Well-characterized CD patients from the adult IBD Center at University of North Carolina were included in this study (IRB Approval # 10-0355, 14-2445 and 11-0359). (Weiser et al., 2016) Colon tissue was obtained from each sample. Tissue samples were confirmed by an independent pathologist to have no active inflammation, only quiescent colitis. Tissue from non-IBD control patients was obtained at time of surgical resection for non-IBD related illness and from a site distant from any pathology (Weiser et al., 2016). 12 CD samples and 18 control samples were submitted for Assay for Transposase Accessible Chromatin (ATAC) sequencing analyses. In addition, all samples were genotyped using the Illumina Immunochip platform, and imputation was carried out with the University of Michigan imputation server. (Das et al., 2016)

Gender	CD	non-IBD
Female	8	11
Male	4	7
	12	18

Table1. Group size and gender information for the Crohn's Disease patient group (CD) and the non-IBD control group (non-IBD).

Custom reference genome creation

The personalized reference genome was created for each sample using the standard hg19 genome. Based on the imputed genotype information, homozygous alternate alleles were directly replaced in the hg19 fasta files. (Buchkovich, Eklund, Duan, Li, Mohlke, & Furey, 2015) 'Snpindex' was used to create index files containing heterozygous positions to tolerate these SNPs during alignment process (Wu, & Watanabe, 2015).

ATAC-seq and analysis pipeline (FureyWiki, 2019)

ATAC-seq was performed according to previously described protocol (Buenrostro, Wu, Chang, & Greenleaf, 2015). Sequence reads were processed by 'Cutadapt' (Martin, 2011) to remove adapter sequence. Low quality ends were trimmed before adapter removal. Five matching bases between the read and adapter were needed for the adapter removal to be performed. Reads with length shorter than 30bps after removal were discarded. FASTX_Toolkit was used to trim reads into length of 50 bases, and only keeping reads that have quality higher than 20 for at least 90% of its bases. Non-mappable sequences such as linker, barcode, fingerprints are filtered and discarded using tagdust. The remaining sequence reads were aligned (Lassmann, Hayashizaki, & Daub, 2009).

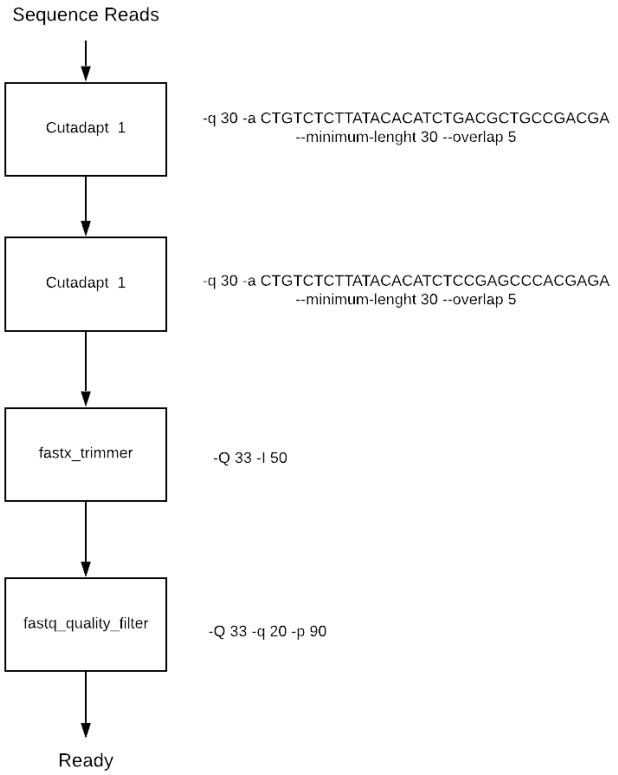

*Figure1. ATAC-seq analysis pipeline and parameters used for processing (FureyWiki, 2019**).** Including parameters and the adapter sequence used for each step.*

Alignment

GSNAP was used to align the 50 bp paired-end reads to the customized reference genome (Wu, & Watanabe, 2015). Alignment results were referenced with the previously generated list of heterozygous SNPs in each individual. Reads containing listed positions would be substituted with the alternate allele at these loci. These converted reads were transformed into fastq files and re-aligned to the customized reference genome using the same parameters as the original reads. Reads aligned to different positions in the second alignment were discarded. Thus, only reads aligned uniquely regardless of the allele present were used to detect allelic imbalance (Buchkovich, Eklund, Duan, Li, Mohlke, & Furey, 2015).

