

A Tutorial for the package GLiMMPS

Keyan Zhao

Department of Microbiology, Immunology, and Molecular Genetics

University of California, Los Angeles

Los Angeles, CA 90095-7278

Email: kyzhao@ucla.edu

Alternative Email: kyzhao@gmail.com

Yi Xing

Department of Microbiology, Immunology, and Molecular Genetics

University of California, Los Angeles

Los Angeles, CA 90095-7278

Email: yxing@ucla.edu

February 20, 2013

Contents

1 Overview	2
2 Prerequisites	3
3 Download GLiMMPS source code	3
4 Preparing Data	3
4.1 RNA-Seq mapping and processing	4
4.2 Obtain the junction sequence read counts for all possible alternative splicing events in the population.	4
4.3 Obtain the genome-wide genotype data for all individuals in population in plink format	4
5 Run statistical models for sQTLs	5
5.1 Input files for sQTL analysis.	5
5.2 One example of running sQTL analysis for all cis SNPs within 200kb of the target exon. . . .	5
5.3 Plot Exon inclusion level for one significant sQTL.	6
5.4 Overlapping GWAS signal with sQTL signal.	6

1 Overview

Alternative splicing (AS) is the process by which the exon sequences from precursor mRNA transcripts are differentially included in mature mRNA resulting in different isoforms, and it serves as a major contributor to both protein diversity and control of gene expression levels. The development of high-throughput RNA sequencing (RNA-Seq) technology provided an efficient way of quantifying alternative splicing variation. This technology has several advantages over exon arrays, including greater dynamic range of exon expression levels, ability to detect novel transcripts not probed on the array, single nucleotide level resolution, and less confounding effects from polymorphisms on the target exons.

Despite the significant findings in previous RNA-Seq based splicing quantitative trait loci (sQTL) studies to characterize the genetic variation of alternative splicing, the statistical models applied didn't take into account many variations in the data. Here, we developed a robust and flexible method for sQTL analysis from RNA sequencing data. Our method, a Generalized Linear Mixed Model prediction of sQTL (GLiMMPS), takes into account the individual variation in sequence depth and overdispersions prevalent in the data. A schematic graph showing the main idea is illustrated in Figure 1

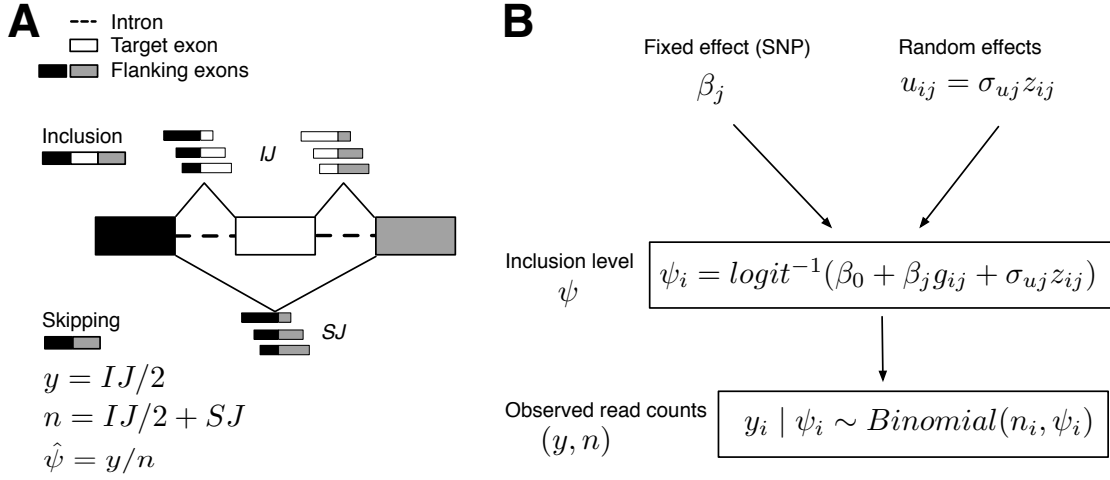


Figure 1: **Generalized Linear Mixed Model prediction of sQTL.**

GLiMMPS can process various types of Alternative Splicing events listed in Figure 2.

