Accurate Estimation of Isoform and Gene Expression Levels from Next Generation Sequencing Data

Marius Nicolae

Dipl.-Ing. Computer Science, University POLITEHNICA of Bucharest, Romania, 2009

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

at the

University of Connecticut

2011

APPROVAL PAGE

Master of Science Thesis

Accurate Estimation of Isoform and Gene Expression Levels from Next Generation Sequencing Data

Presented by

Marius Nicolae, Dipl.-Ing.

Major Advisor _____
Dr. Ion Măndoiu

Associate Advisor _____
Dr. Sanguthevar Rajasekaran

Associate Advisor _____
Dr. Yufeng Wu

University of Connecticut

2011

ACKNOWLEDGEMENTS

I would like to thank my major advisor Dr. Ion Măndoiu and collaborators Dr. Alexander Zelikovsky and Serghei Mangul for their direct contributions to this thesis. Special acknowledgements go to professor Doru Popescu-Anastasiu, my high-school Computer Science teacher, who was a great inspiration for me and played a key role in me getting to UConn. I would also like to thank my associate advisors, professors, colleagues, friends and family for their support.

TABLE OF CONTENTS

Introduction

Massively parallel transcriptome sequencing is quickly replacing microarrays as the technology of choice for performing gene expression profiling due to its wider dynamic range and digital quantitation capabilities. However, accurate estimation of expression levels from sequencing data remains challenging due to the short read length delivered by current sequencing technologies and still poorly understood protocol- and technology-specific biases. To date, two main transcriptome sequencing protocols have been proposed in the literature. The most commonly used one, referred to as RNA-Seq, generates short (single or paired) sequencing tags from the ends of randomly generated cDNA fragments. An alternative protocol, referred to as 3'-tag Digital Gene Expression (DGE), or high-throughput sequencing based Serial Analysis of Gene Expression (SAGE-Seq), generates single cDNA tags using an assay including as main steps transcript capture and cDNA synthesis using oligo(dT) beads, cDNA cleavage with an anchoring restriction enzyme, and release of cDNA tags using a tagging restriction enzyme whose recognition site is ligated upstream of the recognition site of the anchoring enzyme.

In this thesis we present two novel expectation-maximization algorithms for

inference of isoform- and/or gene-specific expression levels from RNA-Seq and DGE data and a comparison of estimation performance of the two transcriptome sequencing protocols.

The first algorithm, IsoEM [1, 3], works on RNA-Seq data and is based on disambiguating of 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 when available. Empirical experiments on both synthetic and real RNA-Seq datasets show that IsoEM has scalable running time and outperforms existing methods of isoform and gene expression level estimation. Simulation experiments confirm previous findings that, for a fixed sequencing cost, using reads longer than 25-36 bases does not necessarily lead to better accuracy for estimating expression levels of annotated isoforms and genes.

The second chapter introduces a rigorous statistical model of DGE data and a novel expectation-maximization algorithm, DGE-EM [2], for inference of gene and isoform expression levels from DGE tags. Unlike previous methods, our algorithm takes into account alternative splicing isoforms and tags that map at multiple locations in the genome, and corrects for incomplete digestion and sequencing errors. Experimental results show that DGE-EM outperforms methods based on unique tag counting on a multi-library DGE dataset consisting of 20bp tags generated from two commercially available reference RNA samples that have been well-characterized by quantitative real time PCR as

part of the MicroArray Quality Control Consortium (MAQC).

We also take advantage of the availability of RNA-Seq data generated from the same MAQC samples to directly compare estimation performance of the two transcriptome sequencing protocols. While RNA-Seq is clearly more powerful than DGE at detecting alternative splicing and novel transcripts such as fused genes, previous studies have suggested that for gene expression profiling DGE may yield accuracy comparable to that of RNA-Seq at a fraction of the cost [38]. We find that the two protocols achieve similar cost-normalized accuracy on the MAQC samples when using state-of-the-art estimation methods. However, the current protocol versions are unlikely to be optimal. Indeed, the results of a comprehensive simulation study assessing the effect of various experimental parameters suggest that further improvements in DGE accuracy could be achieved by using anchoring enzymes with degenerate recognition sites and using partial digest of cDNA with the anchoring enzyme during library preparation.

Chapter 1

Estimation of alternative splicing isoform frequencies from RNA-Seq data

1.1 Background

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 across diverse cell types [33], with as much as 68% of multi-exon genes expressing multiple isoforms in a clonal cell line of colorectal cancer origin [11]. Not surprisingly, the ability to reconstruct full length isoform sequences and accurately estimate their expression levels is widely believed to be critical for unraveling gene functions and transcription regulation mechanisms [25].

Three key interrelated computational problems arise in the context of transcriptome analysis: *gene expression level estimation (GE)*, *isoform expression level estimation (IE)*, and *novel isoform discovery (ID)*. Targeted GE using methods such as quantitative PCR has long been a staple of genetic studies. The com-

pletion of the human genome has been a key enabler for genome-wide GE performed using expression microarrays. Since expression microarrays have limited capability of detecting alternative splicing events, specialized splicing arrays have been developed for genome-wide interrogation of both annotated exons and exon-exon junctions. However, despite sophisticated deconvolution algorithms [4, 28], the fragmentary information provided by splicing arrays is typically insufficient for unambiguous identification of full-length transcripts [14, 18]. Massively parallel whole transcriptome sequencing, commonly referred to as RNA-Seq, is quickly replacing microarrays as the technology of choice for performing GE due to their wider dynamic range and digital quantitation capabilities [34]. Unfortunately, most RNA-Seq studies to date still ignore alternative splicing or, similar to splicing array studies, restrict themselves to surveying the expression levels of exons and exon-exon junctions. The main difficulty in inferring expression levels for full-length isoforms lies in the fact that current sequencing technologies generate short reads (from few tens to hundreds of bases), many of which cannot be unambiguously assigned to individual isoforms.

1.1.1 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 limited 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. [22] 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 obtained by re-estimation based on 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 [24].

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, [18] showed that neither single nor paired read RNA-Seq data can theoretically guarantee unambiguous inference of isoform expression levels, although 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. Estimating isoform expression levels based solely on unambiguous reads, as suggested, e.g., in [11], results in splicing-dependent biases similar to the transcript-length bias noted in [23], further complicating the design of

unbiased differential expression tests based on RNA-Seq data. To overcome this difficulty, [17] proposed a Poisson model of single-read RNA-Seq data explicitly modeling isoform frequencies. Under their model, maximum likelihood estimates are obtained by solving a convex optimization problem, and uncertainty of estimates is obtained by importance sampling from the posterior distribution. Li et al. [20] introduced an expectation-maximization (EM) algorithm similar to that of [24] but applied to isoforms instead of genes. Unlike the method of [17], which estimates isoform frequencies only from reads that map to a unique location in the genome, the algorithm of [20] incorporates multireads as well. The IE problem for single reads is also tackled in [26], who propose an EM algorithm for inferring isoform expression levels from the read coverage of exons (reads spanning exon junctions are ignored).

The related novel isoform discovery (ID) problem is also receiving much interest in the literature. Although showing encouraging results, *de novo* transcriptome assembly algorithms such as [5, 16, 30] have difficulties in identifying transcripts with moderate coverage. Very recently, [10, 12, 32] proposed genome-assisted (i.e., mapping based) methods for simultaneously solving ID and IE based on paired RNA-Seq reads. The method of Feng et al. [10] generates isoform candidates from the splicing graph derived from annotations and reads spanning exon-exon junctions. After discarding multireads, [10] formulates IE for a given set of isoforms as a convex quadratic program (QP) that can be efficiently solved for each gene locus. The set of isoform candidates is

iteratively refined until the p-value of the objective value of the QP, assumed to follow a χ^2 distribution, exceeds an empirically selected threshold of 5%. 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. As in [10], Guttman et al. [12] construct a splicing graph from the mapped reads and filter candidate isoforms using paired-end information. Isoform specific expression levels are inferred using the method of [22]. After performing spliced alignment of (paired) reads onto the genome using TopHat [31], the method of Trapnell et al. [32], referred to as 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 [17].

1.1.2 Our contributions

In this chapter 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 [7] and [21] 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 following ultra-deep paired-end transcriptome sequencing projects such as [12, 32] and the widely anticipated deployment of third-generation sequencing technologies such as [8, 9], which deliver reads with significantly increased length. 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 the 'Experimental results' section, 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 [22], or the EM algorithm of [24].

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 information provided by the distribution of insert sizes, which is tightly controlled during sequencing library preparation under current RNA-Seq protocols. Such information is not modeled in the "exact" information models of [14, 18], challenging the validity of their negative results. Guttman et al. [12] take into account insert lengths derived from paired read data, but only for filtering candidate isoforms in ID. Trapnell et al. [32] is the only other work we are aware of that exploits this information for IE, in conjunction with paired

read data. We show that modeling insert sizes is highly beneficial for IE even for RNA-Seq data consisting of single reads. Insert sizes contribute to increased estimation accuracy in two different ways. On one hand, they can 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 [20], the genomic locations of multireads are also resolved probabilistically in this step, further contributing to improved 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 maximization step of the IsoEM algorithm.

We also present the results of comprehensive experiments conducted to assess the performance of IsoEM on both synthetic and real RNA-Seq datasets. These results show that IsoEM consistently outperforms existing methods under a wide range of sequencing parameters and distribution assumptions. We also report results of experiments empirically evaluating the effect of sequencing parameters such as read length, read pairing, and strand information on estimation accuracy. Our experiments confirm the surprising finding of [20] 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.

1.2 Methods

1.2.1 Read mapping

As with many 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 ungapped aligners (we used Bowtie [19] 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 [31], as done, e.g., in [12, 32]. However, preliminary experiments with TopHat resulted in fewer mapped reads and significantly increased mapping uncertainty, despite providing TopHat with a complete set of annotated junctions. Since further increases in read length coupled with improvements in spliced alignment algorithms could make mapping onto the genome more attractive in the future, we made our IsoEM implementation compatible with both mapping approaches by always converting read alignments to genome coordinates and performing all IsoEM read-isoform compatibility calculations in genome space.

