Accurate Estimation of Gene Expression Levels from DGE Sequencing Data

Marius Nicolae and Ion Măndoiu

Computer Science & Engineering Department, University of Connecticut
371 Fairfield Way, Storrs, CT 06269
{man09004,ion}@engr.uconn.edu

Abstract. Two main transcriptome sequencing protocols have been proposed in the literature: the most commonly used shotgun sequencing of full length mRNAs (RNA-Seq) and 3'-tag digital gene expression (DGE). In this paper we present a novel expectation-maximization algorithm, called DGE-EM, for inference of gene-specific 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. The open source Java/Scala implementation of the DGE-EM algorithm is freely available at `http://dna.engr.uconn.edu/software/DGE-EM/`. Experimental results on real DGE data generated from reference RNA samples show that our algorithm outperforms commonly used estimation methods based on unique tag counting. Furthermore, the accuracy of DGE-EM estimates is comparable to that obtained by state-of-the-art estimation algorithms from RNA-Seq data for the same samples. Results of a comprehensive simulation study assessing the effect of various experimental parameters suggest that further improvements in estimation accuracy could be achieved by optimizing DGE protocol parameters such as the anchoring enzymes and digestion time.

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., [4, 6, 8]), analysis of DGE data still relies on direct estimates obtained from counts of uniquely mapped DGE tags [1, 10]. 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 paper 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 [7]. 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 DGE Protocol

The DGE protocol generates short cDNA tags from a mRNA population in several steps (Figure 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 enzyme cleaves cDNA at sites at which the four nucleotide motif CATG appears). We refer to

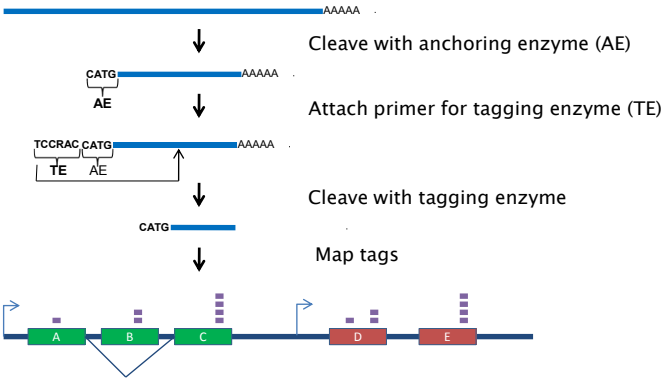


Fig. 1. Schematic representation of the DGE protocol

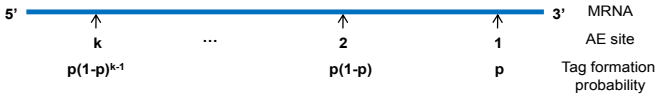


Fig. 2. Tag formation probability: p for the rightmost AE site, geometrically decreasing for subsequent sites

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 sites, or no tag at all. As in [10], 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, 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.

3 DGE-EM Algorithm

Previous studies have either discarded ambiguous DGE tags (e.g. [1, 10]) or used simple heuristic redistribution schemes for rescuing some of them. For example, in [9] 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

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

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. 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 (1) gives the complete data log-likelihood:

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

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

$$n_{i,j}^{(r)} := E(n_{i,j}|y,\theta^{(r)}) = \sum_{y_t \in c_{i,j},(i,j,w)\in y_t} \frac{f_i(1-p)^{j-1}pw}{\sum_{(l,q,z)\in y_t} f_l(1-p)^{q-1}pz} \tag{2}$$

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.

3.2 M-Step

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

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

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

$$f_i = \frac{N_i^{(r)}}{1-(1-p)^{k_i}} \left(\sum_{l=1}^{m} \frac{N_l^{(r)}}{1-(1-p)^{k_l}} \right)^{-1} \tag{3}$$

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 [10]. 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.

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).

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 $\mathrm{n}(\mathrm{iso}, \mathrm{site})$ counter. This greatly reduces the running time and memory footprint.

4 Results

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

Algorithm 1 DGE-EM algorithm

assign random values to all $f(i)$
while not converged **do**
  initialize all $\mathbf{n}(\mathbf{iso}, \mathbf{site})$ 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**
      $\text{n}(\text{iso}, \text{site}) += w \times f(\text{iso}) \times (1-p)^{\text{site}-1}/\text{sum}$
    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

as "Uniq". RNA-Seq data was analyzed using both our IsoEM algorithm [6], which was shown to outperform existing methods of isoform and gene expression level estimation, and the well-known Cufflinks algorithm [8]. As in previous works [4, 6], 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 [1]. 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 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 [2]) . 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 [3] 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 [5], 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 expression 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 [6]. 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.

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 3 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%.

