Estimation of alternative splicing isoform frequencies from RNA-Seq data*

Marius Nicolae[1], Serghei Mangul[2], Ion Măndoiu[1], and Alex Zelikovsky[2]

[1] Computer Science & Engineering Department, University of Connecticut
371 Fairfield Way, Storrs, CT 06269
{man09004,ion}@engr.uconn.edu
[2] Computer Science Department, Georgia State University
University Plaza, Atlanta, Georgia 30303
{serghei,alexz}@cs.gsu.edu

Abstract. In this paper we present a novel expectation-maximization algorithm for inference of alternative splicing isoform frequencies from high-throughput transcriptome sequencing (RNA-Seq) data. Our algorithm exploits disambiguation information provided by the distribution of insert sizes generated during sequencing library preparation, and takes advantage of base quality scores, strand and read pairing information if available. Empirical experiments on synthetic datasets show that the algorithm significantly outperforms existing methods of isoform and gene expression level estimation from RNA-Seq data. The Java implementation of IsoEM is available at http://dna.engr.uconn.edu/software/IsoEM/.

1 Introduction

Ubiquitous regulatory mechanisms such as the use of alternative transcription start and polyadenylation sites, alternative splicing, and RNA editing result in multiple messenger RNA (mRNA) isoforms being generated from a single genomic locus. Most prevalently, alternative splicing is estimated to take place for over 90% of the multi-exon human genes [19], and thought to play critical roles in early stages of development and normal function of cells from diverse tissue types. Thus, the ability to reconstruct full length isoform sequences and accurately estimate their frequencies is critical for understanding gene functions and transcription regulation mechanisms.

Three key interrelated computational problems arise in the context of transcriptome analysis: *gene expression level estimation (GE)*, *isoform discovery (ID)*, and *isoform expression level estimation (IE)*. Targeted GE has long been a staple of genetic studies, and the completion of the human genome has enabled genome-wide GE performed using expression microarrays. Since expression microarrays have limited capability of detecting alternative splicing events, specialized splicing arrays have been developed to interrogate genome-wide both (annotated) exons and exon-exon junctions. However, despite sophisticated deconvolution algorithms [1, 15], the fragmentary information provided by splicing

* Work supported in part by NSF awards IIS-0546457, IIS-0916401, and IIS-0916948.

arrays is often insufficient for unambiguous identification of transcribed isoforms [6, 9]. High-throughput transcriptome sequencing, commonly referred to as RNA-Seq, is quickly replacing microarrays as the technology of choice for performing GE due to the far wider dynamic range and more accurate quantitation capabilities [20]. Unfortunately, most RNA-Seq studies to date either ignore alternative splicing or, similar to splicing array studies, restrict themselves to surveying the presence/expression levels of exons and exon-exon junctions. The main difficulty lies in the fact that current technologies used to perform RNA-Seq generate short reads (from few tens to hundreds of bases), many of which cannot be unambiguously assigned to individual isoforms.

1.1 Related Work

RNA-Seq analyses typically start by mapping sequencing reads onto the reference genome, transcript libraries, exon-exon junction libraries, or combinations thereof. Early RNA-Seq studies have recognized that short read lengths result in a significant percentage of so called *multireads*, i.e., reads that map equally well at multiple locations in the genome. A simple (and still commonly used) approach is to discard multireads, and estimate expression levels using only the so called *unique* reads. Mortazavi et al. [12] proposed a multiread "rescue" method whereby initial gene expression levels are estimated from unique reads and used to fractionally allocate multireads, with final expression levels re-estimated from total counts obtained after multiread allocation. An expectation-maximization (EM) algorithm that extends this scheme by repeatedly alternating between fractional read allocation and re-estimation of gene expression levels was recently proposed in [13].