1.2.2 Finding read-isoform compatibilities

The candidate set of isoforms for each read is obtained by combining all genome coordinates of reads and isoforms, sorting them and using a line sweep technique to detect read-isoform compatibilities (see Figure 1.2.1) As detailed below, during the line sweep reads are grouped into equivalence classes defined by their isoform compatibility sets; this speeds up the E-step 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 $Q_a = \prod_{k=1}^{|r|}[(1-\varepsilon_k)M_{a_k} + \frac{\varepsilon_k}{3}(1-M_{a_k})]$, where $M_{a_k} = 1$ if position k of alignment a matches the reference genome sequence and 0 otherwise, while ε_k denotes the error probability of k-th base of r.

- For paired end reads, F_a represents the probability of the fragment length needed to produce alignment a from isoform j; note that the length of this fragment can be inferred from the genome coordinates of the two aligned reads and the available isoform annotation. For single reads, we can only estimate an upperbound u on the fragment length: if the alignment is on the same strand as the isoform then u is the number of isoform annotated bases between the 5′ end of the aligned read and the 3′ end of the isoform,

otherwise u is the number of isoform annotated bases between the 5′ end of the aligned read and the 5′ end of the isoform. In this case F_a is defined as the probability of observing a fragment with length of u bases or fewer.

- 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 paired-end reads, $O_a = 1$ 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.

1.2.3 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 denote by $n(j)$ the number of reads coming from isoform j and let $p(k)$ denote the probability of a fragment of length k, then

$$E[n(j)] \propto \sum_{k \leq l(j)} p(k)(l(j) - k + 1) \tag{1.2.1}$$

since, the number of fragments of length k is expected to be proportional to the number of valid starting positions for a fragment of that length in the

```
X = all the coordinates of all the entities (isoforms and reads)
sort X (radix sort; for equal values, isoform coordinates come first)
for x in X do
  e = entityFor(x)
  if x is an entity end then
    sig = signature[e]
    gap = getLastGap(sig)
    if x is an isoform end then
      currentIsoformsForGap[gap].remove(e)
    else if x is a read end then
      isoforms = currentIsoformsForGap[gap].keepOnlyMatching(sig)
      if read e is the second read in the pair then
        isoformsForRead[e] = isoformsForRead[e]∩ isoforms
      else
        isoformsForRead[e] = isoforms
      end if
      readClasses[isoformsForRead[e]].add(e)
    end if
    signature.remove(e)
  else
    signature[e].add(x)
  end if
  if x is an exon start then
    sig = signature[e]
    lastButOneGap = getLastButOneGap(sig)
    currentIsoformsForGap[lastButOneGap].remove(e)
    lastGap = getLastGap(sig)
    currentIsoformsForGap[lastGap].add(e, sig)
  end if
end for
```

Figure 1.2.1: The algorithm for identifying isoforms compatible with reads.

isoform. 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 $c(j) = n(j)/\sum_{k \leq l(j)} p(k)(l(j) - k + 1)$ denotes the length-normalized fragment coverage. Note that the length of most isoforms is significantly larger than the mean fragment length μ typical of current sequencing libraries; for such isoforms $\sum_{k \leq l(j)} p(k)(l(j) - k + 1) \approx l(j) - \mu + 1$ and $c(j)$ can be approximated by $n(j)/(l(j) - \mu + 1)$.

Since some reads match multiple isoforms, their isoform of origin cannot be established unambiguously. The IsoEM algorithm (see Figure 1.2.2) 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 $w_{r,j}$ computed as described in the previous section
- 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 the prior E-step

1.2.4 IsoEM optimizations

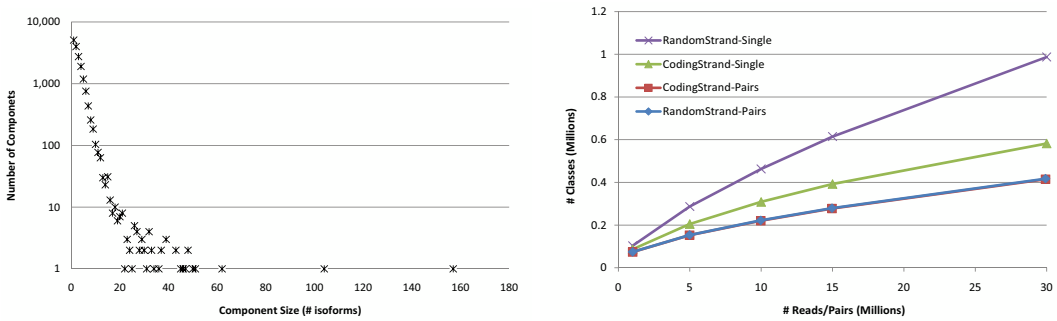
Below we describe two implementation optimizations that significantly improve the performance of IsoEM by reducing both runtime and memory usage. The first optimization consists of partitioning the input into compatibility components. The compatibility between reads and isoforms naturally induces a bipartite read-isoform compatibility graph, with edges connecting each isoform with all reads that can possibly originate from it. Connected components of the compatibility graph can be processed independently in IsoEM since the frequencies of isoforms in one connected component do not affect the frequencies

assign random values to all $f(i)$
while not converged **do**
  E-step:
  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
  M-step:
  $s = \sum_j n(j)/(l(j) - \mu + 1)$
  for each isoform j **do**
    $f(j) = \frac{n(j)/(l(j)-\mu+1)}{s}$
  end for
end while

Figure 1.2.2: The expectation-maximization algorithm used by IsoEM.

of isoforms in any other connected component. Although this optimization can be applied to any EM algorithm, its impact is particularly significant in IsoEM. Indeed, in this context the compatibility graph decomposes in numerous small components (see Figure 1.2.3(a) for a typical distribution of component sizes; a similar distribution of component sizes is reported for Arabidopsis gene models in [15]). The resulting speed-up comes from the fact that in each iteration of IsoEM we update frequencies of isoforms in a single compatibility component, avoiding needless updates for other isoforms.

The second IsoEM optimization consists of partitioning the set of reads within each compatibility component into equivalence classes. Two reads are equivalent for IsoEM if they are compatible with the same set of isoforms and their compatibility weights to the isoforms are proportional. Keeping only a single



(a) (b)

Figure 1.2.3: Distribution of compatibility component sizes (defined as the number of isoforms) for 10 million single reads of length 75 (a) and number of read classes for 1 to 30 million single reads or pairs of reads of length 75 (b).

E-step for read classes:
initialize all $n(j)$ to 0
for each read class R **do**
  sum $= \sum_{j:w_{R,j}>0} w_{R,j} f(j)$
  for each isoform j with $w_{R,j} > 0$ **do**
    $n(j) += m(R) * w_{R,j} f(j)/\text{sum}$
  end for
end for

Figure 1.2.4: The E-Step of IsoEM algorithm based on read classes.

representative from each read class (with appropriately adjusted frequency) drastically reduces the number of reads kept in memory (see Figure 1.2.3(b)). As the number of reads increases, the number of read classes increases much slower. Eventually this reaches saturation and no new read classes appear – at which point the runtime of IsoEM becomes virtually independent of the number of reads. Indeed, in practice the runtime bottlenecks are parsing the reads, computing the compatibility graph and detecting equivalent reads.

Once read classes are constructed, we only need a small modification of the

E-step of IsoEM to use read classes instead of reads (Figure 1.2.4). Next we describe the union-find algorithm used for efficiently finding compatibility components and read classes in IsoEM. A read class is defined as $\langle m, \{(i, w) | i = \text{isoform}, w = \text{weight}\}\rangle$, where m is called the multiplicity of the read class. Given a collection of reads, we want to:

- Find the connected components of the compatibility graph induced by the reads, and
- Collapse equivalent reads into read classes with multiplicity indicating the number of reads in each class.

A straightforward approach is to solve the first problem using a union-find algorithm, then to take the reads corresponding to each connected component and remove equivalent reads, e.g., using hashing. However, there are two drawbacks to this approach:

- First, all reads need to be kept in memory until all connected components have been computed.
- Second, when the number of reads in a connected component is very large the number of collisions increases, which leads to poor performance.

We overcome the two problems presented above using an online version of the union-find algorithm which computes connected components and eliminates equivalent reads on the fly. This way, equivalent reads will never reside too

long in memory. Also, we avoid the problem of large hash tables by using multiple smaller hash tables which are guaranteed to be disjoint.

We start our modified version of union-find with an empty set of trees. A new single-node tree is initialized every time a new isoform is found in a read class. In each node we store a hash-table of read classes. Each read is processed as follows:

- *If the isoforms compatible with the read correspond to nodes in more than one tree* unite the corresponding trees. The root of the tallest tree becomes the root of the union tree. Then create a new read class for this read (we can be sure it was not seen before, otherwise the isoforms would have been in the same tree) and add it to the hash table of the root node. Notice that at this point the root node is also (trivially) the Lowest Common Ancestor (LCA) of the nodes corresponding to the isoforms in the read class

- *If the isoforms correspond to nodes in the same tree* find the LCA of all these nodes. If the class of the read is present in the hash table of the LCA, increment its multiplicity and then drop the read. Otherwise, create a new read class and add it to the LCA's hash table.

Notice that in the second case it suffices to look only in the LCA of the isoforms for an already existing read class. This follows immediately from the fact that we always add reads to the LCA of the nodes (isoforms) compatible with the read. Note that we cannot use path compression to speed up 'find' operations

because this would be altering the structure of existing trees. Thus, 'find' operations will take logarithmic (amortized) time. At the end of the algorithm, each tree in the union-find forest corresponds to a connected component. The read classes in each connected component are obtained by traversing the corresponding tree and collecting all the read classes present in the nodes. At this point we are sure that all the read classes are distinct, so the collection process performs simple concatenations. To further speed up the collection process, we can safely use path compression as we traverse the trees, since we no longer care about the exact topology of the subtrees.

Runtime analysis. Each union operation takes $O(1)$ time, so for a read with k compatible isoforms we spend at most $O(k)$ time doing unions. By always making the root of the taller tree to be the root of a union, we ensure that the height of any tree is not bigger than $O(\log n)$ where n is the number of nodes in the tree. Thus, finding the root of a node's tree takes $O(\log n)$. For a read with k compatible isoforms we spend at most $O(k \log n)$ time processing it. The LCA of two nodes can be computed at constant overhead when performing find operations (by marking the nodes on the paths from isoforms to root). Collecting all the read classes is sped-up by using path compression. The whole collecting phase takes $O(n\alpha(n))$ time where n is the total number of isoforms and $\alpha(n)$ is the inverse of the Ackermann function. Overall, for q reads with an average of k isoforms per read and n total distinct isoforms, computing read classes and compatibility components using the modified union-find algorithm

takes $O(qk \log n + n\alpha(n))$ time.

