An Alignment-free Regression Approach to Estimating Allele-Specific Expression in F1 Animals

Chen-ping Fu, Vladimir Jojic, and Leonard McMillan

December 13, 2012

Abstract

We wish to study allele-specific expression in diploid organisms, specifically in F1 animals with inbred parental strains. Current methods for analyzing allele-specific expression rely on read alignment, which leads to reference bias unless there is prior knowledge of all genomic variants in the parental strains. However, in the case where RNA-seq data is available for both parental strains, we do not need prior knowledge of parental genomic variants. Our approach first uses parental RNA-seq reads to create maternal and paternal versions of transcript sequences, then estimates allele-specific expression levels in the F1 animal for each transcript. Using the parental versions of all candidate transcripts as features, we use a modified lasso penalized linear regression model for estimating abundance levels of expressed transcripts in the F1 animal.

We tested our methods on synthetic data from the mouse transcriptome and compared our results with those of Trinity, a state-of-the-art de novo RNA-seq assembler. Our methods achieved much higher sensitivity and specificity in both identifying expressed transcripts and transcripts exhibiting allele-specific expression. We were also able to separately predict relative expression levels from paternal and maternal strains with more accuracy.

1 Introduction

Recent advances in high-throughput mRNA sequencing (RNA-seq) technology have enabled the generation of massive amounts of data. While this offers exciting potential for studying known transcripts and discovering new ones, it also necessitates new bioinformatic tools that can efficiently and accurately analyze such data.

Current RNA-seq techniques generate short reads from mRNA sequences at high coverage, and the main challenge in RNA-seq analysis lies in reconstructing transcripts and their relative abundances from millions of short read sequences. One approach is to map short reads onto a reference genome, then subsequently estimating the abundance in each gene region. These read-alignment methods include Cufflinks [18] and Scripture [6], which use algorithms such as the Burrows-Wheeler transform to achieve fast read alignment. Although these methods are well established in the RNA-seq community and offer many auxiliary tools [16, 8, 17] for downstream analysis, read alignment to a reference genome offers a couple of disadvantages. First, read alignment assumes that the samples being aligned are genetically similar to the reference genome, and samples deviating greatly from the reference may have a large portion of unmappable reads. Second, mapping methods typically cannot resolve reads...
which map to multiple locations in the genome, resulting in reads being arbitrarily mapped or discarded from analysis. Some suggested workarounds to the first problem of reference bias involve creating new genome sequencies, typically by incorporating known variants, to use in place of the reference genome for read alignment [14]. However, this requires prior knowledge of genomic variants in the targeted RNA-seq sample, which may not always be possible to obtain.

Another class of methods perform de novo assembly of transcriptomes using De Bruijn graphs of k-mers from reads [5, 13, 20]. These methods enable reconstruction of the transcriptome in species for which no reference genomic sequence is available. However, the denovo nature of these methods make it difficult to map assembled transcripts back to known annotated transcripts, and estimation of transcript expression levels in these methods is not straightforward and generally involves a post-process of aligning assembled contigs to a reference genome [5, 13].

Expression level estimation is nontrivial in non-inbred diploid organisms, since each expressed transcript may contain two different sets of alleles, one from the maternal haplotype, and one from the paternal haplotype. In many cases, one parental allele can be preferentially expressed over another, resulting in what is known as allele-specific expression. It is often desirable to identify genes and transcripts exhibiting allele-specific expression, as well as estimate the expression level of the maternal and paternal alleles separately. To our knowledge, current methods that analyze allele-specific expression rely on read alignment [14, 15], which either results in reference bias or requires prior knowledge of genomic variants relative to the reference. Reference bias can be particularly problematic in allele-specific expression analysis, since the parental strain with the genome more similar to the reference genome will often be preferred over the other parental strain.