A number of recent works have addressed the IE problem, namely isoform expression level estimation from RNA-Seq reads. Under a simplified "exact information" model, [9] showed that neither single nor paired read RNA-Seq data can theoretically guarantee unambiguous inference of isoform expression levels, but paired reads may be sufficient to deconvolute expression levels for the majority of annotated isoforms. The key challenge in IE is accurate assignment of ambiguous reads to isoforms. Compared to the GE context, read ambiguity is much more significant, since it affects not only multireads, but also reads that map at a unique genome location expressed in multiple isoforms. To overcome this difficulty, [8] proposed a Poisson model of single-read RNA-Seq data explicitly modeling isoform frequencies. Under this model, maximum likelihood estimates are obtained by solving a convex optimization problem, and uncertainty of estimates are obtained by importance sampling from the posterior distribution. Li et al. [11] introduced an expectation-maximization (EM) algorithm similar to that of [13] but apply it to isoforms instead of genes. Unlike the method of [8], which estimates isoform frequencies only from reads that map to a unique location in the genome, the algorithm of [11] incorporates multireads as well. The IE problem for single reads is also tackled in [14], who propose an EM algorithm for inferring isoform expression levels from read coverage of exons (reads spanning exon junctions are ignored).

The related isoform discovery (ID) problem has also received much interest in the literature. De novo transcriptome assembly algorithms have been proposed in [2, 7]. Very recently, [4] and [18] proposed methods for simultaneously solving ID and IE based on paired RNA-Seq reads. Assuming known genomic positions for alternative transcription start and polyadenylation sites as well as exon boundaries, [4] formulate IE as a convex quadratic program (QP) that can be efficiently solved for each gene locus after discarding multireads. ID is solved by iteratively generating isoform candidates from the splicing graph derived from annotations and reads spanning exon-exon junctions. The process is continued until the p-value of the objective value of the QP corresponding to the set of selected isoforms, assumed to follow a χ^2 distribution, exceeds an empirically selected threshold of 5%. However, pair read information is not directly used in isoform frequency estimation, contributing only as secondary data to filter out false positives in the process of isoform selection. Trapnell et al. [18] also describe a method, referred to as Cufflinks, for simultaneously solving ID and IE. Unlike the method of [4], Cufflinks requires no genome annotations (but can use them if available). After performing spliced alignment of (paired) reads onto the genome using TopHat [17], Cufflinks constructs a read overlap graph and generates candidate isoforms by finding a minimal size path cover via a reduction to maximum matching in a weighted bipartite graph. Reads that match equally well multiple locations in the genome are fractionally allocated to these locations, and estimation is then performed independently at different transcriptional loci, using an extension to paired reads of the methods in [8].

1.2 Our Contributions

In this paper we focus on the IE problem, namely estimating isoform expression levels (interchangeably referred to as frequencies) from RNA-Seq reads, under the assumption that a complete list of candidate isoforms is available. Projects such as [3] and [16] have already assembled large libraries of full-length cDNA sequences for humans and other model organisms, and the coverage of these libraries is expected to continue to increase rapidly. Although an incomplete isoform library may lead to estimation biases [18], statistical tests such as the one in [4] can be used to detect the presence of isoforms not represented in the library. Inferring expression at isoform level provides information for finer-resolution biological studies, and also leads to more accurate estimates of expression at the gene level by allowing rigorous length normalization. Indeed, as shown in Section 3, genome-wide gene expression level estimates derived from isoform level estimates are significantly more accurate than those obtained directly from RNA-Seq data using isoform-oblivious GE methods such as the widely used counting of unique reads, the rescue method of [12], or the EM algorithm of [13].

Our main contribution is a novel expectation-maximization algorithm for isoform frequency estimation from (any mixture of) single and paired RNA-Seq reads. A key feature of our algorithm, referred to as IsoEM, is that it exploits the information provided by the distribution of insert sizes, which is tightly controlled during sequencing library preparation under current RNA-Seq protocols.

The recently published [18] is the only other work we are aware of that exploits this information (that is not captured by the "exact" information models of [6, 9]) in conjunction with paired read data. We show that modeling insert sizes is also highly beneficial in conjunction with single RNA-Seq reads. Insert sizes contribute to increased estimation accuracy in two different ways. On one hand, insert sizes help disambiguating the isoform of origin for the reads. In IsoEM, insert lengths are combined with base quality scores, and, if available, read pairing and strand information to probabilistically allocate reads to isoforms during the expectation step of the algorithm. As in [11], the genomic locations of multireads are also resolved probabilistically in this step, further contributing to overall accuracy compared to methods that ignore or fractionally pre-allocate multireads. On the other hand, insert size distribution is used to accurately adjust isoform lengths during frequency re-estimation in the M step of the IsoEM algorithm; an equivalent adjustment was independently employed in [18].