1.2.5 Hexamer and repeat bias corrections

As noted in [13], some commonly used library preparation protocols result in biased sampling of fragments from isoforms due to the random hexamers used to prime reverse transcription. To correct for possible hexamer bias, we implemented a simple re-weighting scheme similar to that proposed in [13]. Each read is assigned a weight $b(h)$ based on its first six bases and computed as follows. Given a set of mapped reads, let $\hat{p}_i$ be the observed distribution of hexamers starting at position i (spanning positions i to $i + 5$) of all the reads. Thus, $\hat{p}_i(h)$ is the proportion of reads which have hexamer h at position i and $\hat{p}_1(h)$ is the proportion of reads starting with hexamer h. Let l be the read length. We define the weights b by:

$$b(h) = \frac{\frac{1}{6}\sum_{i=l/2-2}^{l/2+3} \hat{p}_i(h)}{\frac{1}{2}(\hat{p}_1(h) + \hat{p}_2(h))}$$

Since we already collapse equivalent reads into read classes, we can seamlessly incorporate hexamer weights in the algorithm by slightly changing the definition of a read class' multiplicity to $m(R) = \sum_{r \in R} b(h(r))$, where $h(r)$ denotes the starting hexamer of r. The effect of this correction procedure is to reduce (respectively increase) the multiplicity of reads with starting hexamers that are overrepresented (respectively under-represented) at the beginning of reads compared to the middle of reads. The underlying assumption is that

the average frequency with which a hexamer appears in the middle of reads is not affected by library preparation biases. Recent methods [27] further target biases in the bases surrounding the sequenced fragments in addition to those at read ends.

To avoid biases from incorrectly mapped reads originating from repetitive regions, IsoEM will also discard reads that overlap annotated repeats. When applying this correction, isoform lengths are automatically adjusted by subtracting the number of positions resulting in reads that would be discarded.

1.3 Experimental results

1.3.1 Comparison of methods on simulated datasets

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 clusters of known isoforms defined by the GNFAtlas2 table. The dataset contains a total of $66,803$ isoforms pertaining to $19,372$ genes. The isoform length distribution and the number of isoforms per genes are shown in Figure 1.3.1.

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, respectively truncated geometric distribution with ratio $r = 1/2$ for the isoforms of each gene. For a gene with k isoforms $p(j) = 1/k$, $j = 1, \ldots, k$, under the uniform distribution. Under the truncated geometric distribution, the respective isoform probabilities are $p(j) = 1/2^j$ for $j = 1, \ldots, k-1$ and $p(k) = 1/2^{k-1}$. Fragment lengths were simulated from a normal probability distribution with mean 250 and standard deviation 25.

We compared IsoEM to several existing algorithms for solving the IE and GE problems. For IE we included in the comparison the isoform analogs of the Uniq and Rescue methods used for GE [22], 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 [32] (version 0.8.2), and the RSEM algorithm of [20] (version 0.6). For the GE problem, the comparison included the Uniq and Rescue methods, our implementation of the GeneEM algorithm described in [24], and estimates obtained by summing isoform expression levels inferred by Cufflinks, RSEM, and IsoEM. All methods use alignments obtained by mapping reads onto the library of isoforms with Bowtie [19] and then converting them to genome coordinates, except for Cufflinks which uses alignments obtained by directly mapping the reads onto the genome with TopHat [31], as suggested

in [32].

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 [20]. However, accuracy was computed against true frequencies, not against estimates derived from true counts as in [20]. If $\hat{f}_i$ is the frequency 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 $EF_\tau = 50\%$.

Since not all compared methods could handle paired reads or strand information we focused our comparisons on single read data. Table 1.3.1 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 1.3.2 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 illustrated by the significant 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 explaining 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 significant IE and GE accuracy gains of IsoEM over RSEM. Indeed, the latest version of the RSEM package, released as this article goes to print, has been updated to include modeling of insert sizes and appears to have accuracy matching that of IsoEM.

For yet another view, Tables 1.3.2 and 1.3.3 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.

1.3.2 Comparison of methods on two real RNA-Seq datasets

In addition to simulation experiments, we validated IsoEM on two real RNA-Seq datasets. The first dataset consists of two samples with approximately 8

million 27bp Illumina reads each, generated from two human cell lines (embryonic kidney and B cells) as described in [29]. Estimation accuracy was assessed by comparison with quantitative PCR (qPCR) expression levels determined in [26] for 47 genes with evidence of alternative isoform expression. To facilitate comparison with these qPCR results, expression levels were determined using transcript annotations in ENSEMBL version 46. The second dataset consists of approximately 5 million 32bp Illumina reads per sample, generated from the RM11-1a strain of *S. cerevisiae* under two different nutrient conditions [6]. Expression levels were determined using transcript annotations for the reference strain (June 2008 SGD/sacCer2) and compared against qPCR expression levels measured for 192 genes (for a total of 394 datapoints).

Since the available implementation of RSEM could not be run on transcript sets other than UCSC known genes, in Figures 1.3.3 and 1.3.4 we only compare Cufflinks and IsoEM estimates against qPCR values in [26], respectively [6]. Estimation accuracy of both Cufflinks and IsoEM is significantly lower than that observed in simulations. Likely explanations include poor quality of the transcript libraries used to perform the inference, sequencing library preparation biases not corrected for by the algorithms, and possible inaccuracies in qPCR estimates. Nevertheless, the relative performance of the two algorithms is consistent with simulation results, with IsoEM outperforming Cufflinks on both datasets.

1.3.3 Influence of sequencing parameters and scalability

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* genome and transcriptome assembly. Surprisingly, [20] found that *shorter* reads are better for IE when the total number of sequenced bases (as a rough approximation for sequencing cost) is fixed. Figure 1.3.5 plots IE estimation accuracy for reads of length between 10 and 100 when the total amount of sequence data is kept constant at 750M bases. Our results confirm the finding of [20], although the optimal read length is somewhat sensitive to the accuracy measure used and to the availability of pairing information. While 25bp reads minimize MPE regardless of the availability of paired reads, the read length that maximizes r^2 is 25 for paired reads and 50 for single reads. Although further experiments are needed to determine how the optimum length depends on the amount of sequence data and transcriptome complexity, our simulations do suggest that for isoform and gene expression analysis, increasing the number of reads may be more useful than increasing read length beyond 50 bases.

Figure 1.3.6(a) 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.

In practice, many RNA-Seq data sets are generated from transcripts with poly(A) tails, and some of the sequenced fragments will contain parts the poly(A) tails. We have added to IsoEM the option to automatically extend annotated transcripts with a poly(A) tail, thus allowing it to use reads coming from such fragments. Table 1.3.4 shows the accuracy of isoform and gene expression levels inferred by IsoEM using 30M reads of length 25 simulated from transcripts with and without poly(A) tails assuming geometric expression of gene isoforms. The accuracy of IsoEM is practically the same under the two simulation scenarios for paired read data, and decreases only slightly for single reads simulated taking poly(A) tails into account, likely due to the fact that reads overlapping poly(A) tails are more ambiguous.

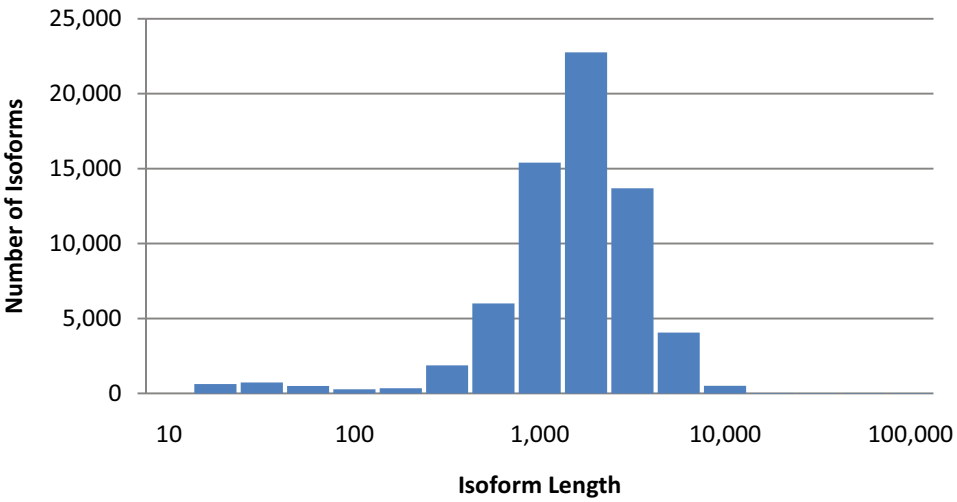
As shown in Figure 1.3.6(b), the runtime of IsoEM scales roughly linearly with the number of *fragments*, and is practically insensitive to the type of sequencing data (single or paired reads, directional or non-directional). IsoEM was tested on a Dell PowerEdge R900 server with 4 Six Core E7450Xeon Processors at 2.4Ghz (64 bits) and 128Gb of internal memory. None of the datasets required more than 16GB of memory to complete. It is also true that increasing the

available memory significantly decreases runtime by keeping the garbage collection overhead to a minimum. The runtimes in Figure 1.3.6 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.