In the case where RNA-seq samples of a mother-father-child trio are available, we can utilize the RNA-seq data from the parental strains and eliminate the need for prior knowledge of their genomic variants. Here, we examine allele-specific expression in F1 mouse strains, which are hybrids of two distinct homogeneous parental strains. We construct maternal and paternal versions of transcripts using RNA-seq reads from the parental strains and annotated reference transcripts, creating a set of candidate transcripts the F1 strain could express. We then estimate the expression level of each candidate transcript in the F1 strain using a modified lasso regression model. The lasso regression model has been proposed by Li et al. [11] for RNA-seq analysis, but not in the context of estimating allele-specific expression without reference alignment. We choose to use lasso regularization since it reduces parameters to zero, enabling us to effectively eliminate non-expressed isoforms that have significant k-mer overlaps with expressed isoforms. We modify the lasso penalty slightly to prefer assigning higher F1 expression levels in transcripts with k-mers that appear frequently in the parental RNA-seq reads, due to the assumption that most highly expressed genes in the parents should also be highly expressed in the F1 strain.

We test our methods on synthetic and real RNA-seq data from the wild-derived mouse strains CAST/EiJ and PWK/EiJ and their F1 combination CASTxPWK, with CAST/EiJ as the maternal strain and PWK/EiJ as the paternal strain. The CAST/EiJ and PWK/EiJ strains are have well-annotated genomes that differ significantly from the mouse reference strain C57BL/6J [19], making them ideal candidates for studying allele-specific expression and the effects of reference bias.
2 Approach

2.1 Notation

We present the following notation for ease of reference:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>F1 k-mer profile. A vector indicating the number of times each possible k-mer occurs in the F1 sample.</td>
</tr>
<tr>
<td>( y_i )</td>
<td>Number of times the ( i^{th} ) k-mer occurs in the F1 sample.</td>
</tr>
<tr>
<td>( z^M )</td>
<td>Maternal k-mer profile.</td>
</tr>
<tr>
<td>( z^P )</td>
<td>Paternal k-mer profile.</td>
</tr>
<tr>
<td>( X^M )</td>
<td>Set of k-mer profiles of candidate transcripts from ( z^M ).</td>
</tr>
<tr>
<td>( X^P )</td>
<td>Set of k-mer profiles of candidate transcripts from ( z^P ).</td>
</tr>
<tr>
<td>( x_j )</td>
<td>K-mer profile of the ( j^{th} ) candidate transcript, for ( \forall x_j \in X^M \cup X^P ).</td>
</tr>
<tr>
<td>( x_{i,j} )</td>
<td>Number of times the ( i^{th} ) k-mer occurs in the ( j^{th} ) candidate transcript.</td>
</tr>
<tr>
<td>( \theta_j )</td>
<td>Estimated expression level for the ( j^{th} ) candidate transcript.</td>
</tr>
</tbody>
</table>

The k-mer profiles of the F1, maternal, and paternal strains are normalized based on total read counts, so that they may be comparable with each other. We will denote the k-mer profiles of maternal candidate transcripts, \( X^M = \{x^M_1, x^M_2, ..., x^M_r\} \), and the k-mer profiles of paternal candidate transcripts, \( X^P = \{x^P_1, x^P_2, ..., x^P_s\} \), jointly as \( X = X^M \cup X^P \), representing the k-mer profiles of all candidate transcripts. Each candidate transcript k-mer profile is labeled as originating from the maternal k-mer profile, the paternal k-mer profile, or both.

2.2 Regression model

We propose a modified lasso penalized regression model for estimating the abundance of each candidate transcript, with the assumption that the F1’s k-mer profile \( y \) can be expressed as a linear combination of its expressed transcripts \( X = \{x_1, ..., x_j, ..., x_m\} \) multiplied by their relative expression levels \( \theta_j \):

\[
y = \sum_{j=1}^{m} \theta_j x_j.
\]

(1)

To filter out non-expressed transcripts and prevent overfitting, each candidate transcript is penalized by an \( l_1 \)-norm, parameterized by \( \lambda \), the lasso regularization parameter, as well as the inverse of \( w_j \), where

\[
w_j = \text{median} \begin{cases} 
\{z^M_i / x_{i,j}, \forall x_{i,j} > 0\}, & x_j \in X^M \\
\{z^P_i / x_{i,j}, \forall x_{i,j} > 0\}, & x_j \in X^P \\
((z^M_i + z^P_i) / x_{i,j}, \forall x_{i,j} > 0\}, & x_j \in X^P \cap X^M 
\end{cases}
\]