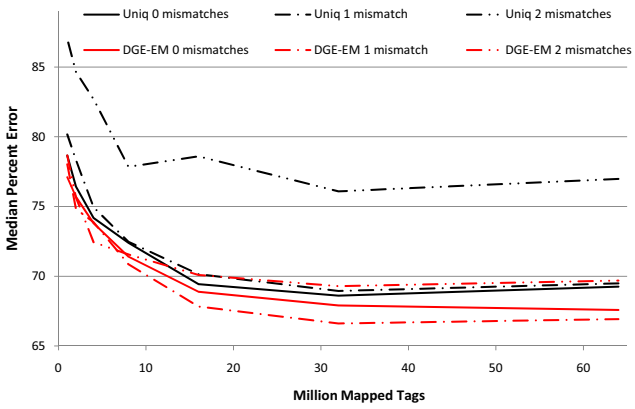


Fig. 3. Median Percent Error of DGE-EM and Uniq estimates for varying number of allowed mismatches and DGE tags generated from the HBRR library 4.

4.3 Comparison of DGE and RNA-Seq Protocols

Figure 4 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 [6] and the well-known Cufflinks algorithm [8]. The cutting probability inferred by DGE-EM is almost the same for all 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 [1], 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.

4.4 Possible DGE Assay Optimizations

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 5(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 5(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

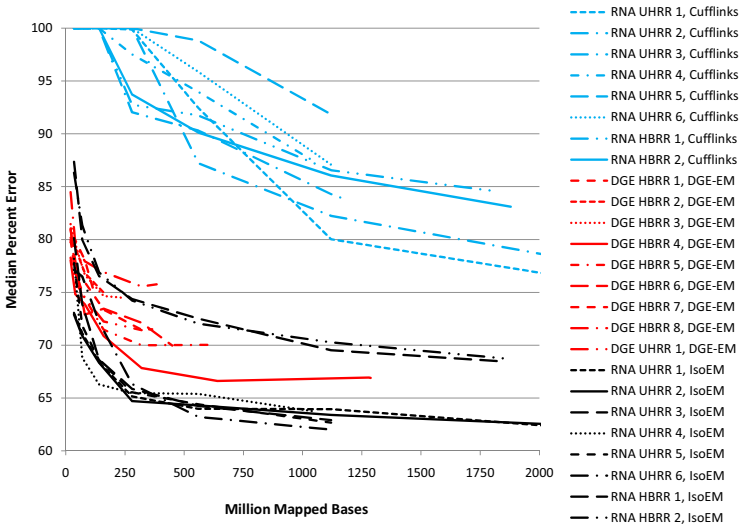


Fig. 4. Median Percent Error of DGE-EM, IsoEM, and Cufflinks estimates from varying amounts of DGE/RNA-Seq data generated from the HBRR MAQC sample.

percentage of genes cut leads to improvements in accuracy. In particular, enzymes like NlaIII (previously used in [9]) 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.

5 Conclusions

In this paper 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

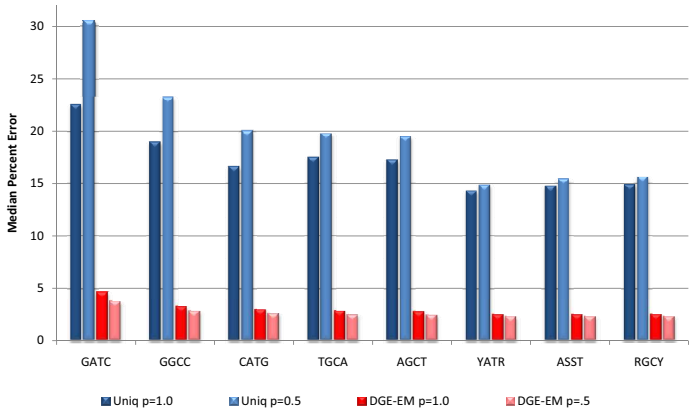
degeneracy of the anchoring enzyme and cutting probability. It would be interesting to experimentally test this hypothesis.

Acknowledgment

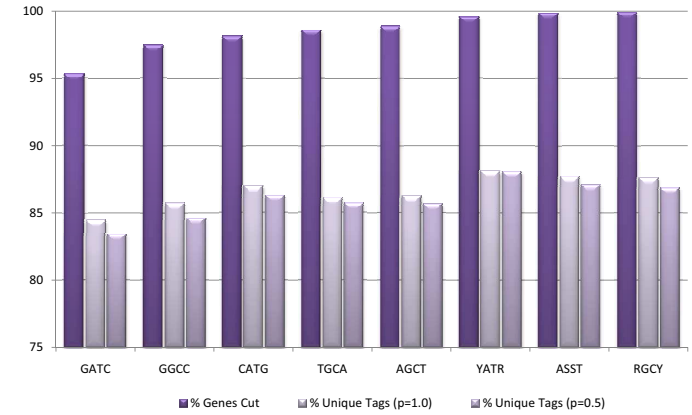
This work has been supported in part by NSF awards IIS-0546457 and IIS-0916948. The authors wish to thank Yan Asmann for kindly providing us with the DGE data from [1].

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.



(a)



(b)

Fig. 5. (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.