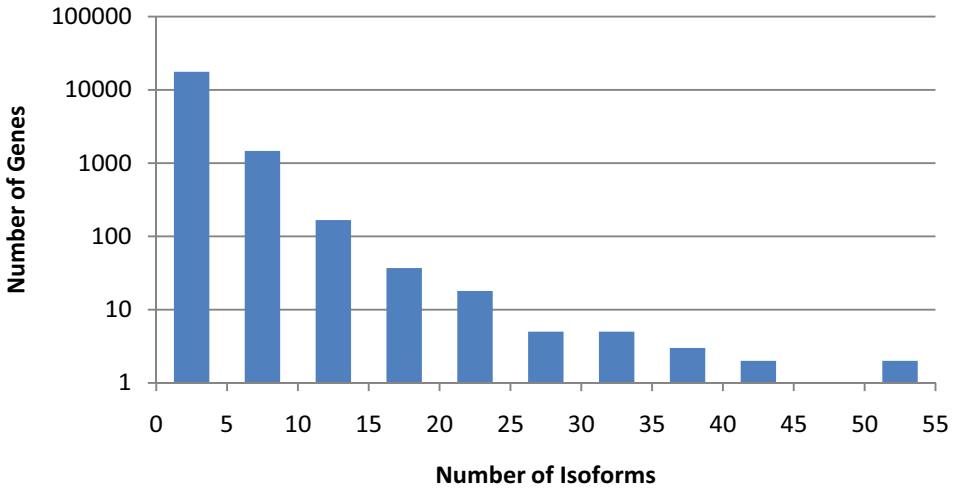
1.4 Conclusions

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



(a)

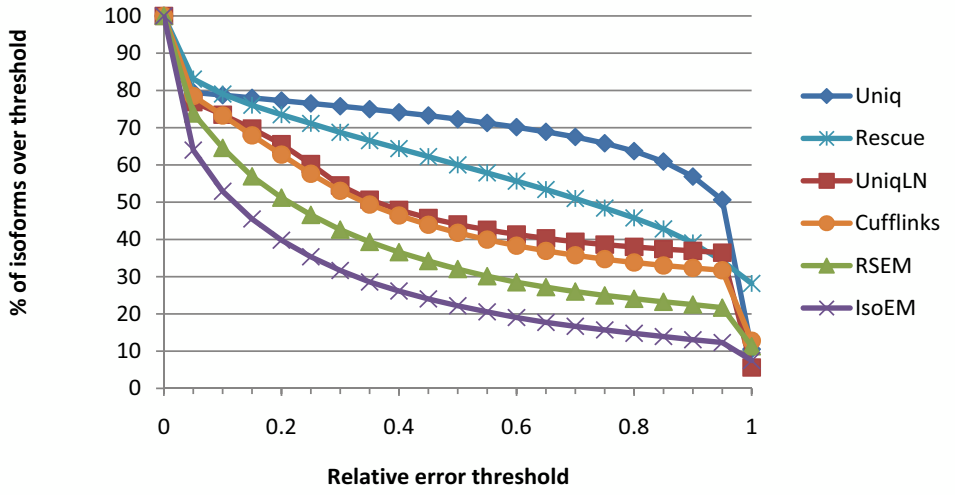


(b)

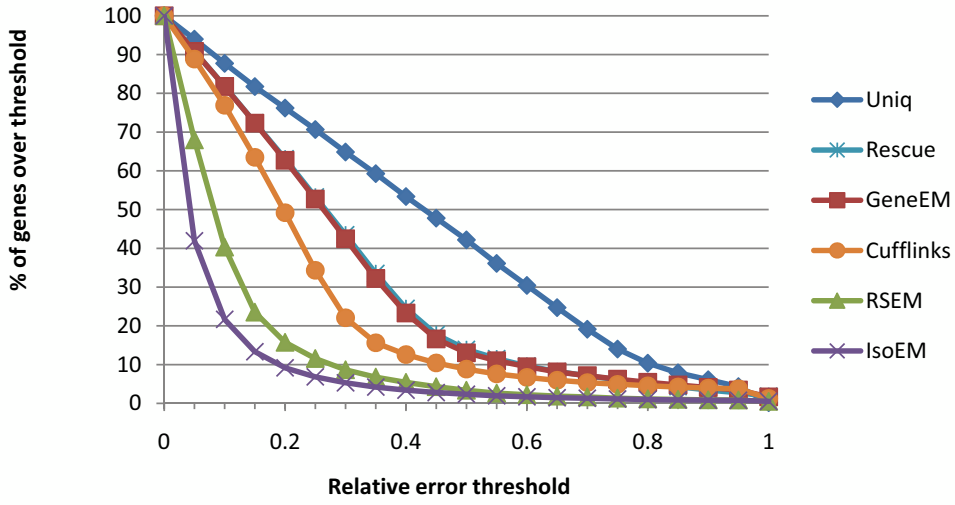
Figure 1.3.1: Distribution of isoform lengths (a) and gene cluster sizes (b) in the UCSC dataset.

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	**0.971**	**0.970**	IsoEM	**0.990**	**0.982**

Table 1.3.1: 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.



(a)



(b)

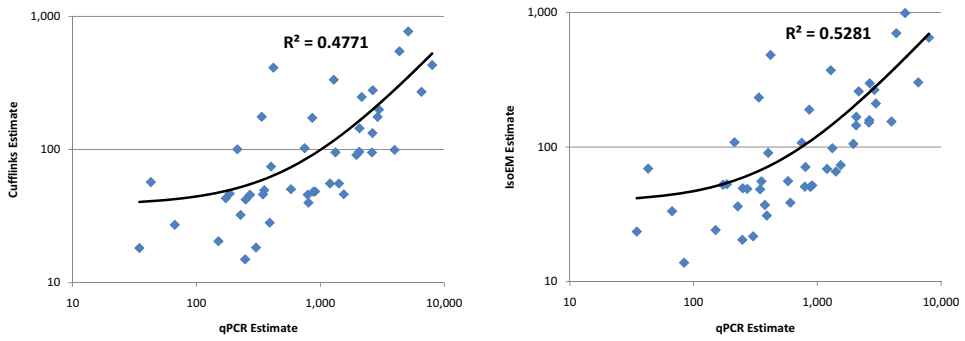
Figure 1.3.2: Error fraction at different thresholds for isoform (a) and gene (b) expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.

Expression range		0	$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
# isoforms		13,290	10,024	23,882	18,359	1,182	66	66,803
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**	**25.3**	**7.3**	**3.2**	**2.2**	**12.0**
$EF_{.15}$	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	3.4	**93.1**	**65.1**	**29.1**	**11.1**	**7.6**	**46.1**

Table 1.3.2: Median percent error (MPE) and 15% error fraction ($EF_{.15}$) for isoform expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.

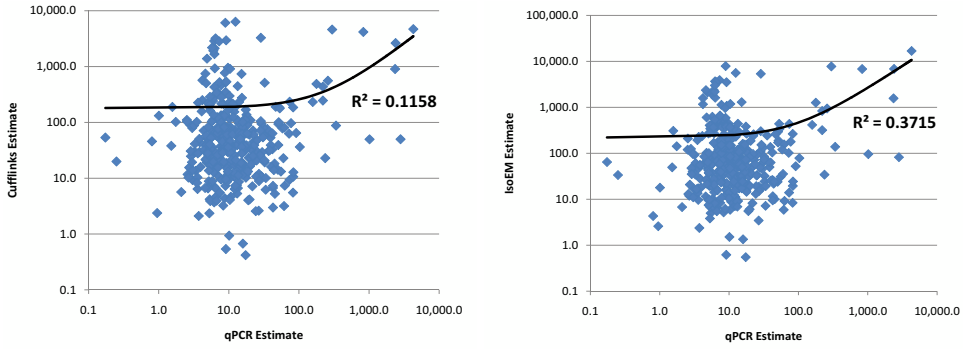
Expression range		$(0, 10^{-6}]$	$(10^{-6}, 10^{-5}]$	$(10^{-5}, 10^{-4}]$	$(10^{-4}, 10^{-3}]$	$(10^{-3}, 10^{-2}]$	All
# genes		120	5,610	11,907	1,632	102	19,372
MPE	Uniq	37.4	43.6	42.7	43.0	48.2	43.0
	Rescue	32.8	28.7	26.0	25.1	28.8	26.7
	GeneEM	30.6	28.2	25.7	25.1	28.0	26.3
	Cufflinks	33.0	21.1	19.0	20.2	40.2	19.7
	RSEM	23.6	11.0	7.2	7.9	11.4	8.1
	IsoEM	**18.2**	**8.4**	**3.2**	**2.0**	**1.9**	**3.9**
$EF_{.15}$	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**	**28.1**	**6.7**	**6.1**	**4.9**	**13.2**

Table 1.3.3: Median percent error (MPE) and 15% error fraction ($EF_{.15}$) for gene expression levels inferred from 30M reads of length 25 simulated assuming geometric isoform expression.



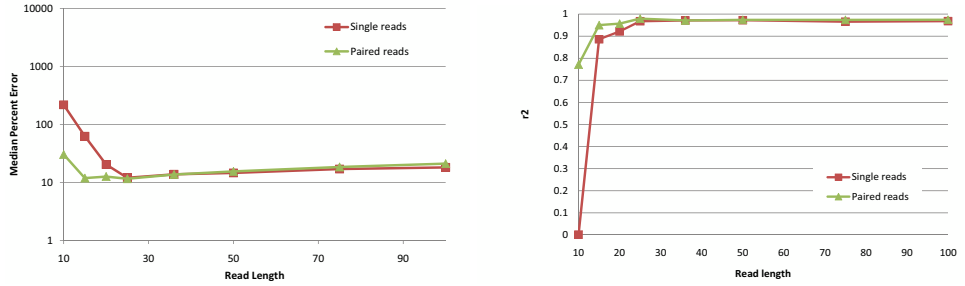
(a) (b)

Figure 1.3.3: Comparison of Cufflinks (a) and IsoEM (b) estimates to qPCR expression levels reported in [26].



(a) (b)

Figure 1.3.4: Comparison of Cufflinks (a) and IsoEM (b) estimates to qPCR expression levels reported in [6].

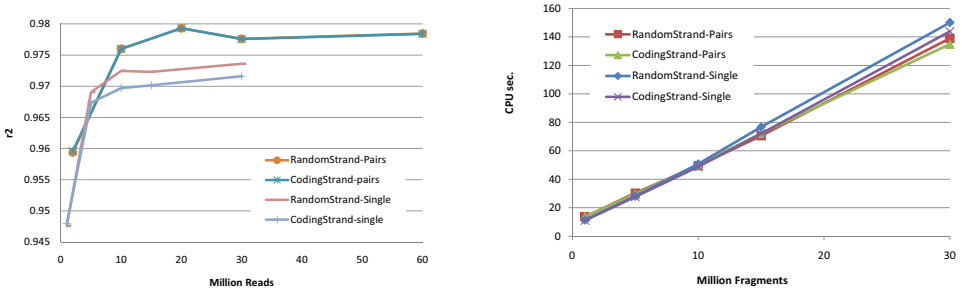


(a) (b)

Figure 1.3.5: IsoEM MPE (a) and r^2 values (b) for 750Mb of simulated data generated using single and paired-end reads of length varying between 10 and 100.