(2)

Therefore, transcripts that are expressed at a high level in the parental samples are more likely to be expressed at a high level in the F1 sample as well. Our objective function then
Figure 1: Our pipeline for estimating allele-specific expression in F1 animals. (a) k-mer profiles are created for the maternal, paternal, and F1 strains, using all available RNA-seq reads from one sample of each strain. Each k-mer is also saved as its reverse compliment, since we do not know the directionality of the read. (b) De Bruijn graphs are created for the maternal and paternal samples. Using annotated reference transcripts and the parental De Bruijn graphs, we select candidate transcripts which incorporate parental alleles from the De Bruijn graphs. (c) The k-mer profile of the F1 sample, \( y \), is then regressed onto the candidate parental transcripts, \( \{x_1^M, x_2^M, ..., x_r^M\} \cup \{x_1^P, x_2^P, ..., x_s^P\} \), and we estimate the expression level \( \theta \) of each candidate transcript.
becomes
\[
\text{argmin}_{\theta} \quad \frac{1}{2} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \theta_j x_{i,j})^2 + \lambda \sum_{j=1}^{m} \frac{\theta_j}{w_j}
\]
subject to \( \theta_j \geq 0, \forall j, \)

with each \( \theta_j \) constrained to be positive since they represent transcript expression levels.

## 3 Methods

### 3.1 Synthetic data

We generated synthetic RNA-seq reads of the CAST/EiJ mouse strain, the PWK/EiJ mouse strain, as well as a CASTxPWK F1 hybrid. We chose these mouse strains because they are well-annotated strains that differ significantly from the reference strain C57BL/6J, compared to other common laboratory strains. Transcript sequences for CAST/EiJ and PWK/EiJ were created by inserting known SNPs and indels from the Wellcome Trust Institute [7] into reference transcripts annotated by the Ensembl Genome Database [2].

Transcripts were separated into fragments with lengths sampled from the Poisson distribution with a mean fragment length of 400 bp. The starting position of each fragment was uniformly sampled across all possible positions. Each transcript had an expression level randomly selected from a uniform distribution between 10 and 3000. We synthesized 100 bp paired-end reads from transcript fragments by incorporating a position-dependent error model based on observed quality scores from real RNA-seq data of the same strains. F1 reads were synthesized using CAST/EiJ and PWK/EiJ transcripts. To model allele-specific expression, the amount of maternal contribution was uniformly sampled between 0 and 1.

### 3.2 Selecting candidate transcripts

We consider two algorithms for selecting candidate transcripts from the De Bruijn graphs of each parental k-mer profile. The first algorithm employs a divide-and-conquer technique to search for all paths between two anchor nodes from each reference transcript, then it selects the best path out of all possible paths, as defined by the Levenshtein distance [9] from each path to the reference transcript. Despite the divide-and-conquer approach, this method is not practical for a large number of transcripts, as the space of paths can grow exponentially, and the computation of the Levenshtein distance is quadratic in time complexity. The second proposed algorithm greedily tries to match the reference transcript to k-mers in the graph, allowing for a maximum number of 5 mismatches within a sliding window of 25 bp. Although efficient, this algorithm cannot accommodate indels or dense SNPs. Algorithm 1 is used for selecting candidate transcripts in synthetic data, while Algorithm 2 should be used for selecting candidate transcripts in real data.

### 3.3 Coordinate descent

To optimize our objective function Eq. (3), we update \( \theta_j \) using coordinate descent:
Algorithm 1 Find candidate transcripts from the created maternal k-mer profile $z^M$

$X^M = \{}$

function findPaths($t$)
    select $i$ nearest to $0$, s.t. $|z^M_i - \text{median}(\forall z^M_s \in t)| < \text{thresh}$
    select $j$ nearest to $\text{len}(t - k + 1)$, s.t. $|z^M_j - \text{median}(\forall z^M_s \in t)| < \text{thresh}$
    find all paths from $z^M_i$ to $z^M_j$
    if number of paths $> \text{thresh}$ then
        $t_1 = t[:\frac{1}{2}\text{len}(t)]$
        $t_2 = t[\frac{1}{2}\text{len}(t):]$
        paths1 = findPaths($t_1$)
        paths2 = findPaths($t_2$)
        return [path1+path2, \forall path1, \forall path2]
    end if
    extend all paths to include k-mers to the left of $z^M_i$ and the right of $z^M_j$
    return paths
end function

for each transcript $t$ do
    paths = findPaths($t$)
    bestpath = path with minimum Levenshtein distance to $t$
    if Levenshtein(bestpath, $t$) < thresh then
        $X^M += \text{kmer profile of bestpath}$
    end if
end for

