Inference of allele specific expression levels from RNA-Seq data

Sahar Al Seesi and Ion Măndoiu

Computer Science and Engineering
University of Connecticut
{sahar,ion}@engr.uconn.edu

Abstract. Accurate allele specific expression estimation requires the availability of a diploid transcriptome, which makes it a challenging problem. Most existing methods rely on simple counting of alleles coverage at heterozygous Single Nucleotide Polymorphic sites. In this work, we present RNA-PhASE, a pipeline for Allele Specific gene and isoform Expression estimation from RNA-Seq Reads. The pipeline integrates methods for SNV detection and phasing with a new diploid version of an Expectation Maximization algorithm for gene/isoform estimation. Within this pipeline, we couple an existing phasing algorithm with a novel method for coverage based phasing.

1 Introduction

Most current methods for estimating gene/isoform expression levels from high-throughput whole transcriptome sequencing (RNA-Seq) data rely on mapping the reads to a reference genome and/or transcriptome and do not consider the difference between the two parental alleles (diploid transcriptome). The diploid transcriptome can be easily inferred when a diploid genome is available, as in recent studies of cis- and trans-regulation [8] and parent-of-origin effects [5] that use hybrids of inbred species or strains. However, reconstructing the diploid genome of human subjects remains a difficult task [3]. Hence, existing studies of allele-specific gene expression rely on simple alleles coverage analysis for heterozygous Single Nucleotide Polymorphic (SNP) sites within transcripts. Such approaches typically do not allow inference of allele-specific expression of individual gene isoforms, result in less robust estimates since they use only RNA-Seq reads that overlap heterozygous SNP sites, and are affected by systematic read mapping biases toward reference alleles [1][6].

In this work, we integrate a recent method for SNV detection and genotyping from RNA-Seq data [4] with the scalable haplotype reconstruction method [2] and a diploid version of the Expectation Maximization (EM) algorithm for isoform expression estimation of [9] into a pipeline for estimation of allele-specific isoform expression levels. Our pipeline, RNA-PhASE, does not require genome sequencing data, but can incorporate such data when available. Inferring the two haplotypes and re-mapping the reads against the diploid transcriptome resolves
the above mentioned bias towards reference alleles, while the EM model improves
inference accuracy by using all reads, including those that map to more than one
isoform, incorporating additional sources of disambiguation information such as
the distribution of RNA-Seq fragment lengths, and correcting biases introduced
by library preparation and sequencing protocols.

Preliminary results show the ability of the proposed pipeline to accurately
infer allele specific isoform expression levels for synthetic hybrids with varying
levels of heterozygosity, generated by pooling whole brain RNA-Seq reads of
different mouse strains studied as part of the Sanger Institute Mouse Genomes
Project [7].

2 Methods

The RNA-PhASE pipeline, depicted in Figure 1, starts by mapping the RNA-Seq
reads against a haploid reference transcriptome and reference genome. Align-
ments from both mappings are merged together, according a set of rules de-
scribed in [5], and the resulting set of alignments are used to call SNVs. The
merging method, referred to as HardMerge, keeps a read if it aligns uniquely to
the genome only, uniquely to the transcriptome only, or to both provided that
the two alignments agree. Results have shown that this hybrid method results
in calling SNVs with very high confidence. We introduce a local alignment ver-
sion of HardMerge that works on the base level, discarding read bases mapped
to multiple locations. It then generates alignments from contiguous stretches of
non-ambiguously mapped bases. This modification enables HardMerge to handle
local alignments of long RNA-Seq reads generated by technologies like 454 and
ION Torrent. SNVs are then called using SNVQ [4], which uses Bayes rule to call
the genotype with the highest probability while taking base quality scores into
account.

For haplotyping, we couple an efficient Single Individual Haplotyping algo-
rithm, RefHap [2], with a novel method for coverage based phasing. Our new
method merges phased blocks in the RefHap output, and it phases called SNVs
that were not phased by RefHap because they are not in close proximity with
other SNVs and consequently there is no read evidence that can be used to
phase them. In coverage based phasing, for two successive heterozygous SNVs
i and j, the i’s allele with highest coverage is paired with j’s allele with highest
coverage in the same haplotype, and similarly lowest coverage alleles are paired
in the other haplotype. When one or both SNVs have equal coverage for the two
alleles, phasing is done arbitrarily. i and j can be two SNVs for which the phase
was not resolved by RefHap. Alternatively, j can be first SNV in a phased block
and i is the last in the most adjacent SNV preceding j.