We also present preliminary experimental results on synthetic datasets generated with various sequencing parameters and distribution assumptions. The results show that IsoEM algorithm significantly outperforms existing methods of isoform and gene expression level estimation from RNA-Seq data. Furthermore, we empirically evaluate the effect of sequencing parameters such as read length, read pairing, and strand information on estimation accuracy. Our experiments confirm the finding of [11] that, for a fixed total number of sequenced bases, longer reads do not necessarily lead to better accuracy for estimation of isoform and gene expression levels.

2 Methods

2.1 Read Mapping

As with most RNA-Seq analyses, the first step of IsoEM is to map the reads. Our approach is to map them onto the library of known isoforms using any one of the many available aligners (we used Bowtie [10] with default parameters in our experiments). An alternative strategy is to map the reads onto the genome using a spliced alignment tool such as TopHat [17], as done in [18]. However, preliminary experiments with TopHat resulted in fewer mapped reads and increased mapping uncertainty. Since further increases in read length coupled with improvements in spliced alignment algorithms could make genome mapping more attractive in the future, we made our IsoEM implementation compatible with both mapping approaches by converting read alignments to genome coordinates and performing all operations in genome space.

2.2 Finding Read-Isoform Compatibilities

The candidate set of isoforms for each read is obtained by putting together all genome coordinates for reads and isoforms, sorting them and using a line sweep technique to detect read-isoform compatibilities. During the line sweep, reads

are grouped into equivalence classes defined by their isoform compatibility sets; this speeds up the E-steps of the IsoEM algorithm by allowing the processing of an entire read class at once.

Some of the reads match multiple positions in the genome, which we refer to as *alignments* (for paired end reads, an alignment consists of the positions where the two reads in the pair align with the genome). Each alignment a can in turn be compatible with multiple isoforms that overlap at that position of the genome. During the line sweep, we compute the relative "weight" of assigning a given read/pair r to isoform j as $w_{r,j} = \sum_a Q_a F_a O_a$, where the sum is over all alignments of r compatible with j, and the factors of the summed products are defined as follows.

- Q_a represents the probability of observing the read from the genome locations described by the alignment. This is computed from the base quality scores as $Q_a = \prod_{k=1}^{|r|}[(1-\varepsilon_k)M_{a_k} + \varepsilon_k(1-M_{a_k})]$, where $M_{a_k} = 1$ if position k of alignment a matches the genome and 0 otherwise, while ε_k denotes the error probability of kth base of r.
- F_a represents the probability of the fragment length needed to produce alignment a from isoform j. For paired end reads, the length of the fragment can be inferred from the positions of the two reads. For single reads, we can only estimate a maximum fragment length: if the alignment is on the same strand as the isoform, we use the distance from the start of the alignment to the end of the isoform, otherwise we use the distance from the end of the alignment to the start of the isoform.
- O_a is 1 if alignment a of r is consistent with the orientation of isoform j, and 0 otherwise. Consistency between the orientations of r and j depends on whether or not the library preparation protocol preserves the strand information. For single reads $O_a = 1$ when reads are generated from fragment ends randomly or, for directional RNA-Seq, when they match the known isoform orientation. For pairs, $O_a = 1$ if the two reads come from different strands, point to each other, and, in the case of directional RNA-Seq, the orientation of first read matches the known isoform orientation.

Weigths $w_{r,j}$ can be further adjusted to account for biases introduced by sequencing library preparation or the sequencing process once a model of this biases, such as the one in [5], is available.