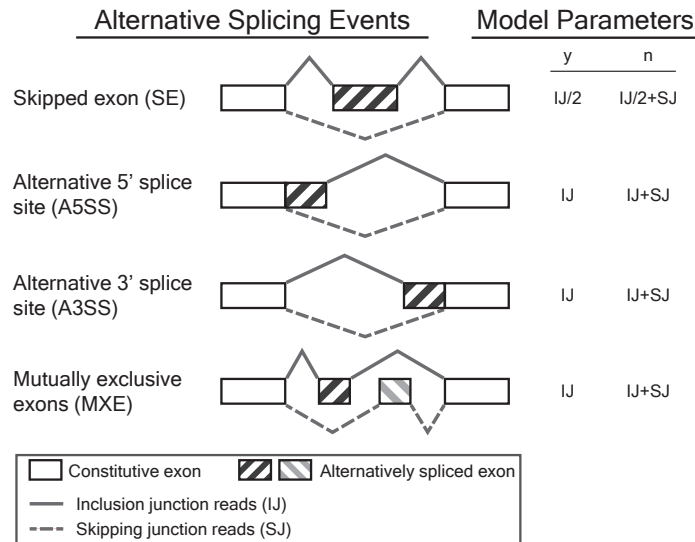


Figure 2: **Various types of alternative splicing events analyzed.**

This vignette walks the user through the process of applying GLiMMPS for sQTL analysis using RNA-Seq data in the population. The methods in this package was introduced in the paper of Zhao et al. (2012).¹

2 Prerequisites

GLiMMPS used **Python** and **Perl** scripts for RNA-sequence analysis and data preparation. Thus you must have **Python** and **Perl** installed on your computer. For RNA sequence mapping and processing, you need to have a few Bioinformatics tools installed.

1. Download and install RNA sequence mapping tools **bowtie** (version 0.12.8 or above) <http://bowtie-bio.sourceforge.net/index.shtml> and **tophat** (version 1.4.0 or above). <http://tophat.cbcb.umd.edu/downloads/>
2. Download and install **samtools**. <http://samtools.sourceforge.net/>
3. Download and install **plink**, a tool for misc genotype processing. <http://pngu.mgh.harvard.edu/~purcell/plink/>

After installing these softwares, make sure their installed binaries are included in your \$PATH environment variable. To add the software path to your \$PATH variable, use export command as follows at command prompt: e.g. , suppose tophat is installed at your local folder /home/you/tophat1.4.0/, then do the following at command line:

```
export PATH=$PATH:/home/you/tophat1.4.0/
```

You need to add command to /etc/profile or /home/you/.bash_profile so that PATH get set automatically after each reboot.

The statistical analysis and plotting were implemented in **R** scripts. Thus you must install R on your computer. R can be downloaded from <http://cran.r-project.org/>. Also there are a few prerequisite R packages needed before running GLiMMPS in **R**.

1. lme4
2. MASS

You can install packages in R in the following way:

```
>install.packages(c("lme4","MASS"))
```

3 Download GLiMMPS source code

Download the source code: GLiMMPScode.tar.gz from GLiMMPS website. For scripts required for the downstream analysis, R scripts are in folder Rscripts/ and other python/perl scripts are in pythonperlsrscs/.

4 Preparing Data

Download the human genome reference files and the human genome annotation GTF file from GLiMMPS website.

Because of the large data files of RNA sequence files, we only provided example datasets after the sequence mapping and junction reads counting.

¹ Zhao Keyan, et al. Robust statistical model for regulatory variation of alternative splicing using RNA-Seq data. Submitted

4.1 RNA-Seq mapping and processing

Mapping the RNA-Seq data using **tophat**:

```
tophat -o NA06985 --transcriptome-index Ensembl_r65.gtf -a 8 -m 0 -I
300000 -p 2 -g 20 --library-type fr-unstranded --initial-read-mismatches
3 --segment-mismatches 2 hg19 fastqfiles/NA06985.fastq
```

We need to run the mapping process for all individuals.

Process tophat mapping results: accepted_hits.bam file with the following commands to get uniquely mappable reads in sam format. Suppose the read length is 50bp, then do the following:

```
samtools view -Xh accepted_hits.bam | awk -F"\t" '($0 ~ "NH:i:1[~0-9]" || $0 ~ "NH:i:1$")
&& ($2 ~ "P" && (($6 == "50M") || ($6 ~ "N" && $6 ! ~ "D" && $6 ! ~ "I")))) || NF < 7' > unique.sam
```

We need to run the process for all individuals, it can be done using the sample batch processing script for all individuals: pythonperlsrscs/batchtophat2uniq.pl

4.2 Obtain the junction sequence read counts for all possible alternative splicing events in the population.

For all individuals in the population, find all possible Alternative Splicing (AS) events in the population based on either the gene annotation in the GTF file or the new splicing site identified from tophat.

```
../../pythonperlsrscs/batch_allASevents.pl config.GLiMMPS.txt
```

Main parameters for all data processing scripts are in this file configuration file : config.GLiMMPS.txt. One example is provided in the example folder.