Allele Specific Expression (ASE) levels are estimated through realigning the
reads against the diploid transcriptome and feeding the mapping results into a
diploid version of IsoEm [9], an EM algorithm that makes use of information
such as insert size, quality scores, and read pairing, if available, to handle read
Fig. 1. RNA-PhASE: Pipeline for Allele Specific Expression inference from RNA-Seq data through calling and phasing expressed Single Nucleotide Variations

mapping ambiguities. Finally, allelic expression imbalance is inferred through applying Fisher’s Exact test.

3 Experimental Results

We test RNA-PhASE against synthetic hybrids data created by merging whole brain RNA-Seq reads from the Sanger Institute Mouse Genomes project. Four synthetic hybrids data sets were created by merging equal number of reads from C57BL/6NJ with each of the following strains: BALB/cJ, A/J, CAST/EiJ, and SPRET/EiJ. The four strains were selected to provide the test of RNA-PhASE performance with varying levels of heterozygosity. As a measure of strain variation compared to C57BL/6NJ, and thus heterozygosity level of the synthetic hybrids, we use the number of genomic SNVs reported in [7]. The strains are listed here in an increasing variation order, compared to C57BL/6NJ.

Testing is done on two levels. First, we test the ability of the diploid IsoEM to accurately estimate ASE given the diploid transcriptome. This is done by creating diploid transcriptomes for the hybrids using the SNVs reported in [7]. The inferred expression level for each allele of an isoform or gene is compared with the expression level of that isoform/gene estimated from the corresponding strain reads when processed separately. We measure Pearson coefficient of correlation, error fractions (EF) and median percent errors (MPE). EF at a certain threshold $t$ is the percentage of isoforms (or genes) with relative error larger than given threshold $t$, where the relative error is calculated as the difference in estimated
to actual expression levels divided by the actual expression level. MPE is the threshold \( t \) for which \( EF \) is 50%. Figure 2 and tables 1 and 2 show these results.

![Fig. 2. Isoform and Gene Error Fractions. Error Fractions at different threshold values for expression levels estimated for strains in synthetic hybrids vs. corresponding separate strain.](image)

The second level of testing is for the whole pipeline, starting from the synthetic hybrid reads and haploid reference. In this case, a direct comparison of the ASE from the hybrids against the corresponding separate strain expression levels will not be feasible. Results accuracy will be determined by comparing which isoforms and/or genes are detected to have allelic imbalance in the hybrid vs. the corresponding separate strains. The allelic imbalance will be determined using Fisher’s Exact test. These results are currently being generated.

<table>
<thead>
<tr>
<th>Hybrid</th>
<th>C57BL Strain IE</th>
<th>Strain IE</th>
<th>C57BL Strain GE</th>
<th>Strain GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57BLxBALBc</td>
<td>0.705</td>
<td>0.675</td>
<td>0.706</td>
<td>0.675</td>
</tr>
<tr>
<td>C57BLxAJ</td>
<td>0.855</td>
<td>0.902</td>
<td>0.856</td>
<td>0.903</td>
</tr>
<tr>
<td>C57BLxCAST</td>
<td>0.872</td>
<td>0.824</td>
<td>0.924</td>
<td>0.882</td>
</tr>
<tr>
<td>C57BLxSPRET</td>
<td>0.952</td>
<td>0.726</td>
<td>0.951</td>
<td>0.725</td>
</tr>
</tbody>
</table>

**Acknowledgments.** This project was supported by in part by awards IIS-0546457 and IIS-0916948 from NSF, and Agriculture and Food Research Initiative Competitive Grant no. 2011-67016-30331 from the USDA National Institute of Food and Agriculture.
Table 2. MPE for isoform expression levels estimated for strains in synthetic hybrids vs. corresponding separate strains.

<table>
<thead>
<tr>
<th>Hybrid</th>
<th>C57BLxStrain</th>
<th>C57BL</th>
<th>Strain</th>
</tr>
</thead>
<tbody>
<tr>
<td>C57BLxBALBc</td>
<td>0.3874</td>
<td>0.9075</td>
<td></td>
</tr>
<tr>
<td>C57BLxAJ</td>
<td>0.6281</td>
<td>0.4339</td>
<td></td>
</tr>
<tr>
<td>C57BLxCAST</td>
<td>0.2276</td>
<td>0.1840</td>
<td></td>
</tr>
<tr>
<td>C57BLxSPRET</td>
<td>0.1871</td>
<td>0.1753</td>
<td></td>
</tr>
</tbody>
</table>

References