2.3 The IsoEM Algorithm

The IsoEM algorithm starts with the set of N known isoforms. For each isoform we denote by $l(j)$ its length and by $f(j)$ its (unknown) frequency. If we ignore library preparation and amplification biases, the probability that a read is sampled from isoform j is proportional with $(l(j) - \mu + 1)f(j)$ where μ is the mean fragment length from the sample preparation. To see why this is true, we write the expected number of reads coming from an isoform by summing over all possible fragment lengths. For each fragment length k we expect the number

Algorithm 1 IsoEM algorithm

assign random values to all $f(i)$
while not converged **do**
  initialize all $n(j)$ to 0
  for each read r **do**
    sum $= \sum_{j:w_{r,j}>0} w_{r,j} f(j)$
    for each isoform j with $w_{r,j} > 0$ **do**
      $n(j) += w_{r,j} f(j)/\text{sum}$
    end for
  end for
  $s = \sum_j n(j)/(l(j) - \mu + 1)$
  for each isoform j **do**
    $f(j) = \frac{n(j)/(l(j)-\mu+1)}{s}$
  end for
end while

of fragments of that length to be proportional to the number of valid starting positions for a fragment of that length in the isoform. If $p(k)$ denotes the probability of a fragment of length k and $n(j)$ denotes the number of reads coming from isoform j then $E[n(j)] \propto \sum_k p(k)(l(j) - k + 1) = l(j) - \mu + 1$. Thus, if the isoform of origin is known for each read, the maximum likelihood estimator for $f(j)$ is given by $c(j)/(c(1) + \ldots + c(N))$, where $c(j) = n(j)/(l(j) - \mu + 1)$ denotes the length-normalized fragment coverage.

Unfortunately, some reads match multiple isoforms, so their isoform of origin cannot be established unambiguously. The IsoEM algorithm (see Algorithm 1) overcomes this difficulty by simultaneously estimating the frequencies and imputing the missing read origin within an iterative framework. After initializing frequencies $f(j)$ at random, the algorithm repeatedly performs the next two steps until convergence:

- E-step: Compute the expected number $n(j)$ of reads that come from isoform j under the assumption that isoform frequencies $f(j)$ are correct, based on weights $w_{r,j}$
- M-step: For each j, set the new value of $f(j)$ to $c(j)/(c(1) + \ldots + c(N))$, where normalized coverages $c(j)$ are based on expected counts computed in previous step

3 Experimental Results

3.1 Simulation Setup

We tested IsoEM on simulated human RNA-Seq data. The human genome sequence (hg18, NCBI build 36) was downloaded from UCSC together with the coordinates of the isoforms in the KnownGenes table. Genes were defined as

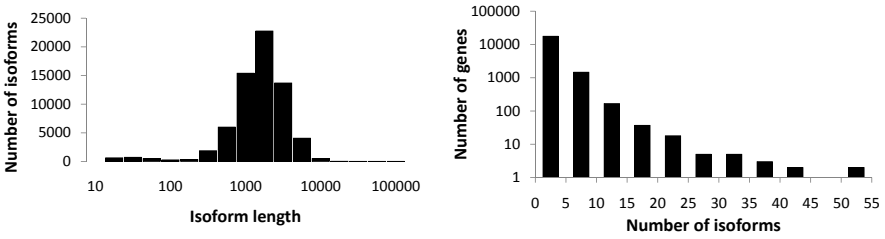

Fig. 1. Distribution of isoform lengths (left panel) and gene cluster sizes (right panel) for the UCSC KnownGenes dataset.

clusters of known isoforms defined by the GNFAtlas2 table. The dataset contains a total of 66803 isoforms pertaining to 19372 genes. The isoform length distribution and the number of isoforms per genes are shown in Figure 1.

Single and paired-end reads were randomly generated by sampling fragments from the known isoforms. Each isoform was assigned a *true frequency* based on the abundance reported for the corresponding gene in the first human tissue of the GNFAtlas2 table, and a probability distribution over the isoforms inside a gene cluster. Thus, the true frequency of isoform j is $a(g)p(j)$, where $a(g)$ is the abundance of the gene g for which j is an isoform and $p(j)$ is the probability of isoform j among all the isoforms of g. We simulated datasets with uniform and geometric ($p = 0.5$) distributions for the isoforms of each gene. Fragment lengths were simulated from a normal probability distribution with mean 250 and standard deviation 25. We simulated between 1 and 60 million single and paired reads of lengths ranging from 25 to 100 base pairs, with or without strand information.