Allelic Imbalance Analyses

The filtered alignment results were processed with the GATK toolkit (McKenna et al., 2010) using the ASEReadCounter function, with the -U ALLOW_N_CIGAR_READS flag used and a list of heterozygous positions provided as reference. This function calculates allele counts at a set of positions after applying filters that are tuned for enabling allele-specific expression (ASE) analysis. The function's output file contained allele read counts at all heterozygous SNP positions.

Imbalance significance was assessed using binomial probability with the p-value threshold set to 0.05 (Buchkovich, Eklund, Duan, Li, Mohlke, & Furey, 2015). To ensure statistical power, a minimum threshold for each allele was set at 5. This makes the smaller count between the two alleles of each SNP be at least 5 for the loci to be considered for testing as a significantly imbalanced site. With a p-value less than 0.05, it is unlikely that the allelic imbalance was caused by chance. However, it has to be noticed that significance in one subject cannot imply significance among the population.

Results

Establishment of allelic imbalance detection pipeline

During this study, a next generation sequencing data based pipeline was developed to detect allelic imbalance sites. The pipeline takes each sample's ATAC-seq reads and genotype

information as input, and produces the detected imbalance loci and their corresponding p-value as the output.

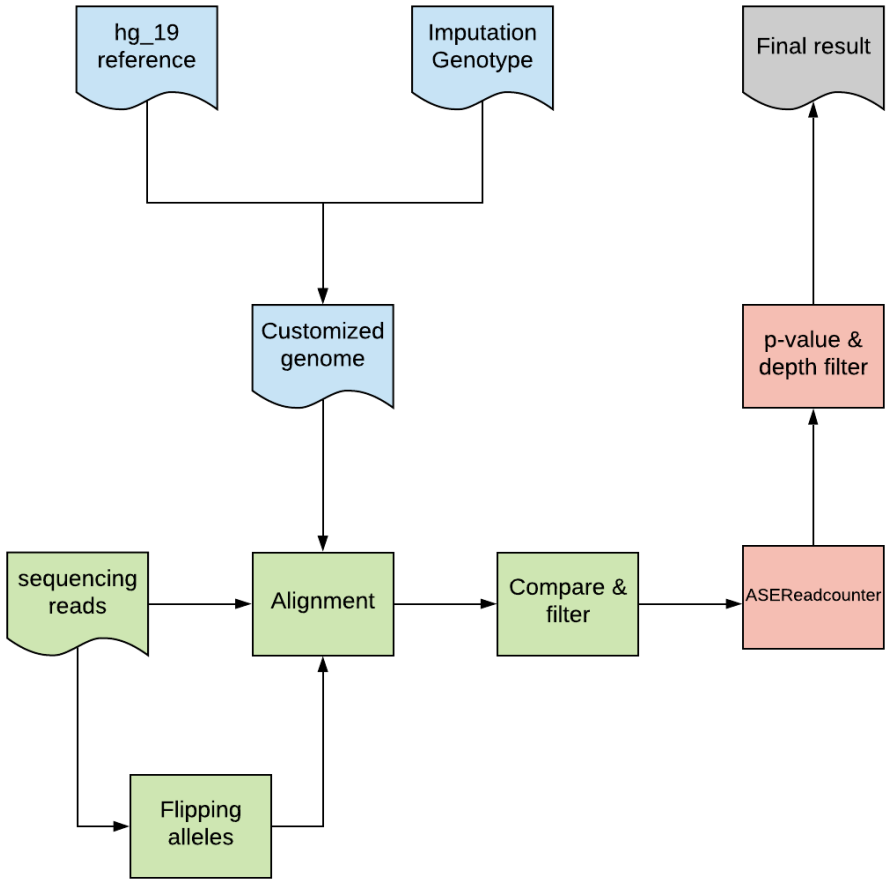

Figure 2. Allelic imbalance detection pipeline process flowchart. The pipeline consists of three main parts: customized reference genome creation (blue), allele-aware alignment (green), and allelic imbalance detection (red).

The goal for creating a customized reference genome was to eliminate the allelic bias introduced during the alignment. By referencing the standard genome, reads containing alleles not included in the reference genome will be penalized due to mismatch and reduced similarity. This bias is important for the allelic specific expression level study because it can introduce false positive, favoring reads containing same alleles with the reference genome. By creating the customized

reference genome, both copy of alleles can receive equal consideration during alignment and therefore eliminate this source of bias.