Algorithm 2 Find candidate transcripts from the maternal k-mer profile $z^M$ - heuristic method

$X^M = \{}$

for each transcript $t$ do
    $t' = \{}$
    for each k-mer $s$ in $t$ do
        if $s \in z^M$ then
            $t' += s$
        else if
            for dothen $q \in \{\text{‘A’}, \text{‘C’}, \text{‘G’}, \text{‘T’}\}$
            if $s[:\text{k-1}] + q \in z^M$ then
                $s := s[:\text{k-1}] + q$
                $t' += s$
            end if
        end if
    end for
    if size($t'$) $> \text{thresh} \times \text{len}(t - k + 1)$ then
        $X^M += \text{k-mer profile of } t'$
    end if
end for
\[ \theta_j = \max \left( \frac{\sum_{i=1}^{n} y_i^{(-j)} x_{i,j} - \lambda w_j}{\sum_{i=1}^{n} x_{i,j}^2}, 0 \right), \text{ where } y_i^{(-j)} = y_i - \sum_{k \neq j} \theta_j x_{i,j}. \] \tag{4}

The lasso regularization parameter \( \lambda \) is chosen via 4-fold cross validation. It is important to note that the value of \( \lambda \) depends on the mean observed values for \( w_j \), so different values of \( \lambda \) should be chosen for each sample.

4 Results

4.1 Synthetic data results

In our synthetic F1 sample, we generated 193 unique transcript sequences from both the maternal and paternal haplotypes, representing 100 reference transcripts. We used algorithm 1 to identify 217 candidate transcript sequences, 177 of which were truly expressed, representing 98 out of 100 truly expressed reference transcripts.

We selected the lasso regularization parameter \( \lambda \) to be \( 5.0 \times 10^7 \) using 4-fold cross validation. We took \( \theta_j = 0 \) to mean transcript \( j \) was not expressed and calculated the sensitivity and specificity of our method in identifying which transcripts were expressed. For the chosen value of \( \lambda \), we found the sensitivity to be 0.81 and the specificity to be 0.68. To allow for comparison of relative expression levels, we normalized both true and predicted expression levels to have a mean value of 1 across all expressed transcripts. The mean absolute error between true and predicted expression levels was 0.2780 for the chosen value of \( \lambda \) True positive rates, false positive rates, and mean absolute error of predicted expression levels for different values of \( \lambda \) are summarized in Fig. (2).

Out of 73 transcripts exhibiting allele-specific expression, as defined by having a maternal contribution outside the range [0.4, 0.6], our model correctly identified 69, achieving a sensitivity of 0.95. Out of the 25 transcripts not exhibiting allele-specific expression, our model correctly identified 18, achieving a specificity of 0.72. Of the 69 true positives, we correctly attributed the parent with the higher expression level in 65 cases.

We chose to compare our results with Trinity [5], since its de novo assembly methods are able to separate maternal and paternal versions of transcripts better than reference alignment-based methods.

We ran Trinity with its default parameters. Transcript sequences were assembled by Trinity, and expression level estimation was done with auxiliary downstream analysis tools. Per Trinity’s downstream analysis guidelines, reads were aligned to the assembled transcript sequences using Bowtie [8], then expression level estimation was performed using RSEM [10].