We compared IsoEM to several existing IE and GE algorithms. For IE we included in the comparison the isoform analogs of the Uniq and Rescue methods used for GE [12], an improved version of Uniq (UniqLN) that estimates isoform frequencies from unique read counts but normalizes them using adjusted isoform lengths that exclude ambiguous positions, the Cufflinks algorithm of [18], and the RSEM algorithm of [11]. For the GE problem, the comparison included the Uniq and Rescue methods, our implementation of the EM algorithm described in [13] (GeneEM), and estimates obtained by summing isoform expression levels inferred by Cufflinks, RSEM, and IsoEM. All methods except Cufflinks use alignments obtained by mapping reads onto the library of isoforms with Bowtie [10] and then converting them to genome coordinates. As suggested in [18], Cufflinks uses alignments obtained by mapping the reads onto the genome with TopHat [17], which was provided with a complete set of annotated junctions.

Frequency estimation accuracy was assessed using the coefficient of determination, r^2, along with the *error fraction (EF)* and *median percent error (MPE)* measures used in [11]. However, accuracy was computed against true frequencies, not against estimates derived from true counts as in [11]. If $\hat{f}_i$ is the frequency

Isoform Expression			Gene Expression		
Algorithm	Uniform	Geometric	Algorithm	Uniform	Geometric
Uniq	0.466	0.447	Uniq	0.579	0.586
Rescue	0.693	0.675	Rescue	0.724	0.724
UniqLN	0.856	0.838	GeneEM	0.636	0.637
Cufflinks	0.661	0.618	Cufflinks	0.778	0.757
RSEM	0.919	0.911	RSEM	0.939	0.934
IsoEM	**0.979**	**0.964**	IsoEM	**0.988**	**0.978**

Table 1. r^2 for isoform and gene expression levels inferred from 30M reads of length 25 from reads simulated assuming uniform, respectively geometric expression of gene isoforms.

estimate for an isoform with true frequency f_i, the *relative error* is defined as $|\hat{f}_i - f_i|/f_i$ if $f_i \neq 0$, 0 if $\hat{f}_i = f_i = 0$, and ∞ if $\hat{f}_i > f_i = 0$. The error fraction with threshold τ, denoted EF_τ is defined as the percentage of isoforms with relative error greater or equal to τ. The median percent error, denoted MPE, is defined as the threshold τ for which $EF_\tau = 50\%$.

3.2 Comparison Between Methods

Table 1 gives r^2 values for isoform, respectively gene expression levels inferred from 30M reads of length 25, simulated assuming both uniform and geometric isoform expression. IsoEM significantly outperforms the other methods, achieving an r^2 values of over .96 for all datasets. For all methods the accuracy difference between datasets generated assuming uniform and geometric distribution of isoform expression levels is small, with the latter one typically having a slightly worse accuracy. Thus, in the interest of space we present remaining results only for datasets generated using geometric isoform expression.

For a more detailed view of the relative performance of compared IE and GE algorithms, Figure 2 gives the error fraction at different thresholds ranging between 0 and 1. The variety of methods included in the comparison allows us to tease out the contribution of various algorithmic ideas to overall estimation accuracy. The importance of rigorous length normalization is demonstrated by the IE accuracy gain of UniqLN over Uniq – clearly larger than that achieved by ambiguous read reallocation as implemented in the IE version of Rescue. Proper length normalization is also the main reason for the accuracy gain of isoform-aware GE methods (Cufflinks, RSEM, and IsoEM) over isoform oblivious GE methods. Similarly, the importance of modeling insert sizes even for single read data is underscored by the IE and GE accuracy gains of IsoEM over RSEM.

For yet another view, Tables 2 and 3 report the MSE and $EF_{.15}$ measures for isoform, respectively gene expression levels inferred from 30M reads of length 25, computed over groups of isoforms with various expression levels. IsoEM consistently outperforms the other IE and GE methods at all expression levels except for isoforms with zero true frequency, where it is dominated by the more conservative Uniq algorithm and its UniqLN variant.