Another source of possible bias is the difference in mappability during alignments based on which allele is present (Buchkovich, Eklund, Duan, Li, Mohlke, & Furey, 2015). The allele-aware alignment step was performed in order to eliminate this bias. For each read that contained an heterozygous position during the first alignment, we flipped the allele so that it contained to the alternate allele, and performed the alignment again with same reference genome and parameters. Comparing this alignment result to the original alignment, we filtered out reads that have a changed alignment result. These reads' alignments were influenced by the allele they carry. To exclude the alignment bias, we only kept reads that uniquely aligned to the same position regardless of the allele contained. Also, we filtered out reads aligned to multiple locations. By doing this, it is possible that we lowered the sensitivity of the detection by filtering out reads. But we were trying to ensure that we eliminated as much false positive results as possible.

To account for the possible bias introduced by the overlapping ends of paired reads, we used the ASEReadcounter from the GATK toolkit (McKenna et al., 2010). Its -U ALLOW_N_CIGAR_READS flag automatically removes one copy of the overlapping ends during allele counting and therefore prevents counting alleles in these regions twice and which can cause false positive signals. A binomial distribution model was used for the significance test. We chose the standard p-value threshold of 0.05. On average only 0.004% of reads were identified as significant.

Allelic imbalance detected in Crohn's Disease

The previously described allelic imbalance detection pipeline was used to detect imbalance sites potentially related to Crohn's Disease. Data collected from two groups of sample—Crohn's disease patients and non-IBD controls were ran through the pipeline. 239 allelic imbalance sites were detected from the CD group, and 273 imbalance sites from the non-IBD group (Appendix A, B). 54 sites were identified in both groups. Among these commonly identified loci, 22 loci show different imbalance tendency in two groups (i.e. one copy of allele has higher expression level in the non-IBD group but has lower expression level in the CD group compared to the alternate copy) (Table 2).

Position	CD Group	non-IBD Group
10:52384485	9:17	21:12
1:12040324	14:23	19:9
11:64085148	11:23	15:7
11:95523030	5:12	19:10
12:58087737	5:17	17:5
13:101327245	39:19	18:27
14:39901596	6:17	14:6
15:75230502	6:14	18:8
17:18266975	11:25	12:5
19:1490888	16:8	7:15
19:56165921	5:12	17:9
20:31407661	5:14	18:7
2:224822043	14:25	17:7
22:50964153	18:6	7:15
3:119813282	17:26	23:8
3:195809080	5:14	18:9
3:81810725	13:23	12:5
7:123174842	6:22	19:9
7:27779729	10:19	14:5
8:104427359	21:9	10:22
8:11660168	27:13	5:17
8:145026450	11:22	15:6
9:100818720	13:5	9:17

Table 2. Common imbalanced sites with different tendency in CD group and non-IBD group. Numbers shown are the reference allele counts versus alternate allele counts of that loci for each group.

To further interpret the identified positions, we looked at genes close to them. Since genes close to the SNP sites are more likely to be regulated, their function partially reflects what possible change in phenotype these alleles, by regulating gene expression, may cause. Because Crohn's disease involves chronic inflammation in the colon tissue, we especially looked for gene functions that are connected with immune system pathways.

There are several interesting genes being detected. The GSK3B gene near position chr3:119813282 is involved in energy metabolism and inflammation. The TAX1BP1 gene near chr 7:27779729 have a role in the inhibition of inflammatory signaling pathways. The PLEC gene near chr8:145026450 is related with immune system. Although these gene ontology cannot be used as direct evidence for connection with the disease, the related sites should be considered to have a higher possibility to contribute.

One of the detected positions especially attracted our attention. The alternate allele on chr2:219151926 was detected to have a higher expression level. This loci falls within the intron region of gene TMBIM1, which is related to innate immune system pathways. Also, this position has been associated with Crohn's Disease by previous GWAS studies. Suggesting a higher probability of contributing to Crohn's Disease.

Discussion

The allelic imbalance detection on sequencing data can be used as a tool to identify effects of genetic variation on gene regulation, and possibly detect connection between non-coding regions

with disease phenotypes (FureyWiki, 2019). In this study, we used it on Crohn's disease patients, trying to identify gene regulation sites that may contribute to the disease.