Reads	Poly(A)	Isoform Expression	Gene Expression
1× 25	Yes	0.956	0.977
	No	0.970	0.982
2× 25	Yes	0.972	0.990
	No	0.976	0.985

Table 1.3.4: r^2 for isoform and gene expression levels inferred from 30M single, respectively paired reads of length 25, simulated assuming geometric expression of gene isoforms with and without poly(A) tails.



(a) (b)

Figure 1.3.6: IsoEM r^2 (a) and CPU time (b) for 1-60 million single/paired reads of length 75, with or without strand information.

Chapter 2

Accurate Estimation of Gene Expression Levels from DGE Sequencing Data

2.1 Introduction

Massively parallel transcriptome sequencing is quickly replacing microarrays as the technology of choice for performing gene expression profiling due to its wider dynamic range and digital quantitation capabilities. However, accurate estimation of expression levels from sequencing data remains challenging due to the short read length delivered by current sequencing technologies and still poorly understood protocol- and technology-specific biases. To date, two main transcriptome sequencing protocols have been proposed in the literature. The most commonly used one, referred to as RNA-Seq, generates short (single or paired) sequencing tags from the ends of randomly generated cDNA fragments. An alternative protocol, referred to as 3'-tag Digital Gene Expression (DGE), or high-throughput sequencing based Serial Analysis of Gene Expression (SAGE-Seq), generates single cDNA tags using an assay including as main steps transcript capture and cDNA synthesis using oligo(dT) beads, cDNA cleavage with an anchoring restriction enzyme, and release of

cDNA tags using a tagging restriction enzyme whose recognition site is ligated upstream of the recognition site of the anchoring enzyme.

While computational methods for accurate inference of gene (and isoform) specific expression levels from RNA-Seq data have attracted much attention recently (see, e.g., [1, 20, 32]), analysis of DGE data still relies on direct estimates obtained from counts of uniquely mapped DGE tags [35, 40]. In part this is due to salient features of the DGE protocol, which, unlike RNA-Seq, guarantees that each mRNA molecule in the sample generates at most one tag and obviates the need for length normalization. Nevertheless, ignoring ambiguous DGE tags (which, due to the severely restricted tag length, can represent a sizeable fraction of the total) is at best discarding useful information, and at worst may result in systematic inference biases. In this chapter we seek to address this shortcoming of existing methods for DGE data analysis. Our main contribution is a rigorous statistical model of DGE data and a novel expectation-maximization algorithm for inference of gene and isoform expression levels from DGE tags. Unlike previous methods, our algorithm, referred to as DGE-EM, takes into account alternative splicing isoforms and tags that map at multiple locations in the genome, and corrects for incomplete digestion and sequencing errors. Experimental results show that DGE-EM outperforms methods based on unique tag counting on a multi-library DGE dataset consisting of 20bp tags generated from two commercially available reference RNA samples that have been well-characterized by quantitative real time PCR as

part of the MicroArray Quality Control Consortium (MAQC).

We also take advantage of the availability of RNA-Seq data generated from the same MAQC samples to directly compare estimation performance of the two transcriptome sequencing protocols. While RNA-Seq is clearly more powerful than DGE at detecting alternative splicing and novel transcripts such as fused genes, previous studies have suggested that for gene expression profiling DGE may yield accuracy comparable to that of RNA-Seq at a fraction of the cost [38]. We find that the two protocols achieve similar cost-normalized accuracy on the MAQC samples when using state-of-the-art estimation methods. However, the current protocol versions are unlikely to be optimal. Indeed, the results of a comprehensive simulation study assessing the effect of various experimental parameters suggest that further improvements in DGE accuracy could be achieved by using anchoring enzymes with degenerate recognition sites and using partial digest of cDNA with the anchoring enzyme during library preparation.

2.2 DGE Protocol

The DGE protocol generates short cDNA tags from a mRNA population in several steps (Figure 2.1.1). First, PolyA+ mRNA is captured from total RNA using oligo-dT magnetic beads and used as template for cDNA synthesis. The double stranded cDNA is then digested with a first restriction enzyme, called *Anchoring Enzyme* (AE), with known sequence specificity (e.g., the NlaIII en-

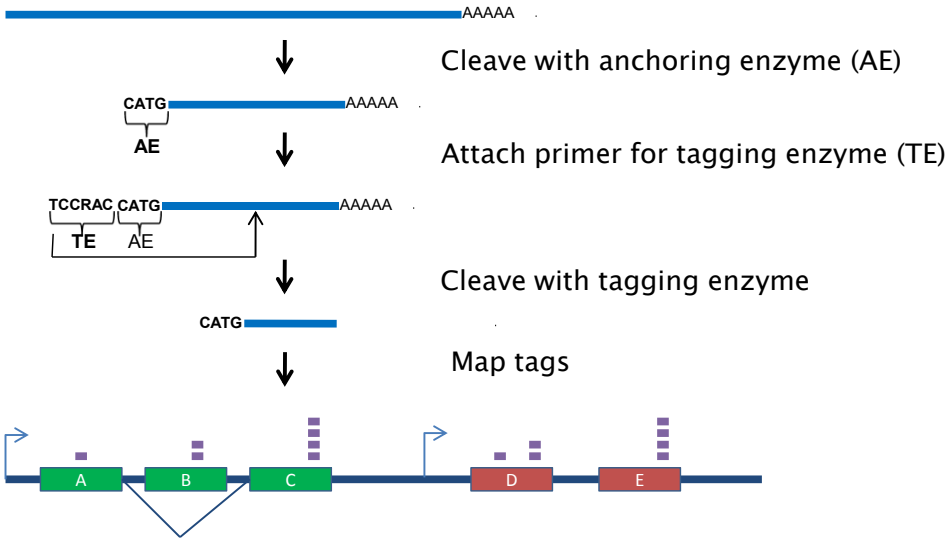


Figure 2.1.1: Schematic representation of the DGE protocol

zyme cleaves cDNA at sites at which the four nucleotide motif CATG appears). We refer to the cDNA sites cleaved by the anchoring enzyme as *AE sites*. The recognition site of a second restriction enzyme, called *Tagging Enzyme* (TE) is ligated to the fragments of cDNA that remain attached to the beads after cleavage with the AE, immediately upstream of the AE site. The cDNA fragments are then digested with TE, which cleaves several bases away from its recognition site. This results in very short cDNA tags (10 to 26 bases long, depending on the TE used), which are then sequenced using any of the available high-throughout technologies.

Since the recognition site of AE is only 4 bases long, most transcripts contain multiple AE sites. Under perfect experimental conditions, full digest by AE would ensure that DGE tags are generated only from the most 3′ AE site of each transcript. In practice some mRNA molecules release tags from other AE

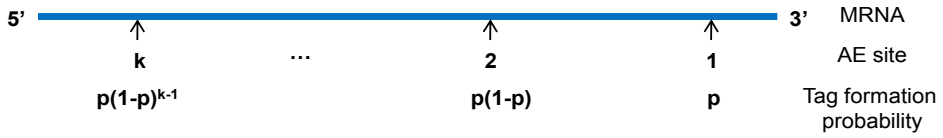


Figure 2.2.1: Tag formation probability: p for the rightmost AE site, geometrically decreasing for subsequent sites

sites, or no tag at all. As in [40], we assume that the cleavage probability of the AE, denoted by p, is the same for all AE sites of all transcripts. Since only the most 3′ cleaved AE site of a transcript releases a DGE tag, the probability of generating a tag from site $i = 1, \ldots, k$ follows a geometric distribution with ratio $1 - p$ as shown in Figure 2.2.1, where sites are numbered starting from the 3′ end. Note that splicing isoforms of a gene are likely to share many AE sites. However, the probability of generating a tag from a site is *isoform specific* since it depends on the number downstream AE sites on each isoform. Thus, although the primary motivation for this work is inference of gene expression levels from DGE tags, the algorithm presented in next section must take into account alternative splicing isoforms to properly allocate ambiguous tags among AE sites.

2.3 DGE-EM Algorithm

Previous studies have either discarded ambiguous DGE tags (e.g. [35, 40]) or used simple heuristic redistribution schemes for rescuing some of them. For example, in [39] the rightmost site in each transcript is identified as a "best" site. If a tag matches several locations, but only one of them is a best site, then

the tag is assigned to that site. If a tag matches multiple locations, none of which is a best site, the tag is equally split between these locations. In this section we detail an Expectation Maximization algorithm, referred to as DGE-EM, that probabilistically assigns DGE tags to candidate AE sites in different genes, different isoforms of the same gene, as well as different sites within the same isoform.

In a pre-processing step, a weight is assigned to each (DGE tag, AE site) pair, reflecting the conditional probability of the tag given the site that releases it. This probability is computed from base quality scores assuming that sequencing errors at different tag positions arise independently of one another. Formally, the weight for the alignment of tag t with the j^{th} rightmost AE site in isoform i is $w_{t,i,j} \propto \prod_{k=1}^{|t|}[(1-\varepsilon_k)M_{t_k} + \frac{\varepsilon_k}{3}(1-M_{t_k})]$, where $M_{t,k}$ is 1 if position k of tag t matches the corresponding position at site j in the transcript, 0 otherwise, while ε_k denotes the error probability of the k-th base of t, derived from the corresponding Phred quality score reported by the sequencing machine. In practice we only compute these weights for sites at which a tag can be mapped with a small (user selected) number of mismatches, and assume that remaining weights are 0. To each tag t we associate a "tag class" y_t which consists of the set of triples (i, j, w) where i is an isoform, j is an AE site in isoform i, and $w > 0$ is the weight associated as above to tag t and site j in isoform i. The collection of tag classes, $y = (y_t)_t$, represents the observed DGE data.

