RNA-Seq Splicing Analysis Pipeline

Ryan Schubert, Bree Coffey, Ally Miley

**Abstract**

RNA-Sequencing (RNA-Seq) allows for whole transcriptome analysis and profiling using high-throughput sequencing, and has become the preferred method for transcriptome analysis, over previous methods such as gene expression microarray. Using RNA-Seq, we can measure gene expression, annotate complete transcripts, and characterize alternative splicing, among other things. Alternative splicing of RNA is caused by altered excision of introns into translated mRNA, resulting in altered final protein structure. This alternative splicing is one of the main links between genetic variation and disease, and further understanding of the sQTL’s (splicing Quantitative Trait Loci) associated with the traits they encode will allow for targeted study and treatment of diseases linked to alternative splicing. Here, we present a pipeline that takes in aligned RNA-Seq reads (.bam files), and outputs intron excision ratios and sQTL mapping results. We outline the challenges, required parameters, and analysis of the application of our tool.

**Introduction**

RNA-Seq data is a fast and useful way to analyze the transcriptome of an organism. One of the details revealed in the sequencing of RNA is the abundance of certain transcripts. The introns of a gene can be transcribed in many different ways due to alternative splicing. Different quantities of spliced RNAs leads to a changed expression level of specific proteins. One of the causes of alternative splicing is single nucleotide polymorphisms (SNPs). The alteration of a single nucleotide can change the site at which splicing of an intron occurs. These SNPs are referred to as splicing quantitative trait loci (sQTLs).

One way to derive information from sQTLs is to map them to the human genome. By locating the SNPs which differentiate between splice sites in two individuals, information about each sample can be gathered. Some sQTLs may have a connection to genetic variations, like diseases. Additionally, determining the ratio at which the introns are excised can provide a number which can be compared between samples. Humans with a higher intron excision ratio in a specific gene may be more inclined to have a specific disorder.

The goal for this pipeline is to take an input of mapped RNA-Seq reads and quantify the ratios at which introns have been excised, as well as to map the sQTLs based on the corresponding SNP variants. Currently, several softwares exist that tackle portions of this problem. One study outlined three protocols for RNA-Seq pipeline data analysis for Differential Gene Expression Analysis of RNA-Seq, Differential Expression and Usage of Isoforms, and Cryptic Splicing (Yalamanchili et al. 2017). However, their step-by-step protocols have neither intron excision quantification, nor a mapping sQTL function. On the other hand, SUPPA was a software pipeline specifically created to calculate relative inclusion values of alternative splicing events from many samples, but this pipeline cannot identify sQTLs (Trincado et al. 2018). Finally, a Java visualization tool called SpliceSeq can analyze alternative splicing events from RNA-Seq data by making splice graphs, and predicting protein isoforms (Ryan et al. 2012). Many tools have been created for RNA-Seq data analysis, but this Splicing Analysis Pipeline both differs from and utilizes existing technologies to create a user-friendly pipeline for quantifying intron excision ratios and mapping SNPs at alternative splicing sites.

**Implementation**

The original planned workflow involved the use of WASP to filter biased reads from sample data. After WASP filtering was performed, reads were to be analyzed by Leafcutter for intron excision ratios and sQTL’s were to be mapped and identified using FastQTL **(Fig 1)**. This original plan was meant to mimic the experimental workflow used in the original Leafcutter documentation for the the analysis of alternative splicing. However, while these methods were suggested by the original Leafcutter documentation, due to poorly updated documentation, a different path was decided.

WASP’s application in particular was found to have poor documentation as many of its dependencies were found to be outdated. The filtering step by WASP only served to make Leafcutter more efficient, and updating each of its dependencies throughout the server was found to be too time consuming. So, it was decided to abandon the filtering step. FastQTL also was found to have poor documentation. After testing its installation for several days it was decided to use MatrixEQTL as a comparable replacement.