Among the significant imbalance sites we detected, we found chr2:219151926 especially interesting. Its correlation with Crohn's Disease was corroborated by GWAS results as well as gene ontology. This SNP position was identified to be an eQTL loci in colon tissue for gene TMBIM1. TMBIM1 negatively regulates MPP9, which is a gene commonly being up-regulated in IBD patients. According to our discovery and these previous studies, we hypothesized that a higher accessibility in one of the copies means a regulatory factor, likely an enhancer for TMBIM1, have less binding to this position due to the presence of the alternate allele. Therefore, it is possible that TMBIM1 gets suppressed and MPP9 gets up-regulated. However, because this imbalance was only detected in one sample from the CD group, we have no strong evidence to directly prove its association with Crohn's Disease.

There are several other detected sites close to genes relevant to the immunization pathway. Thus, if these genes are regulated by the SNPs, the regulation of their expression may send abnormal signals to the immunization system, causing it to respond as if there is an infection. Or, the regulation might cause an overexpression in immunization behavior in the form of constant inflammation. Those possibilities provide a potential link between these sites and the disease phenotype.

Again, with all the imbalance sites only identified in one samples, we could not verify the hypothesis that those sites are connected with the Crohn's disease. The study can be improved by

expanding group size to avoid bias and by combining reads from different subjects for more statistical power. Given these improvements, a future study will provide a better evidence on whether the allelic imbalance sites may contribute to the Crohn's disease by regulating gene expression or not.

Possible future improvements

One of the most important limiting factors to the established pipeline is the sequencing depth for each subject. Sequencing depth influences the statistical power of our imbalance detection, so that by increasing the sequencing depth, more statistical power can be granted to our test, providing more reliable results. One way of improving statistical power for our test is, instead of using individual reads, combine reads from multiple subjects for imbalance detection. With reads from multiple subjects, one thing that we can avoid is subject bias, such as imbalanced sites that are only significant in specific subject due to conditions other than Crohn's disease. These subject specific imbalance reads are likely to be diluted by balanced reads from other subjects and therefore lose significance, leaving the imbalance sites that are prevalent among population. Finally, since more reads are present, larger differences between allelic counts is needed for the imbalance to be significant.

If, with future studies, we can determine whether these imbalanced loci in non-coding regions contribute to Crohn's disease by regulating nearby genes' expression, we can gain a better understanding of the cause of Crohn's Disease. Also, further research on this subject can help

with the unclassified IBD problem by classifying patients according to genetic resemblance in addition to phenotype information.

Acknowledgement

Great thanks to professor Terry Furey for his instructions during the course of this research and comments that greatly improved the manuscript. I'm also grateful for the help provided by the members of the Furey Lab, Ben Keith, Viraj Rapolu, and Nur Shahir. I would also like to extend my thanks to the IBD center of UNC hospital for providing samples that supported this study.