Trinity assembled 366 transcript sequences. Following their guidelines to eliminate false positives, we retained 314 transcripts representing at least 1% of the per-component expression level. We used a criterion of Levenshtein distance less than 10% of the true transcript length to match annotated transcripts to the de novo transcripts reported by Trinity. With this criterion, only 28 of the true expressed transcripts were present in the set of expressed transcripts found by Trinity. In this set, the mean Levenshtein distance from each true transcript sequence to the Trinity sequences was 0.12% of the true transcript length, with the maximum distance being 2.6% of the true transcript length, suggesting our matching criterion of 10% Levenshtein distance was generous.
Figure 2: True positive rate vs. false positive rate for different values of $\lambda$. Here we only consider transcripts appearing in the set of 217 candidate transcripts on which we ran our regression model. Each point is colored by the mean absolute error between normalized true and estimated expression levels for all transcripts correctly classified as expressed. $\lambda$ was select via 4-fold cross-validation to be $5.0 \times 10^7$. Trinity had a low sensitivity of 0.14, and the mean absolute error for correctly identified transcripts was 0.4664.

Out of the 28 assembled transcripts correctly identified, 24 had nonzero expression levels, making the recall rate 0.14. To compare expression levels determined by the Trinity-Bowtie-RSEM pipeline with true expression levels, we again normalized both sets of expression levels to have means of 1 across all transcripts. Out of the 24 transcript sequences that were correctly identified by Trinity, the mean absolute error between the estimated and the true expression levels was 0.4664 (Fig. 2).

Of the 24 transcripts correctly identified by Trinity, 21 originated from reference transcripts with allele-specific expression. Trinity correctly identified 17 true positives and 1 true negative, achieving a sensitivity of 0.81 and specificity of 0.33. Of the 17 true positives, Trinity attributed the higher expression level to the correct parent in 14 cases.

5 Discussion

We have developed methods to accurately estimate separate expression levels for maternal and paternal versions of transcripts. Our need for such methods arose when we realized that though we have RNA-seq data of many biological trios and wish to analyze allele-specific expression of F1 mouse strains, current methods, both alignment-based and de novo, are not able to directly take advantage of available RNA-seq data from parental strains. Our proposed model is able to exploit the information from the maternal and paternal RNA-seq
reads and built candidate transcripts that accurately reflect the F1 strain’s transcriptome.

Our methods performed well when compared to Trinity, a state-of-the-art de novo assembler for RNA data. We were able to achieve high sensitivity and specificity in both identifying expressed transcripts and detecting allele-specific expression. We were also able to more accurately assign expression levels to both maternal and paternal versions of transcripts. Since we utilize information from the maternal and paternal haplotypes, we are able to clearly distinguish maternal and paternal versions of transcripts in the F1 strain. Trinity cannot distinguish between maternal and paternal versions of transcripts as well, since F1 k-mers are treated as independent once they are constructed from reads, and phasing independent k-mers into maternal and paternal haplotypes can be difficult.

The dimensionality of our data can be large in the k-mer space. In our synthetic data, there were $8 \times 10^5$ possible k-mers appearing in the maternal, paternal, and F1 reads. In our real data of CAST/EiJ, PWK/EiJ, and CASTxPWK mouse liver samples sequenced at 30X coverage in the form of 100 bp paired-end reads, the number of possible k-mers grows to $5 \times 10^8$. We can significantly decrease our k-mer space by merging k-mers into contigs when (k-1)-mer nodes in the associated De Bruijn graph have an in degree and out degree of one.

We can also end up with many candidate transcripts, generating a large feature space. Although our synthetic data set contained only a small number of transcripts, in real data, the number of candidate transcripts grow by orders of magnitude to tens of thousands. In these cases, it would be advantageous to move toward a more computational efficient algorithm such as least angle regression [4] or random coordinate descent [12].

Currently, our methods cannot assemble novel transcript sequences, since our candidate transcripts are generated from annotated reference transcripts. However, we can model the k-mer profiles of all novel transcripts as the residual of our linear regression, and de novo assembly of the residual k-mers using methods similar to Trinity would then allow us to generate sequences of novel transcripts. Our current methods also do not exploit paired-end information since all k-mers are considered independent, but we plan on developing methods to incorporate paired-end information.

Although our methods are still in the early stages of development and testing, they show great promise for estimating allele-specific expression levels in RNA-seq data involving biological trios, which is often easily available in model organisms with inbred parents. We hope to improve the efficiency of our methods and test them on real data in the very near future.

References


