Quantifying gene and allele-specific expression simultaneously using personalized human genomes

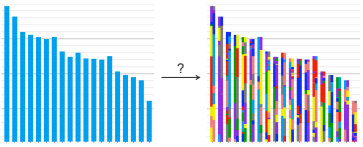
Narayanan Raghupathy, Kwangbom Choi, Steven C. Munger, and Gary A. Churchill

The Jackson Laboratory, Bar Harbor, Maine 04609 USA

The Jackson Laboratory
Leading the search for tomorrow's cures

Summary

Large-scale genome sequencing efforts have characterized millions of common genetic variants across human populations. However development of tools that can effectively utilize this individual-specific variation to inform quantitation of gene expression abundance have lagged behind. We present tools to utilize individual genetic variation in RNA-seq and quantify gene expression and allele-specific expression (ASE) simultaneously.



- Seqnature: Builds individualized genomes using known genetic variations: SNPs and Indels (Munger, Raghupathy et. al. submitted to **Genetics**)
- EMASE: Simultaneous quantitation of gene expression and allele-specific expression (Raghupathy, Choi *et. al.* to be submitted to ***Genetics***)
- We use Seqnature to build personalized diploid human genome using 1000 Genomes variation data and apply EMASE to quantify gene expression and allele specific expression simultaneously.

Genetic Variations Matter in RNA-seq Analysis

If one is working with RNA-seq data from a genetically diverse population, a large number of genes can have one or more SNPs and indel.

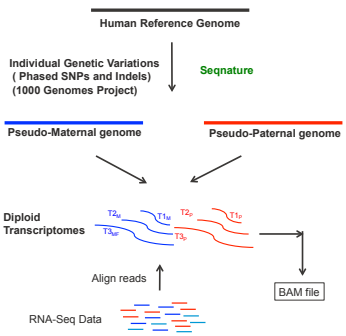
Genomic SNPs	~ 3.6 Million
Genomic Indels	~2.5*10^6
Coding SNPs	~25000
Coding Indels	~200

SNP/Indel Statistics: An YRI Individual from 1000 Genomes project

Aligning short sequencing reads to a common reference genome is the first step in RNA-seq analysis. Genetic variations present in the sample, but not in the reference genome can lead to misalignments and incorrect expression estimates and biased allele-specific expression.

SEQNATURE

Seqnature incorporates known SNPs and indels in to reference genomes to construct individualized diploid or haploid genomes, for human, genetically diverse heterozygous model organisms and inbred strains. Seqnature also updates the annotation file and it is readily usable for read alignment by common aligners.



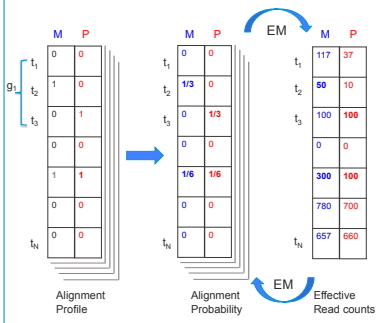
Seqnature is an open access tool written in Python and a beta version is available at https://github.com/jaxcs/Seqnature

EMASE: EM for Allele Specific Expression

EMASE is a model-based quantitation approach that employs an Expectation Maximization (EM) algorithm to apportion

- gene multireads
- isoform multireads
- haplotype multireads

EMASE takes alignment results from diploid transcriptome from a common aligner and uses an EM algorithm to estimate gene expression and ASE simultaneously.



Alignment profile and probability are sparse matrices of size NxRx2, where N is number of transcripts, R is number of reads.

Aligning every read to the diploid transcriptome after accounting for genetic variations help to avoid alignment biases and in accurate quantitation.

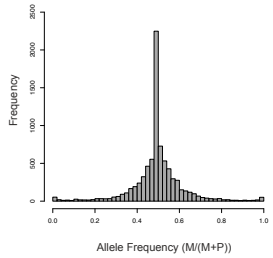
EMASE Offers Tools to Mine RNA-seq Alignments

In addition to estimating expression and ASE simultaneously, EMASE offers functions to mine RNA-seq alignments, like information content in terms of uniquely aligned reads, alignment probabilities, mappability, and analyzing alignments from simulation.

EMASE is implemented in Python. It is available upon request now and will be available soon from http://cgd.jax.org/.

Allele Specific Gene Expression

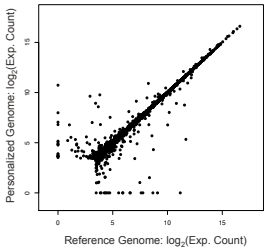
A common problem in quantifying ASE from RNA-seq data is the alignment bias due to common reference genome. Our diploid model accounting for known genetic variations removes the alignment bias. The allele frequency histogram below is symmetric without any reference bias and it shows allele specific expression quantitative.



Gene Expression: Common Reference Genome vs. Personalized Genome

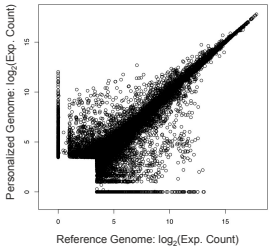
Comparing alignment and gene expression results from using common reference genome and personalized diploid genome show improvement in alignments and expression estimates.

In a single sample, typically about 500-1000 genes show improved expression abundance estimate from aligning to diploid transcriptome (Simulation results not shown). This is mainly due to gained alignments from reads that did not align when a single reference transcriptome is used and the ability to resolve alignments to pseudo-genes and homologous gene families correctly. For example, SNPs/Indels between a gene and its pseudo-gene, and a gene and its homologus family members help differentiate the alignments.



Population-level gene expression analysis

Using personalized diploid genome/transcriptome for expression quantitation is highly relevant for population level RNA-seq analysis. Comparing over 50 RNA-seq samples from HapMap/1000 Genomes Yoruban population show that about 4,000 genes improve in expression estimate and lead to a better understanding of *Cis* regulation.



References

- Steven C. Munger, **Narayanan Raghupathy**, *et. al.* RNA-seq alignment to individualized genomes improves transcript abundance estimates in multiparent populations, submitted to *Genetics*
- **Narayanan Raghupathy**, K.B. Choi, *et. al.* EMASE: Quantifying allele-specific gene expression. To be submitted to *Genetics*.
- The 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, *Nature* 2012
- Joseph Pickrell, John Marioni, *et. al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010.
- Lappalainen *et. al.* Transcriptome and genome sequencing uncovers human functional variations, *Nature* 2013

Funding

- This project is funded by National Institutes of Health (NIH) grant P50GM076468.

Conclusions

Current RNA-seq approaches employ two steps to quantify gene expression abundance and ASE from RNA-seq data;

- Gene-level abundance is estimated from genome-wide alignments
- ASE is assessed separately by analyzing only reads that overlap known SNP locations. Often multimapping reads are ignored in ASE quantitation.

Both the gene expression and ASE estimates suffer from alignment biases if one is working with genetically diverse population. We developed complimentary software (open access) tools Seqnature and EMASE that can account for individual genetic variations in a diploid model and quantify gene expression & ASE simultaneously by using gene, isoform, and haplotype level multireads.

Our results show that using diploid transcriptome at the alignment step and dealing with multireads using an EM approach results in better estimates of gene expression and ASE. Our results also show the potential use of diploid model in understanding *cis*-regulation using population data.

Contact: Narayanan.Raghupathy@jax.org

www.jax.org