Appendix

A. Imbalanced sites detected in CD samples

1:110881432	3:180630191	6:31697558	10:3215111	16:11349604	19:11071560
1:115300685	3:183354524	6:32806584	10:52384485	16:1359217	19:1490888
1:12040324	3:183735286	6:41888827	10:5530385	16:2014935	19:17337223
1:151966348	3:195809080	6:52285014	10:65280994	16:2033941	19:18263871
1:169863165	3:195809116	6:75994572	10:75545063	16:2033970	19:18392894
1:183604905	3:47021124	7:112090360	10:94050988	16:2034137	19:18530161
1:207494679	3:49761613	7:12251262	11:105893240	16:23193934	19:18530172
1:207999200	3:81810725	7:123174842	11:118992177	16:3451030	19:18747696
1:209979613	4:103748797	7:149321803	11:125462616	16:50775745	19:3435254
1:22109796	4:108641464	7:158649280	11:125462625	16:72042442	19:38806772
1:226271290	4:119199911	7:23145289	11:125462634	16:77224656	19:38826947
1:26758748	4:130014703	7:2393947	11:13484873	16:77224827	19:41284398
1:40506239	4:1340939	7:27779729	11:535649	16:87425683	19:4402474
1:43855578	4:140222537	7:2884116	11:64085148	16:89724268	19:50380858
1:85742389	4:140375145	7:44646185	11:64546391	16:89768558	19:54641388
1:93646207	4:185655402	7:76022497	11:86749027	16:89768560	19:55770488
2:122288656	4:26859258	7:76178517	11:95523030	17:12921450	19:55850763
2:128284073	4:3076181	8:104427359	12:110486272	17:18128822	19:56165194
2:135675908	4:57843863	8:11660168	12:49110835	17:18266975	19:56165475
2:173420746	4:85503996	8:145026450	12:51441897	17:27056025	19:56165921
2:174830206	5:108084178	8:146012260	12:57119236	17:27056026	19:7894732
2:175113512	5:122847678	8:38854041	12:58087737	17:34900836	20:30192948
2:176033215	5:131831942	8:49427167	13:101327245	17:34900847	20:31407661
2:190526205	5:133513644	8:74791162	13:101327345	17:43212963	20:33292127
2:190648805	5:153570088	8:74884530	13:27998719	17:5323125	20:34542365
2:219134950	5:78280597	8:81083860	13:27998769	17:5390128	21:30365322
2:219135013	5:79950781	8:8860276	13:98085349	17:54911403	21:35014316
2:219151926	6:109776903	8:9413307	14:35452126	17:56429764	21:37528844
2:224822043	6:16129072	9:100818720	14:39901596	17:57184162	21:43648423
2:241500508	6:170102276	9:110045257	14:71374702	17:57696773	21:43648430
2:27651375	6:26026599	9:111882365	14:74416945	17:61851435	21:44313182
2:43278660	6:26044064	9:132597621	14:74551373	17:66454017	22:20104853
2:44395093	6:26056708	9:20684205	14:74551518	17:73285334	22:31031643
2:44588941	6:26285867	9:33024917	14:77279361	17:8062199	22:39096602
2:73461483	6:26458525	9:35749014	14:90849847	17:81009636	22:39101619
2:73613341	6:26521435	10:102673063	15:41576159	18:21083297	22:39101633
2:86422660	6:29716901	10:104263675	15:45021227	18:29672559	22:41777100
3:119813282	6:31126230	10:104677887	15:55700625	18:56338678	22:50964153
3:121468885	6:31239937	10:112679036	15:75230502	18:77793734	22:51021579
3:156392135	6:31588384	10:17271110	15:80444873	19:10981811	

B. Imbalanced sites detected in non-IBD samples

1:10270386	3:47422152	6:31430799	10:1102710	14:39901596	19:11266584
1:107599258	3:48754877	6:32145707	10:111985946	14:55583477	19:13044544
1:110880945	3:49044713	6:32821859	10:126605316	14:55738018	19:13056557
1:110881432	3:57583265	6:33290872	10:16859618	14:70054406	19:1490703
1:12040324	3:81810725	6:43737794	10:18948212	14:74416945	19:1490888
1:151763246	3:9438746	6:47445789	10:31320553	14:74551373	19:16582937
1:155108287	4:103748797	6:49430974	10:31320967	14:75725769	19:17337223
1:156698346	4:108641437	6:96025384	10:31321291	15:31618590	19:18263871
1:16162472	4:113558640	7:100450224	10:3215348	15:31618659	19:18314917
1:17338460	4:120375727	7:100472826	10:52384485	15:41709195	19:18439383
1:203274618	4:120375782	7:100487520	10:6205827	15:55700348	19:20162985
1:22109796	4:120375980	7:104653265	10:69644335	15:57900059	19:3435254
1:227127618	4:140375296	7:107384127	10:69644341	15:75230502	19:37808765
1:26362001	4:185570553	7:112090279	10:88855300	15:77363451	19:41869756
1:26758748	4:185655402	7:112090360	10:99079023	15:78441769	19:50093572
1:36851843	4:26859258	7:123174842	11:105893240	15:89089657	19:54641388
1:43814864	4:38665818	7:127032807	11:46722221	16:1458381	19:55770488
1:85742264	4:48271762	7:149321803	11:535649	16:1832615	19:55896795
2:101179341	4:48271963	7:150755205	11:5706311	16:2933037	19:55897327
2:10953001	4:7069901	7:27702670	11:57479864	16:30934075	19:56098846
2:113403306	5:130233745	7:27779729	11:64085148	16:4817296	19:56111036
2:148778272	5:131826413	7:36192401	11:64546106	16:68344945	19:56165194
2:174830206	5:176730678	7:39772916	11:74660292	16:72042635	19:56165921
2:176033215	5:57755936	7:95226257	11:77531890	16:88729573	19:5904069
2:190648805	5:6713403	8:104427359	11:809666	16:89883208	19:8578860
2:198365043	5:74532782	8:11660168	11:95523030	17:1359680	20:30161123
2:201390750	5:98109148	8:145008443	11:95523123	17:18266975	20:31407661
2:216176780	6:109330859	8:145022877	12:104323950	17:20059342	20:32254831
2:219433273	6:111580561	8:145026450	12:104324148	17:2296014	20:37376734
2:224822043	6:12009256	8:19674725	12:109125339	17:2304795	20:42086540
2:242254737	6:12011664	8:41386674	12:109125340	17:26368839	20:47663382
2:27651375	6:122720806	8:8243927	12:123451018	17:28705875	20:47894966
2:46926719	6:122793033	9:100818720	12:49076036	17:43138597	21:30365322
2:71357634	6:135818897	9:100819070	12:58087737	17:4870893	21:30375105
2:85645545	6:143858750	9:106856293	12:6580249	17:5323125	21:43648366
2:9771200	6:163149024	9:111882365	12:6651681	17:61851435	21:43648423
3:100120269	6:24721165	9:130159333	12:66696255	17:73521599	21:46707897
3:119182598	6:24721584	9:139839297	12:9102575	17:73901143	22:19166263
3:119813282	6:26123208	9:34049224	12:96252619	17:79980756	22:29137870
3:12598600	6:26124303	9:34178974	13:101327245	17:79980993	22:31063939
3:142720348	6:27440932	9:35749014	13:22178258	17:81009595	22:50639636
3:167452504	6:29855320	9:6681229	13:76123641	18:3261875	22:50964153
3:179065488	6:29910189	9:74979966	13:77903392	18:3261899	22:50964862
3:195809080	6:31126230	10:101380224	14:103851775	18:32870210	
3:39448419	6:31126232	10:104263675	14:20811588	18:77794410	
3:42003698	6:31367838	10:104677887	14:35591726	19:10946416	