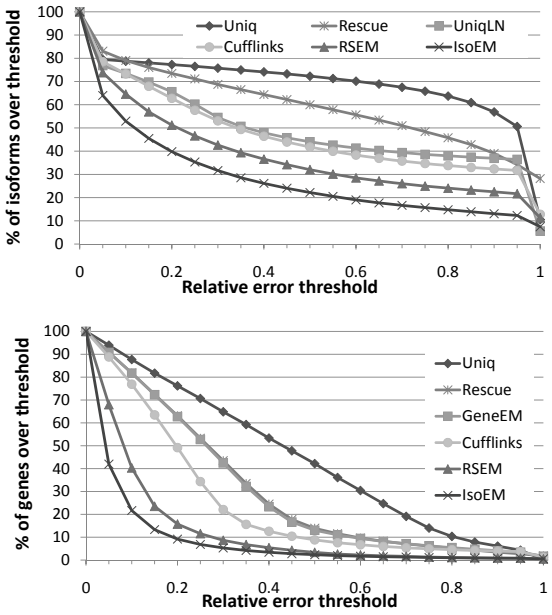

Fig. 2. Error fraction at different thresholds for isoform (top panel) and gene (bottom panel) expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.

3.3 Influence of Sequencing Parameters

Although high-throughput technologies allow users to make tradeoffs between read length and the number of generated reads, very little has been done to determine optimal parameters even for common applications such as RNA-Seq. The intuition that longer reads are better certainly holds true for many applications such as de novo assembly. Surprisingly, [11] found that *shorter* reads are better for IE when the total number of sequenced bases is fixed. Figure 3 plots IE estimation accuracy for reads of length between 25 and 100 when the total amount of sequence data is kept constant at 750M bases. Our results confirm the finding of [11], although the optimal read length is somewhat sensitive to the accuracy measure used and to the availability of pairing information. While 25bp reads optimize the MPE measure regardless of the availability of paired reads, the read length that maximizes r^2 is 36 for paired reads and 50 for single reads. While more experiments are needed to determine how the optimum length depends on the amount of sequence data and transcriptome complexity, this does suggest that, for isoform and gene expression estimation accuracy, increasing the number of reads may be more useful than increasing read length beyond a certain limit.

Expression range		0	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
# isoforms		13290	10024	23882	18359	1182	66	66803
MPE	Uniq	**0.0**	**100.0**	98.4	97.1	98.5	96.6	95.4
	Rescue	**0.0**	294.7	75.5	49.2	30.4	28.3	71.9
	UniqLN	**0.0**	**100.0**	80.8	30.3	26.4	24.8	36.0
	Cufflinks	**0.0**	**100.0**	49.7	25.5	27.2	44.6	34.1
	RSEM	**0.0**	**100.0**	31.9	13.5	11.4	13.0	21.2
	IsoEM	**0.0**	**100.0**	**22.7**	**7.3**	**3.5**	**2.5**	**11.8**
$EF_{.15}$	Uniq	**0.2**	98.4	97.2	96.9	97.0	95.5	78.0
	Rescue	48.4	95.5	86.2	73.1	61.5	56.1	76.0
	UniqLN	**0.2**	97.2	86.2	82.8	83.3	77.3	69.8
	Cufflinks	17.6	96.4	81.3	71.0	74.7	80.3	67.9
	RSEM	19.9	93.7	71.1	46.4	39.8	47.0	56.9
	IsoEM	5.1	**91.2**	**62.8**	**29.3**	**15.8**	**7.6**	**45.5**

Table 2. Median percent error (MPE) and 15% error fraction ($EF_{.15}$) for isoform expression levels inferred from 30M reads of length 25.

The top panel of Figure 4 shows, for reads of length 75, the effects of paired reads and strand information on estimation accuracy as measured by r^2. Not surprisingly, for a fixed number of reads, paired reads yield better accuracy than single reads. Also not very surprisingly, adding strand information to paired sequencing yields no benefits to genome-wide IE accuracy (although it may be helpful, e.g., in identification of novel transcripts). Quite surprisingly, performing strand-specific single read sequencing is actually *detrimental* to IsoEM IE (and hence GE) accuracy under the simulated scenario, most likely due to the reduction in sampled transcript length.