Then for each individual, Run the following script to get all the junction read counts for all the AS events in the population found from above.

```
../../pythonperlsrscs/batch_getASreads.pl config.GLiMMPS.txt

## summarize all types of AS events for all individuals.
../../pythonperlsrscs/summarizeallexoninc.pl config.GLiMMPS.txt
```

```
#Filter to get psi range >0.1 and median.n >=5 and unique major events for each target exon.
cd Exon_Inc_Simple
./summarystat_exonmin5.R
```

4.3 Obtain the genome-wide genotype data for all individuals in population in plink format

The downstream sQTL analysis will use genotypes from plink formatted genotype files with each chromosome in one file. Users can refer to the example/Genotype folder. The .ped and .map file format description can be found at the **plink** website: <http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped> .

5 Run statistical models for sQTLs

5.1 Input files for sQTL analysis.

After Step 4, we have the AS exon information and the Inclusion junction and total junction count matrix for all individuals, as well as the plink-format genotype files. Please refer to the example/ folder for examples. Doing the analysis for all AS exons is very time consuming. Running the jobs on the cluster or machines with many CPUs are highly recommended.

5.2 One example of running sQTL analysis for all cis SNPs within 200kb of the target exon.

Here, Use the example dataset, we will illustrate the analysis using one exon in SP140 on chr2.

```
cd example/CheungCEU/Exon_Inc_Simple/  
  
sQTLregress.oneexon.R alltype/bychrs/exonsinfor.plink.5reads.chr2.txt  
alltype/bychrs/plink.5reads.allreads.chr2.txt alltype/bychrs/plink.5reads.IJ.chr2.txt  
../Genotype/HAPMAP1000G.CheungCEU41.chr2 1117
```

The arguments for **sQTLregress.oneexon.R** are :

1. name of exon information file.
2. name of file for all junction read counts.
3. name of file for exon inclusion junction read counts.
4. name prefix of genotype plink file (suffix .ped or .map not included).
5. The exon index of the target exon to be analyzed in the the exon information file (Integer).
6. Pvalue cutoff defines the significance cutoff of GLiMMPS for plotting. (This argument is optional, with default = 1e-5)

For details on the arguments to the program, run the program with no arguments (**./sQTLregress.oneexon.R**).

sQTLregress.oneexon.R read the exon information file, the junction read count files and the genotype files, and extract only the 1117th exon (SE_2110) and local SNPs within 200kb region around the exon. It calls functions defined in **GLiMMPS_functions.R**. Then multiple sQTL analysis models were carried out : linear model(lm), generalized linear model(glm), generalized linear model using quasibinomial family (glmquasi), generalized linear mixed model with LRT test (glmm), generalized linear mixed model with a Wald test(glmmWald), and P-values are written to the result file: tmpasso/chr2/SE_2110.asso

Chr	SNPID	Pos	pvals.glm	pvals.glmquasi	pvals.glmm	pvals.lm	pvals.glmmWald	Beta
2	rs4973246	230911791	9.172e-01	9.647e-01	6.138e-01	5.601e-01	6.109e-01	0.205
2	rs6722416	230912274	1.449e-01	5.409e-01	1.769e-01	4.814e-01	1.741e-01	-0.468
2	rs34529492	230913066	1.152e-01	4.936e-01	4.810e-01	1.665e-01	4.716e-01	-0.25
2	rs13402579	230913420	9.172e-01	9.647e-01	6.138e-01	5.601e-01	6.109e-01	0.205
...								

Where, column pvals.glmm is the P-value for our GLiMMPS model, and Beta is the fixed SNP effect estimate for the minor allele from GLiMMPS. After associations, the program will also check if there is any SNP passing the pval.cutoff, and plot exon inclusion levels against the significant sQTL SNP that is closest to the target exon Splice Site (like Figure 3).

5.3 Plot Exon inclusion level for one significant sQTL.

In Step 5.2, after we read in the genotype and junction reads files, we can also manually plot the exon inclusion levels for a SNP of interest by running the following script inside an R session. A plot file like Figure 3 will be generated.

```
>source("GLiMMPS_functions.R")  
>psi.geno.plot(y,n,"rs28445040",genesymbol, targetexonID,exon.coordinate)
```

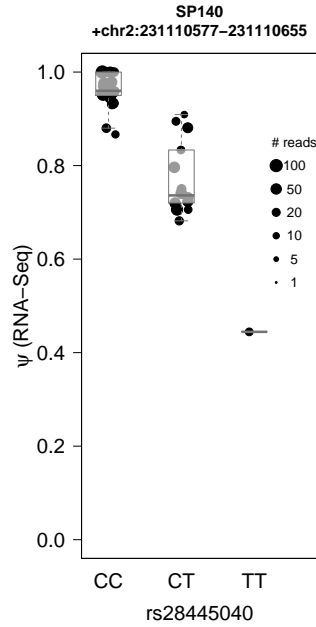


Figure 3: **Boxplot showing Exon inclusion level as a function of SNP genotype.** Dot sizes are proportional to total read counts for each individual.