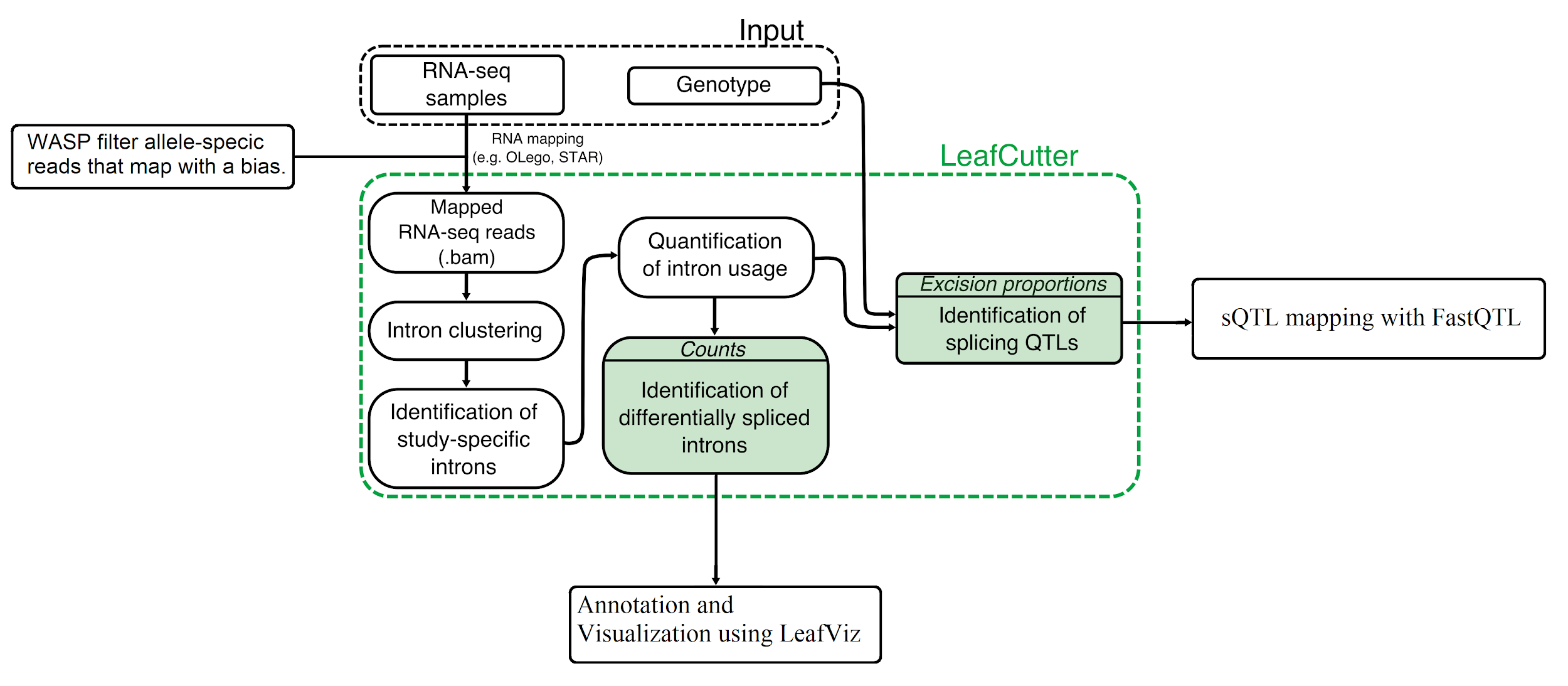
The new path then follows the route of Bam file, into Leafcutter, into MatrixEQTL **(Fig 1b)**. By itself, Leafcutter is capable of quantifying intron excision ratios using the novel method of mapping split reads. The data output by Leafcutter also boasts, upon analysis, the increased identification of sQTLS. This is how MatrixEQTL is applied. The output data from Leafcutter as well as the accompanying SNP data are parsed independently for input into MatrixEQTL. The strength of this pipeline rests in not simply binding two softwares together, but in its smooth translation from one software to another. Data does not need to be manually parsed by the user, as scripts for this will be provided.

Parsing the SNP files for each chromosome was found to take 23 minutes for one chromosome. Parsing the data for all 23 SNP files is estimated to then take approximately 9 hours. Leafcutter meanwhile took approximately 20 minutes to calculate excision ratios from five bam files. Calculating the excision ratios from the entire data set of 495 bam files is estimated to take 33 hours. Exact estimates for MatrixEQTL processing remains to be tested. The parser for the vcf file is of complexity at worst case O(m\*n), where m is the number of SNPs, and n is the number of samples. If users are able to successfully install WASP, it is recommended that bam files are filtered for bias reads as this will both improve the pipeline’s accuracy and improve Leafcutter’s efficiency. Additionally, FastQTL is purported to be several magnitudes faster than MatrixEQTL. Should it be successfully implemented, the Leafcutter output may be easily altered for FastQTL.