Let m be the number of isoforms. The parameters of the model are the relative

frequencies of each isoform, $\theta = (f_i)_{i=1,\ldots,m}$. Let $n_{i,j}$ denote the (unknown) number of tags generated from AE site j of isoform i. Thus, $x = (n_{i,j})_{i,j}$ represents the complete data. Denoting by k_i the number of AE sites in isoform i, by $N_i = \sum_{j=1}^{k_i} n_{i,j}$ the total number of tags from isoform i, and by $N = \sum_{i=1}^{m} N_i$ the total number of tags overall, we can write the complete data likelihood as

$$g(x|\theta) \propto \prod_{i=1}^{m} \prod_{j=1}^{k_i} \left[\frac{f_i (1-p)^{j-1} p}{S} \right]^{n_{i,j}} \tag{2.3.1}$$

where $S = \sum_{i=1}^{m} \sum_{j=1}^{k_i} f_i (1-p)^{j-1} p = \sum_{i=1}^{m} f_i \left(1 - (1-p)^{k_i}\right)$. Put into words, the probability of observing a tag from site j in isoform i is the frequency of that isoform (f_i) times the probability of not cutting at any of the first $j-1$ sites and cutting at the j^{th} [$(1-p)^{j-1}p$]. Notice that the algorithm effectively down-weights the matching AE sites far from the 3′ end based on the site probabilities shown in Figure 2.2.1. Since for each transcript there is a probability that no tag is actually generated, for the above formula to use proper probabilities we have to normalize by the sum S over all observable AE sites.

Taking logarithms in (2.3.1) gives the complete data log-likelihood:

$$\begin{aligned}
\log g(x|\theta) &= \sum_{i=1}^{m} \sum_{j=1}^{k_i} n_{i,j} \left[\log f_i + (j-1) \log (1-p) + \log p - \log S \right] + \text{constant} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{k_i} n_{i,j} \left[\log f_i + (j-1) \log (1-p) \right] \\
&\quad + N \log p - N \log \left(\sum_{i=1}^{m} f_i \left(1 - (1-p)^{k_i}\right) \right) + \text{constant}
\end{aligned}$$

2.3.1 E-Step

Let $c_{i,j} = \{y_t | \exists w \text{ s.t. } (i,j,w) \in y_t\}$ be the collection of all tag classes that are compatible with AE site j in isoform i. The expected number of tags from each cleavage site of each isoform, given the observed data and the current parameter estimates $\theta^{(r)}$, can be computed as

$$n_{i,j}^{(r)} := E(n_{i,j}|y,\theta^{(r)}) = \sum_{y_t \in c_{i,j},(i,j,w)\in y_t} \frac{f_i(1-p)^{j-1}pw}{\sum_{(l,q,z)\in y_t} f_l(1-p)^{q-1}pz} \tag{2.3.1}$$

This means that each tag class is fractionally assigned to the compatible isoform AE sites based on the frequency of the isoform, the probability of cutting at the cleavage sites where the tag matches, and the confidence that the tag comes from each location.

2.3.2 M-Step

In this step we want to select θ that maximizes the Q function,

$$Q(\theta|\theta^{(r)}) = E\left[\log g(x|\theta)|y,\theta^{(r)}\right] = \sum_{i=1}^{m}\sum_{j=1}^{k_i} n_{i,j}^{(r)}\left[\log f_i + (j-1)\log(1-p)\right]$$
$$+ N\log p - N\log\left(\sum_{i=1}^{m} f_i\left(1-(1-p)^{k_i}\right)\right) + \text{constant}$$

Partial derivatives of the Q function are:

$$\frac{\delta Q(\theta|\theta^{(r)})}{\delta f_i} = \frac{1}{f_i}\sum_{j=1}^{k_i} n_{i,j}^{(r)} + N\frac{1-(1-p)^{k_i}}{\sum_{l=1}^{m} f_l\left(1-(1-p)^{k_l}\right)}$$

Letting $C = N/(\sum_{l=1}^{m} f_l \left(1-(1-p)^{k_l}\right))$ and equating partial derivatives to 0 gives

$$\frac{N_i^{(r)}}{f_i} + C\left(1-(1-p)^{k_i}\right) = 0 \Longrightarrow f_i = -\frac{N_i^{(r)}}{C\left(1-(1-p)^{k_i}\right)}$$

Since $\sum_{i=1}^{m} f_i = 1$ it follows that

$$f_i = \frac{N_i^{(r)}}{1-(1-p)^{k_i}} \left(\sum_{l=1}^{m} \frac{N_l^{(r)}}{1-(1-p)^{k_l}} \right)^{-1} \tag{2.3.1}$$

2.3.3 Inferring p

In the above calculations we assumed that p is known, which may not be the case in practice. Assuming the geometric distribution of tags to sites, the observed tags of each isoform provide an independent estimate of p [40]. However, the presence of ambiguous tags complicates the estimation of p on an isoform-by-isoform basis. In order to globally capture the value of p we incorporate it in the DGE-EM algorithm as a hidden variable and iteratively re-estimate it as the distribution of tags to isoforms changes from iteration to iteration.

We estimate the value of p as N^1/D, where D denotes the total number of RNA molecules with at least one AE site, and $N^1 = \sum_{i=1}^{m} n_{i1}$ denotes the total number of tags coming from first AE sites. The total number of RNA molecules representing an isoform is computed as the number of tags coming from that isoform divided by the probability that the isoform is cut. This gives $D = \sum_{i=1}^{m} N_i/(1-(1-p)^{k_i})$, which happens to be the normalization term used in the

M step of the algorithm.

2.3.4 Implementation

For an efficient implementation, we pre-process AE sites in all the known isoform sequences. All tags that can be generated from these sites, assuming no errors, are stored in a trie data structure together with information about their original locations. Searching for a tag is performed by traversing the trie, permitting for as many jumps to neighboring branches as the maximum number of mismatches allowed. The Expectation Maximization part of DGE-EM, which follows after mapping, is given in Algorithm 1 (for simplicity, the re-estimation of p is omitted).

Algorithm 1 DGE-EM algorithm

assign random values to all $f(i)$
while not converged **do**
  initialize all **n**(**iso**, **site**) to 0
  for each tag class t **do**
    sum = $\sum_{(\text{iso},\text{site},w)\in t} w \times f(\text{iso}) \times (1-p)^{\text{site}-1}$
    for (iso, site, w) $\in t$ **do**
      n(iso, site)+ = $w \times f(\text{iso}) \times (1-p)^{\text{site}-1}$/sum
    end for
  end for
  for each isoform i **do**
    $N_i = \sum_{j=1}^{\text{sites}(i)} n(i,j)$
    $f(i) = N_i/(1-(1-p)^{\text{sites}(i)})$
  end for
end while

In practice, for performance reasons, tags with the same matching sites and weights are collapsed into one, keeping track of their multiplicity. Then the EM algorithm can process them all at once by factoring in their multiplicity

when increasing the n(iso, site) counter. This greatly reduces the running time and memory footprint.

2.4 Results

2.4.1 Experimental Setup

We conducted experiments on both real and simulated DGE and RNA-Seq datasets. In addition to estimates obtained by DGE-EM, for DGE data we also computed direct estimates from uniquely mapped tags; we refer to this method as "Uniq". RNA-Seq data was analyzed using both our IsoEM algorithm [1], which was shown to outperform existing methods of isoform and gene expression level estimation, and the well-known Cufflinks algorithm [32]. As in previous works [1, 20], estimation accuracy was assessed using the *median percent error (MPE)*, which gives the median value of the relative errors (in percentage) over all genes.

Real DGE datasets included nine libraries kindly provided to us (in fastq format) by the authors of [35]. These libraries were independently prepared and sequenced at multiple sites using 6 flow cells on Illumina Genome Analyzer (GA) I and II platforms, for a total of 35 lanes. The first eight libraries were prepared from the Ambion Human Brain Reference RNA, (Catalog #6050), henceforth referred to as HBRR and the ninth was prepared from the Stratagene Universal Human Reference RNA (Catalog #740000) henceforth referred to as UHRR. *DpnII*, with recognition site GATC, was used as anchoring enzyme

and *MmeI* as tagging enzyme, resulting in approximately 238 million tags of length 20 across the 9 libraries. Unless otherwise indicated, Uniq estimates are based on uniquely mapped tags with 0 mismatches (63% of all tags) while for DGE-EM we used all tags mapped with at most 1 mismatch (83% of all tags) since preliminary experiments (Section 2.4.2) showed that these are the optimal settings for each algorithm.

For comparison, we downloaded from the SRA repository two RNA-Seq datasets for the HBRR sample and six RNA-Seq datasets for the UHRR sample (SRA study SRP001847 [36]) . Each RNA-Seq dataset contains between 47 and 92 million reads of length 35. We mapped RNA-Seq reads onto Ensembl known isoforms (version 59) using bowtie [19] after adding a polyA tail of 200 bases to each transcript. Allowing for up to two mismatches, we were able to map between 65% and 72% of the reads. We then ran IsoEM and Cufflinks assuming a mean fragment length of 200 bases with standard deviation 50.

To assess accuracy, gene expression levels estimated from real DGE and RNA-Seq datasets were compared against TaqMan qPCR measurements (GEO accession GPL4097) collected by the MicroArray Quality Control Consortium (MAQC). As described in [37], each TaqMan Assay was run in four replicates for each measured gene. POLR2A (ENSEMBL id ENSG00000181222) was chosen as the reference gene and each replicate CT was subtracted from the average POLR2A CT to give the log2 difference (delta CT). For delta CT calculations, a CT value of 35 was used for any replicate that had $CT > 35$. Normalized expres-

sion values are reported: $2^{(\text{CT of POLR2A})-(\text{CT of the tested gene})}$. We used the average of the qPCR expression values in the four replicates as the ground truth. After mapping gene names to Ensembl gene IDs using the HUGO Gene Nomenclature Committee (HGNC) database, we got TaqMan qPCR expression levels for 832 Ensembl genes. Expression levels inferred from DGE and RNA-Seq data were similarly divided by the expression level inferred for POLR2A prior to computing accuracy.

Synthetic error-free DGE and RNA-Seq data was generated using an approach similar to that described in [1]. Briefly, the human genome sequence (hg19, NCBI build 37) was downloaded from UCSC and used as reference. We used isoforms in the UCSC KnownGenes table ($n = 77,614$), and defined genes as clusters of known isoforms in the GNFAtlas2 table ($n = 19,625$). We conducted simulations based on gene expression levels for five different tissues in GNFAtlas2. The simulated frequency of isoforms within gene clusters followed a geometric distribution with ratio 0.5. For DGE we simulated data for all restriction enzymes with 4-base long recognition sites from the Restriction Enzyme Database (REBASE), assuming either complete digestion ($p = 1$) or partial digestion with $p = 0.5$. For RNA-Seq we simulated fragments of mean length 250 and standard deviation 25 and simulated polyA tails with uniform length of 250bp. For all simulated data mapping was done without allowing mismatches.