As shown in the bottom panel of Figure 4, the runtime of our Java implementation of IsoEM scales roughly linearly with the number of *fragments*, and is largely insensitive to the type of sequencing data (single or paired reads, directional or non-directional). IsoEM was tested on a DELL PowerEdge R900 server with 4 Six Core E7450Xeon Processors at 2.4Ghz (64 bits) and 128Gb of internal memory. None of the datasets require more than 16GB of memory to complete, however, increasing the amount of memory made available to the Java virtual machine significantly decreases runtime by reducing the time needed for garbage collection. The runtimes in Figure 4 were obtained by allowing IsoEM to use up to 32GB of memory, in which case none of the datasets took more than 3 minutes to solve.

4 Conclusions and Ongoing Work

In this paper we have introduced an expectation-maximization algorithm for isoform frequency estimation assuming a known set of isoforms. Our algorithm, called IsoEM, explicitly models base quality scores, insert size distribution, strand and read pairing information. Experiments on synthetic data sets generated using two different assumptions on the isoform distribution show that

Expression range		$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
# genes		120	5610	11907	1632	102	19372
MPE	Uniq	37.4	43.6	42.7	43.0	48.2	43.0
	Rescue	32.8	28.7	26.0	25.1	28.8	26.7
	GeneEM	30.6	28.2	25.7	25.1	28.0	26.3
	Cufflinks	33.0	21.1	19.0	20.2	40.2	19.7
	RSEM	23.6	11.0	7.2	7.9	11.4	8.1
	IsoEM	**18.3**	**8.4**	**3.3**	**2.2**	**2.1**	**4.0**
$EF_{.15}$	Uniq	77.5	82.4	81.7	79.7	82.4	81.7
	Rescue	74.2	74.0	71.6	72.8	76.5	72.4
	GeneEM	72.5	73.8	71.5	73.0	74.5	72.3
	Cufflinks	73.3	64.7	62.3	66.2	82.3	63.5
	RSEM	64.2	37.3	17.4	16.3	41.2	23.5
	IsoEM	**57.5**	**28.3**	**6.8**	**6.5**	**4.9**	**13.3**

Table 3. Median percent error (MPE) and 15% error fraction ($EF_{.15}$) for gene expression levels inferred from 30M reads of length 25.

IsoEM consistently outperforms existing algorithms for isoform and gene expression level estimation with respect to a variety of quality metrics.

The open source Java implementation of IsoEM is freely available for download at http://dna.engr.uconn.edu/software/IsoEM/. In ongoing work we are extending IsoEM to perform allelic specific isoform expression and exploring integration of isoform frequency estimation with identification of novel transcripts using the iterative refinement framework proposed in [4].

References

1. M. Anton, D. Gorostiaga, E. Guruceaga, V. Segura, P. Carmona-Saez, A. Pascual-Montano, R. Pio, L. Montuenga, and A. Rubio. SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays. *Genome Biology*, 9(2):R46, 2008.
2. I. Birol, S.D. Jackman, C.B. Nielsen, J.Q. Qian, R. Varhol, G. Stazyk, R.D. Morin, Y. Zhao, M. Hirst, J.E. Schein, D.E. Horsman, J.M. Connors, R.D. Gascoyne, M.A. Marra, and S.J.M. Jones. De novo transcriptome assembly with ABySS. *Bioinformatics*, 25(21):2872–2877, 2009.
3. P. Carninci *et al.* The Transcriptional Landscape of the Mammalian Genome. *Science*, 309(5740):1559–1563, 2005.
4. J. Feng, W. Li, and T. Jiang. Inference of isoforms from short sequence reads. In *Proc. RECOMB*, pages 138–157, 2010.
5. K.D. Hansen, S.E. Brenner, and S. Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucl. Acids Res.*, page gkq224, 2010 (advance access).
6. D. Hiller, H. Jiang, W. Xu, and W.H. Wong. Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics*, 25(23):3056–3059, 2009.
7. B. Jackson, P. Schnable, and S. Aluru. Parallel short sequence assembly of transcriptomes. *BMC Bioinformatics*, 10(Suppl 1):S14+, 2009.