Reference

Buchkovich, M. L., Eklund, K., Duan, Q., Li, Y., Mohlke, K. L., & Furey, T. S. (2015). Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. *BMC medical genomics*, *8*, 43. doi:10.1186/s12920-015-0117-x

Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current protocols in molecular biology*, *109*, 21.29.1-9. doi:10.1002/0471142727.mb2129s109

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature genetics*, *48*(10), 1284-1287.

FureyWiki. (n.d.). Retrieved March 17, 2019, from http://fureylabwiki-dept-fureylab.cloudapps.unc.edu/index.php/Main_Page

Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2009). TagDust--a program to eliminate artifacts from next generation sequencing data. *Bioinformatics (Oxford, England)*, *25*(21), 2839-40.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1), pp. 10-12. doi:https://doi.org/10.14806/ej.17.1.200

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., … DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-303.

Stevenson, K. R., Coolon, J. D., & Wittkopp, P. J. (2013). Sources of bias in measures of allele-specific expression derived from RNA-sequence data aligned to a single reference genome. *BMC genomics*, *14*, 536. doi:10.1186/1471-2164-14-536

Thomas D. Wu, & Colin K. Watanabe (2015). GMAP: a genomic mapping and alignment program for mRNA and EST sequences, Bioinformatics, Volume 21, Issue 9, Pages 1859–1875, https://doi.org/10.1093/bioinformatics/bti310

Verstockt, B., Smith, K. G., & Lee, J. C. (2018). Genome-wide association studies in Crohn's disease: Past, present and future. *Clinical & translational immunology*, *7*(1), e1001. doi:10.1002/cti2.1001

Wagner, J. R., Ge, B., Pokholok, D., Gunderson, K. L., Pastinen, T., & Blanchette, M. (2010). Computational analysis of whole-genome differential allelic expression data in human. *PLoS computational biology*, *6*(7), e1000849. doi:10.1371/journal.pcbi.1000849

Weiser, M., Simon, J. M., Kochar, B., Tovar, A., Israel, J. W., Robinson, A., Gipson, G. R., Schaner, M. S., Herfarth, H. H., Sartor, R. B., McGovern, D., Rahbar, R., Sadiq, T. S., Koruda, M. J., Furey, T. S., … Sheikh, S. Z. (2016). Molecular classification of Crohn's disease reveals two clinically relevant subtypes. *Gut*, *67*(1), 36-42.