2.4.2 DGE-EM Outperforms Uniq

The algorithm referred to as Uniq quantifies gene expression based on the number of tags that match one or more cleavage sites in isoforms belonging to the same gene. These tags are unique with respect to the source gene. Figure 2.4.1 compares the accuracy of Uniq and DGE-EM on library 4 from the HBRR sample, with the number of allowed mismatches varying between 0 and 2. As expected, counting only perfectly mapped tags gives the best accuracy for Uniq, since with the number of mismatches we increase the ambiguity of the tags, and thus reduce the number of unique ones. When run with 0 mismatches, DGE-EM already outperforms Uniq, but the accuracy improvement is limited by the fact that it cannot tolerate any sequencing errors (tags including errors are either ignored, or, worse, mapped at an incorrect location). Allowing 1 mismatch per tag gives the best accuracy of all compared methods, but further increasing the number of mismatches to 2 leads to accuracy below that achieved when using exact matches only, likely due to the introduction of excessive tag ambiguity for data for which the error rate is well below 10%.

2.4.3 Comparison of DGE and RNA-Seq Protocols

Figure 2.4.2 shows the gene expression estimation accuracy for 9 DGE and 8 RNA-Seq libraries generated from the HBRR and UHRR MAQC sample. All DGE estimates were obtained using the DGE-EM algorithm, while for RNA-Seq data we used both IsoEM [1] and the well-known Cufflinks algorithm [32]. The cutting probability inferred by DGE-EM is almost the same for all

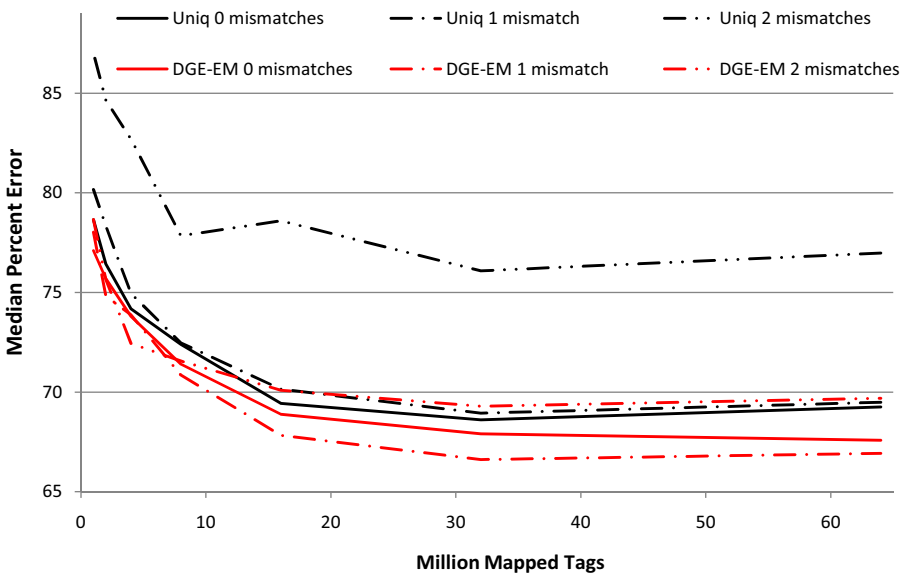


Figure 2.4.1: Median Percent Error of DGE-EM and Uniq estimates for varying number of allowed mismatches and DGE tags generated from the HBRR library 4.

libraries, with a mean of 0.8837 and standard deviation 0.0049. This is slightly higher than the estimated value of 70 – 80% suggested in the original study [35], possibly due to their discarding of non-uniq or non-perfectly matched tags. Normalized for sequencing cost, DGE performance is comparable to that of RNA-Seq estimates obtained by IsoEM, with accuracy differences between libraries produced using different protocols within the range of library-to-library variability within each of the two protocols. The MPE of estimates generated from RNA-Seq data by Cufflinks is significantly higher than that of IsoEM and DGE-EM estimates, suggesting that accurate analysis methods are at least as important as the sequencing protocol.

2.4.4 Possible DGE Assay Optimizations

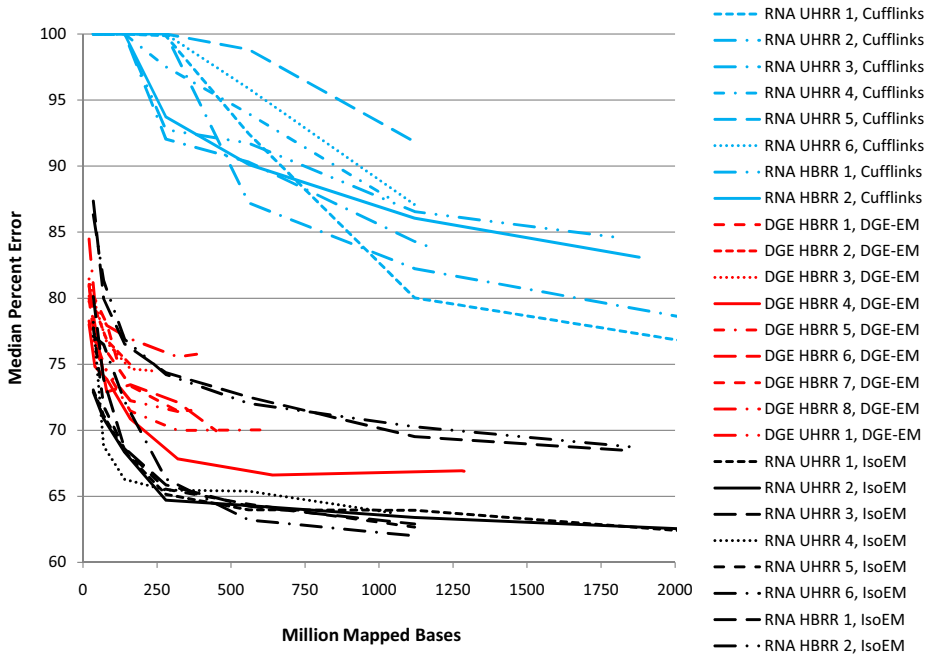


Figure 2.4.2: Median Percent Error of DGE-EM, IsoEM, and Cufflinks estimates from varying amounts of DGE/RNA-Seq data generated from the HBRR MAQC sample.

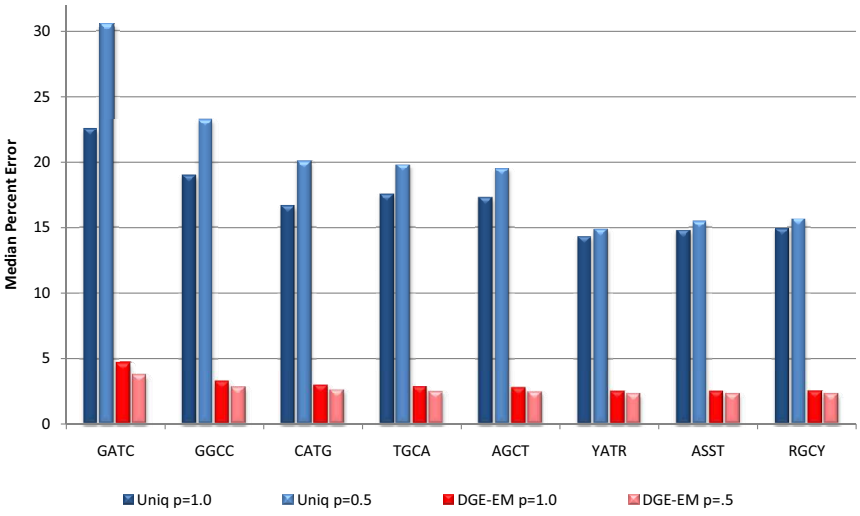
To assess accuracy of DGE estimates under various protocol parameters, we conducted an extensive simulation study where we varied the anchoring enzyme used, the number of tags, the tag length and the cutting probability. We tested all restriction enzymes with 4-base long recognition sites from REBASE. Figure 2.4.3(a) gives MPE values obtained by the Unique and DGE-EM algorithms for a subset of these enzymes on synthetic datasets with 30 million tags of length 21, simulated assuming either complete or $p = .5$ partial digest. Figure 2.4.3(b) gives the percentage of genes cut and the percentage of uniquely mapped DGE tags for each of these enzymes. These results suggest that using enzymes with high percentage of genes cut leads to improvements in accuracy.

In particular, enzymes like NlaIII (previously used in [39]) with recognition site CATG and CviJI with degenerate recognition site RGCY (R=G or A, Y=C or T) cut more genes than the DpnII (GATC) enzyme used to generate the MAQC DGE libraries, and yield better accuracy for both Uniq and DGE-EM estimates. Furthermore, for every anchoring enzyme, partial digestion with $p = .5$ yields an improved DGE-EM accuracy compared to complete digestion. Interestingly, Unique estimates are less accurate for partial digest due to the smaller percentage of uniquely mapped reads. For comparison, IsoEM estimates based on 30 million RNA-Seq tags of length 21 yield an MPE of 8.3.

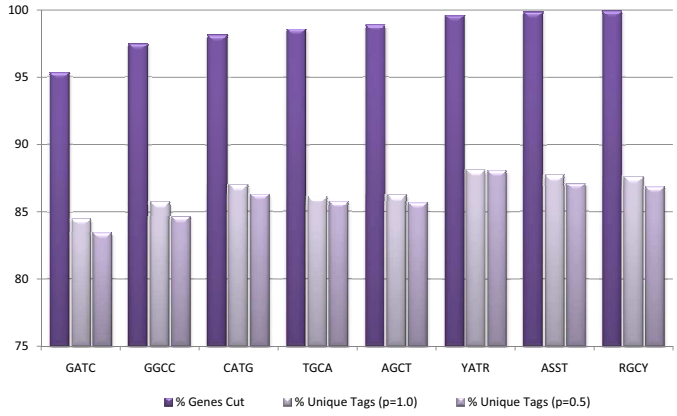
2.5 Conclusions

In this chapter we introduce a novel expectation-maximization algorithm, called DGE-EM, for inference of gene-specific expression levels from DGE tags. Our algorithm takes into account alternative splicing isoforms and tags that map at multiple locations in the genome within a unified statistical model, and can further correct for incomplete digestion and sequencing errors. Experimental results on both real and simulated data show that DGE-EM outperforms commonly used estimation methods based on unique tag counting. DGE-EM has cost-normalized accuracy comparable to that achieved by state-of-the-art RNA-Seq estimation algorithms on the tested real datasets, and outperforms them on error-free synthetic data. Simulation results suggest that further accuracy improvements can be achieved by tuning DGE protocol parameters such

as the degeneracy of the anchoring enzyme and cutting probability. It would be interesting to experimentally test this hypothesis.