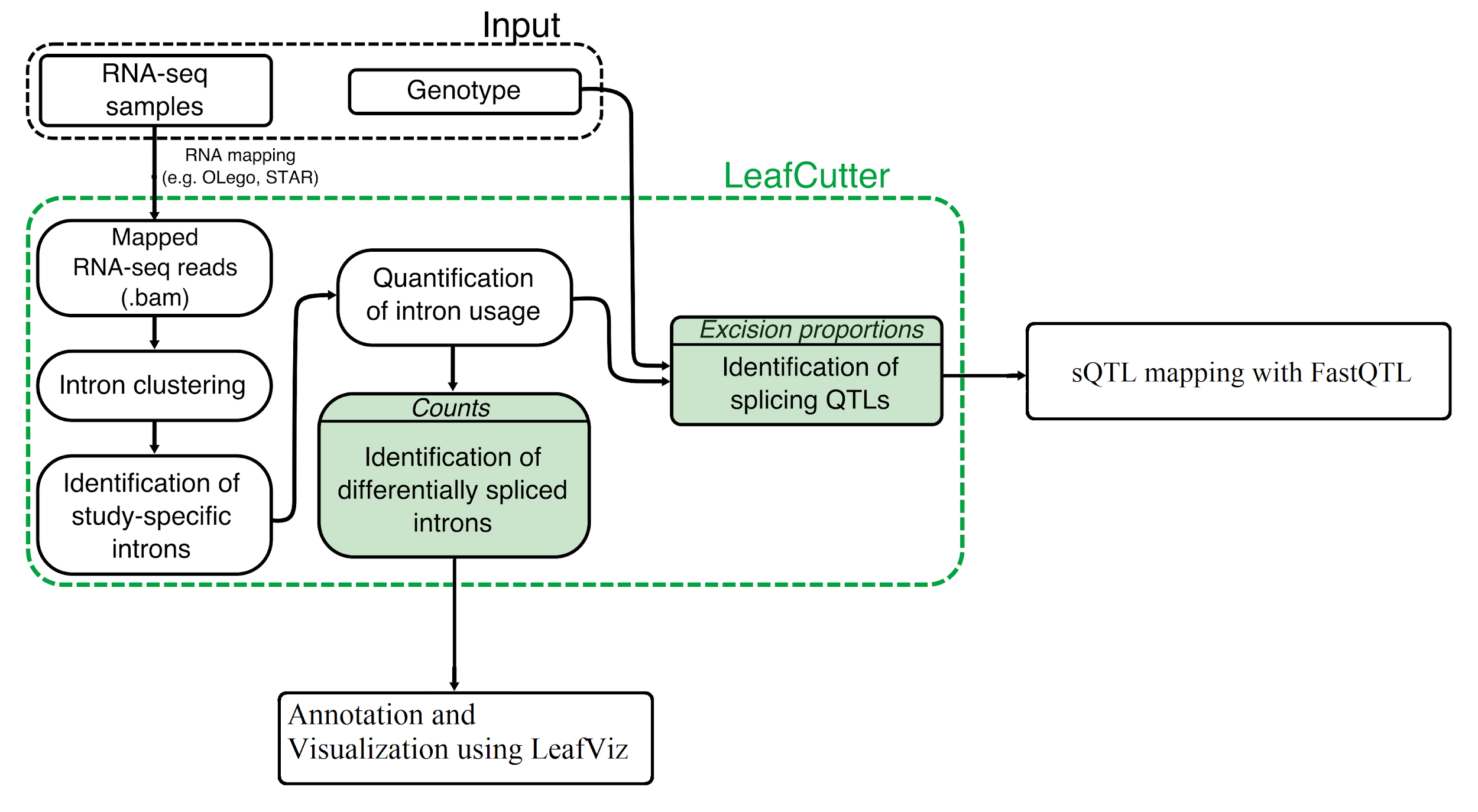
**Results**

So far, the pipeline still remains in its pieces. However, the next steps will be to finalize a shell script that enables a smooth running of the analysis. Individually, the LeafCutter software has shown success on testing four of the 456 total bam files from the GEUVADIS input data. Leafcutter is able to effectively measure intron excision ratios on a small scale and larger scale testing is underway. Once successful analysis of these bam files is complete, the output is easily parsed into the necessary MatrixEQTL inputs Next, the variant call format files for the SNPs for each chromosome were run through a parser to output the genotype file necessary for MatrixEQTL. Additionally, a file was created that contained the SNP id, chromosome number, and position of the SNP on that chromosome. Then, the results from the genotype file are combined with the output of LeafCutter to reorder the samples, because MatrixEQTL requires them to match. So far, our group has had success with these first steps. The outputs from the vcf files have been parsed to match the LeafCutter data. Currently, we are working on a script to run our final data through MatrixEQTL. We have estimates of our performance statistics, and are working on ways to speed up the process. However, these file types are large, and iterating through them can take a long time.

**Figure 1a - Original planned workflow of pipeline**



**Figure 1b - New planned workflow**



**References**

https://cran.r-project.org/web/packages/MatrixEQTL/MatrixEQTL.pdf

https://davidaknowles.github.io/leafcutter/

Li Y. I., Knowles D. A., Humphrey J., Barbeira A. N., Dickinson S. P., Im H. K., Pritchard J. K. (2017) LeafCutter: annotation-free quantification of RNA splicing. bioRXiv, 044107.

Ryan, M. C., Cleland, J., Kim, R., Wong, W. C., & Weinstein, J. N. (2012). SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics*, *28*(18), 2385–2387.

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biology, 19(1), 40.

Yalamanchili H. K., Wan Y. W., Liu Z. (2017). Data Analysis Pipeline for RNA-seq Experiments: From Differential Expression to Cryptic Splicing. Curr Protoc Bioinformatics, 59(11), 15.1-15.21.