5.4 Overlapping GWAS signal with sQTL signal.

After the sQTL association is done in Step 5.2, we can search against the GWAS catalog SNPs and look for sQTL signals overlapping GWAS signals.

```
cd example/CheungCEU/Exon_Inc_Simple/

sQTLGWAS.oneexon.R alltype/bychrs/exonsinfor.plink.5reads.chr2.txt ../gwascatalog.txt
../reference/Ensembl_r65.gtf.gff 1117 3.7e-6
```

The arguments for **sQTLGWAS.oneexon.R** are :

1. name of exon information file.
2. name of tab delimited file for GWAS SNPs downloaded from <https://www.genome.gov/26525384#download>
3. name of gene annotation file in gtf format
4. The exon index of the target exon to be analyzed in the the exon information file (Interger).
5. Pvalue cutoff defines the significance cutoff of GLiMMPS for plotting. (This argument is optional, with default = 1e-5)

For details on the arguments to the program, run the program with no arguments (**./sQTLGWAS.oneexon.R**).

It will generate a file with the subset of GWAS signals that overlapping sQTL signals and also a plot showing the GLiMMPS P-values for SNPs around 20kb of the target exon (as in Figure 4).

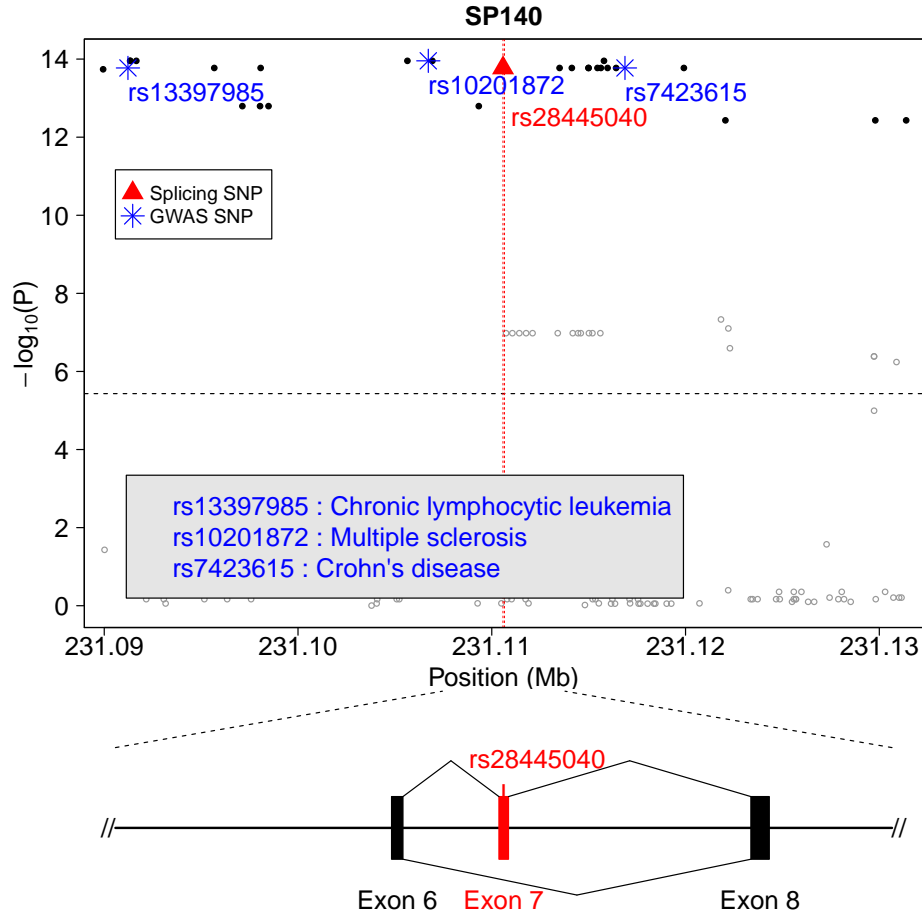


Figure 4: **sQTL signal overlapping with GWAS signal near the target exon.** GLiMMPS P-value distribution around the sQTL exon from gene SP140. The black horizontal dashed line reflects the p-value cutoff and red vertical lines mark the location of the sQTL exon. Those SNPs in linkage disequilibrium ($r^2 > 0.8$) with the GWAS SNPs are shown in solid black dots, while other SNPs are shown in grey circles. The corresponding traits for the GWAS SNPs are listed in the grey box. Exon structure is shown in the bottom with splicing SNP marked at corresponding location.