(a)



(b)

Figure 2.4.3: (a) Median Percent Error of Unique and DGE-EM estimates obtained from 30 million 21bp DGE tags simulated for anchoring enzymes with different restriction sites (averages over 5 GNF-Atlas tissues) (b) Percentage of genes cut and uniquely mapped tags for each anchoring enzyme.

Bibliography

[1] Marius Nicolae, Serghei Mangul, Ion I. Mandoiu, and Alexander Zelikovsky. Estimation of alternative splicing isoform frequencies from RNA-Seq data. In Vincent Moulton and Mona Singh, editors, *Proc. WABI*, volume 6293 of *Lecture Notes in Computer Science*, pages 202–214. Springer, 2010.

[2] M. Nicolae and I.I. Mandoiu. Accurate estimation of gene expression levels from dge sequencing data. In *Proc. 7th International Symposium on Bioinformatics Research and Applications*, volume 6674 of *Lecture Notes in Computer Science*, pages 392–403, 2011.

[3] M. Nicolae, S. Mangul, I.I. Mandoiu, and A. Zelikovsky. Estimation of alternative splicing isoform frequencies from rna-seq data. *Algorithms for Molecular Biology*, 6:9, 2011.

[4] 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.

[5] 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.

[6] Joshua Bloom, Zia Khan, Leonid Kruglyak, Mona Singh, and Amy Caudy. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics*, 10(1):221, 2009.

[7] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta,

D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). The Transcriptional Landscape of the Mammalian Genome. *Science*, 309(5740):1559–1563, 2005.

[8] J. Clarke, H.-C. Wu, L. Jayasinghe, A. Patel, S. Reid, and H. Bayley. Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4(4):265–270, 2009.

[9] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex Dewinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Ali-

cia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, 2009.

[10] Jianxing Feng, Wei Li, and Tao Jiang. Inference of Isoforms from Short Sequence Reads. In Bonnie Berger, editor, *Research in Computational Molecular Biology*, volume 6044 of *Lecture Notes in Computer Science*, pages 138–157. Springer, Berlin, Germany, 2010.

[11] Malachi Griffith, Obi L. Griffith, Jill Mwenifumbo, Rodrigo Goya, A. Sorana Morrissy, Ryan D. Morin, Richard Corbett, Michelle J. Tang, Ying-Chen Hou, Trevor J. Pugh, Gordon Robertson, Suganthi Chittaranjan, Adrian Ally, Jennifer K. Asano, Susanna Y. Chan, Haiyan I. Li, Helen McDonald, Kevin Teague, Yongjun Zhao, Thomas Zeng, Allen Delaney, Martin Hirst, Gregg B. Morin, Steven J. M. Jones, Isabella T. Tai, and Marco A. Marra. Alternative expression analysis by RNA sequencing. *Nature Methods*, 7(10):843–847, 2010.

[12] M. Guttman, M. Garber, J.Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M.J. Koziol, A. Gnirke, C. Nusbaum, J.L. Rinn, E.S. Lander, and A. Regev. *Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5):503–510, 2010.

[13] Kasper D. Hansen, Steven E. Brenner, and Sandrine Dudoit. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucl. Acids Res.*, 38(12):e131+, 2010.

[14] 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.

[15] Brian E. Howard and Steffen Heber. Towards reliable isoform quantification using RNA-SEQ data. *BMC bioinformatics*, 11 Suppl 3(Suppl 3):S6+, 2010.

[16] B. Jackson, P. Schnable, and S. Aluru. Parallel short sequence assembly of transcriptomes. *BMC Bioinformatics*, 10(Suppl 1):S14+, 2009.

[17] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009.

[18] Vincent Lacroix, Michael Sammeth, Roderic Guigo, and Anne Bergeron. Exact transcriptome reconstruction from short sequence reads. In Keith Crandall and Jens Lagergren, editors, *Algorithms in Bioinformatics*, volume

5251 of *Lecture Notes in Computer Science*, pages 50–63. Springer, Berlin, Germany, 2008.

[19] 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.

[20] 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.

[21] MGC Project Team, Gary Temple, Daniela S. Gerhard, Rebekah Rasooly, Elise A. Feingold, Peter J. Good, Cristen Robinson, Allison Mandich, Jeffrey G. Derge, Jeanne Lewis, Debonny Shoaf, Francis S. Collins, Wonhee Jang, Lukas Wagner, Carolyn M. Shenmen, Leonie Misquitta, Carl F. Schaefer, Kenneth H. Buetow, Tom I. Bonner, Linda Yankie, Ming Ward, Lon Phan, Alex Astashyn, Garth Brown, Catherine Farrell, Jennifer Hart, Melissa Landrum, Bonnie L. Maidak, Michael Murphy, Terence Murphy, Bhanu Rajput, Lillian Riddick, David Webb, Janet Weber, Wendy Wu, Kim D. Pruitt, Donna Maglott, Adam Siepel, Brona Brejova, Mark Diekhans, Rachel Harte, Robert Baertsch, Jim Kent, David Haussler, Michael Brent, Laura Langton, Charles L. Comstock, Michael Stevens, Chaochun Wei, Marijke J. van Baren, Kourosh Salehi-Ashtiani, Ryan R. Murray, Lila Ghamsari, Elizabeth Mello, Chenwei Lin, Christa Pennacchio, Kirsten Schreiber, Nicole Shapiro, Amber Marsh, Elizabeth Pardes, Troy Moore, Anita Lebeau, Mike Muratet, Blake Simmons, David Kloske, Stephanie Sieja, James Hudson, Praveen Sethupathy, Michael Brownstein, Narayan Bhat, Joseph Lazar, Howard Jacob, Chris E. Gruber, Mark R. Smith, John McPherson, Angela M. Garcia, Preethi H. Gunaratne, Jiaqian Wu, Donna Muzny, Richard A. Gibbs, Alice C. Young, Gerard G. Bouffard, Robert W. Blakesley, Jim Mullikin, Eric D. Green, Mark C. Dickson, Alex C. Rodriguez, Jane Grimwood, Jeremy Schmutz, Richard M. Myers, Martin Hirst, Thomas Zeng, Kane Tse, Michelle Moksa, Merinda Deng, Kevin Ma, Diana Mah, Johnson Pang, Greg Taylor, Eric Chuah, Athena Deng, Keith Fichter, Anne Go, Stephanie Lee, Jing Wang, Malachi Griffith, Ryan Morin, Richard A. Moore, Michael Mayo, Sarah Munro, Susan Wagner, Steven J. Jones, Robert A. Holt, Marco A. Marra, Sun Lu, Shuwei Yang, James Hartigan, Marcus Graf, Ralf Wagner, Stanley Letovksy, Jacqueline C. Pulido, Keith Robison, Dominic Esposito, James Hartley, Vanessa E. Wall, Ralph F. Hopkins, Osamu Ohara, and Stefan Wiemann. The completion of the Mammalian Gene Collection (MGC). *Genome Research*, 19(12):2324–2333, 2009.

[22] Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and

Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, May 2008.

[23] A. Oshlack and M. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4(1):14, 2009.

[24] B. Paşaniuc, N. Zaitlen, and E. Halperin. Accurate estimation of expression levels of homologous genes in RNA-seq experiments. In B. Berger, editor, *Proc. 14th Annual Intl. Conf. on Research in Computational Molecular Biology (RECOMB)*, volume 6044 of *Lecture Notes in Computer Science*, pages 397–409, Berlin, Germany, 2010. Springer.

[25] Chris P. Ponting and T. Grant Belgard. Transcribed dark matter: meaning or myth? *Human Molecular Genetics*, 19(R2):R162–R168, August 2010.

[26] 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.

[27] Adam Roberts, Cole Trapnell, Julie Donaghey, John Rinn, and Lior Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biology*, 12(3):R22, 2011.

[28] Y. She, E. Hubbell, and H. Wang. Resolving deconvolution ambiguity in gene alternative splicing. *BMC Bioinformatics*, 10(1):237, 2009.

[29] Marc Sultan, Marcel H. Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure L. Yaspo. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891):956–960, 2008.

[30] Y. Surget-Groba and J.I. Montoya-Burgos. Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research*, 20(10):1432–1440, October 2010.

[31] C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[32] 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.

[33] 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.

[34] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.

[35] Y. Asmann, E.W. Klee, E.A. Thompson, E. Perez, S. Middha, A. Oberg, T. Therneau, D. Smith, G. Poland, E. Wieben, and J.-P. Kocher. 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics*, 10(1):531, 2009.

[36] J. Bullard, E. Purdom, K. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 2010.

[37] MAQC Consortium. The Microarray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–1161, September 2006.

[38] Peter A. 't Hoen, Yavuz Ariyurek, Helene H. Thygesen, Erno Vreugdenhil, Rolf H. Vossen, Renée X. de Menezes, Judith M. Boer, Gert-Jan J. van Ommen, and Johan T. den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic acids research*, 36(21):e141+, 2008.

[39] Zhenhua Jeremy Wu, Clifford A. Meyer, Sibgat Choudhury, Michail Shipitsin, Reo Maruyama, Marina Bessarabova, Tatiana Nikolskaya, Saraswati Sukumar, Armin Schwartzman, Jun S. Liu, Kornelia Polyak, and X. Shirley Liu. Gene expression profiling of human breast tissue samples using SAGE-Seq. *Genome Research*, 20(12):1730–1739, 2010.

[40] R. Zaretzki, M. Gilchrist, W. Briggs, and A. Armagan. Bias correction and Bayesian analysis of aggregate counts in SAGE libraries. *BMC Bioinformatics*, 11(1):72, 2010.