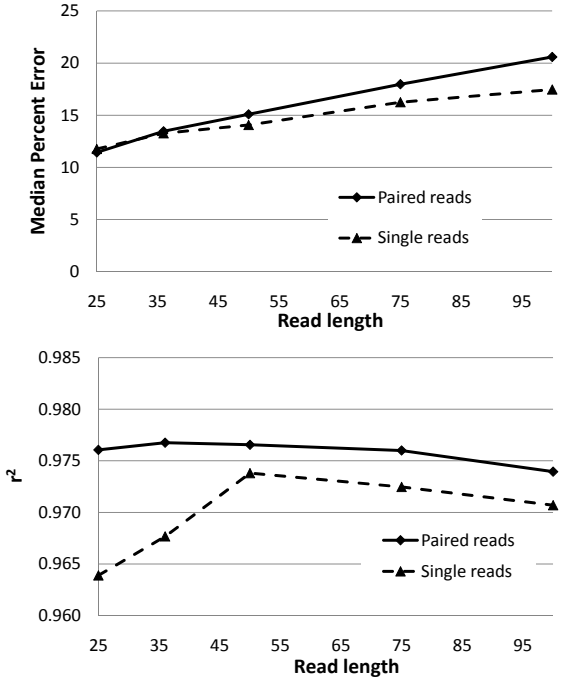

Fig. 3. IsoEM MPE (top panel) and r^2 values (bottom panel) for 750Mb of data generated using single and paired-end sequencing with read length between 25 and 100.

8. H. Jiang and W.H. Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009.
9. V. Lacroix, M. Sammeth, R. Guigo, and A. Bergeron. Exact transcriptome reconstruction from short sequence reads. In *Proc. WABI*, pages 50–63, 2008.
10. B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.
11. B. Li, V. Ruotti, R.M. Stewart, J.A. Thomson, and C.N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.
12. A. Mortazavi, B.A.A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 2008.
13. B. Paşaniuc, N. Zaitlen, and E. Halperin. Accurate estimation of expression levels of homologous genes in RNA-seq experiments. In *Proc. RECOMB*, pages 397–409, 2010.
14. H. Richard, Marcel H. Schulz, M. Sultan, A. Nurnberger, S. Schrinner, D. Balzereit, E. Dagand, A. Rasche, H. Lehrach, M. Vingron, S.A. Haas, and M.-L. Yaspo. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucl. Acids Res.*, 38(10):e112+, 2010.

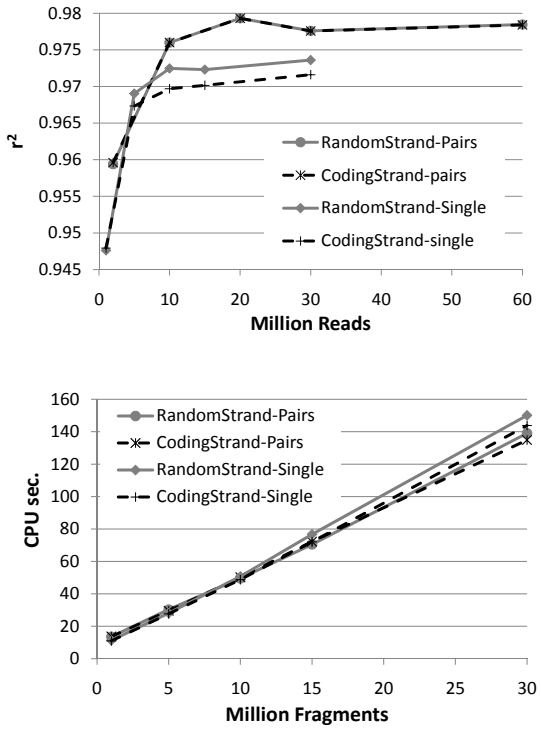

Fig. 4. IsoEM r^2 (top panel) and CPU time (bottom panel) for 1-60 million single/paired reads of length 75, with or without strand information.

15. Y. She, E. Hubbell, and H. Wang. Resolving deconvolution ambiguity in gene alternative splicing. *BMC Bioinformatics*, 10(1):237, 2009.
16. G. Temple *et al.* The completion of the Mammalian Gene Collection (MGC). *Genome Research*, 19(12):2324–2333, 2009.
17. C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.
18. C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010.
19. E.T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, and C.B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.
20. Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.