Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(1) 2010 19-31

Resolving Isoform Expression using Digital Gene Expression Data

Naomi S. Altman[1*], Qingyu Wang[2], Vishesh Karwa[1] and Aleksandra Slavkovic[1]
[1]*Department of Statistics, Pennsylvania State University, University Park, PA 16802*
[2]*Integrated Biology Program, Pennsylvania State University, University Park, PA 16802*

SUMMARY

In many organisms, alternative splicing increases proteomic complexity by generating multiple mRNA (and protein) isoforms from a single gene. The ability to quantify specific mRNA isoform expression levels is therefore more important to the understanding of biological function than quantifying overall gene expression. Next generation ultra-high throughput sequencing technologies make it possible to measure overall gene expression directly by identifying mRNAs in a sample (RNA-seq and digital gene expression). However, because the technologies typically sequence only short fragments of mRNA, and because mRNA isoforms encoded by the same gene often share substantial sequence regions, quantifying isoform expression from sequencing data requires resolving counts of mRNA fragments into mRNA isoform counts. In this paper, we discuss statistical methods to resolve isoform expression from digital gene expression data using restriction enzyme fragmentation. Methodology for determining the margins of contingency tables are used to deconvolve the fragment counts and infer isoform counts.

Keywords : RNA-seq, DGE, Splice variant, Count data, Marginal distribution, Contingency table, Alternative splicing.

1. INTRODUCTION

A fundamental problem in biological studies is understanding the expression of proteins in a biological sample. mRNA expression is often used as a proxy for protein expression, as mRNA is the intermediary between the genes and the expressed proteins and is more readily measured with high throughput quantification. The set of all possible species of mRNAs from an organism is called the (mRNA) transcriptome. mRNA molecules can be distinguished from other RNA molecules in the sample because mRNAs are marked by a string of "A" bases, called the poly-A tail at what is called the 3' end of the mRNA.

Genes in complex organisms consist of discrete transcribing regions called exons separated by non-transcribing regions called introns. When the gene is expressed, the exons are spliced together. In genes which express multiple isoforms, some exons may be dropped during the splicing operation leading to alternate forms of the resulting mRNA called splice variants or mRNA isoforms (Wang *et al*. 2008). Following the paper by Jiang and Wong (2009) we will use the term isoform to mean a splice variant even if the mRNA species is not known to encode a protein. Since mRNA isoforms expressed by the same gene may have shared or unique exons, quantifying isoform expression requires deconvolution of the exon expression levels.

The advent of ultra-high throughput next generation sequencing (NGS) has made it possible to directly sequence large numbers of RNA fragments obtaining an almost direct measure of what RNA sequences are in a sample (for example, Morozova and Marra 2008). The RNA is fragmented as part of the sample preprocessing because NGS sequences only short pieces of RNA. Fragmentation may be done using mechanical means, which leads to random fragments - a method often called *RNA-seq*. Fragmentation may also

[*]*Corresponding author* : Naomi S. Altman
E-mail address : naomi@stat.psu.edu

be done using restriction enzymes which cut the RNA at selected sequences called restriction sites. The sequences of prespecified length adjacent to the restriction sites are called tags, and the measurement process is called digital gene expression (DGE) profiling. Using either technology, the fragments must be captured by the sequencing instrument, sequenced and then matched to the transcriptome to determine to which exon(s) the fragment belongs. Each fragment that is sequenced is called a read. The numbers of reads belonging to each isoform are the data used for gene expression studies. For several model organisms, exon and mRNA catalogs, consisting of the sequences of exons and mRNA species that have been detected in RNA samples, are now available. This information makes it possible to match reads back to exons, and to map exons to isoforms. However, even for the best understood species, the available information on exons and isoforms is incomplete.

In this paper, we discuss the analysis of isoform expression from DGE data. With very short reads, the tags may not uniquely identify the restriction site. However, with sequence lengths of 30 base pairs or more, this problem is extremely rare (Wall *et al*. 2009). Although the data we analyze in this paper come from shorter sequences, we do not address non-uniqueness here. The sequences in the exon library are searched to determine all of the valid tags, and these are used to identify the reads.

In DGE studies, typically the fragments attached to the poly-A tail of the mRNA are retrieved for sequencing, leading to at most one RNA fragment sequenced for each mRNA in the sample; see Fig. 1. The reads are mapped back to a library of known tags and restriction sites for the organism, which produces reliable data with up to 80% of the tags mapping to known mRNAs (’t Hoen *et al*. 2008). Because each mRNA produces at most 1 read and because of the reliability of the matches, this method can be used to quantify exon expression. However, to accurately quantify isoform expression, the read counts must be attributed to the appropriate isoform.

To illustrate the use of DGE data to resolve isoform counts, we use a data set collected by ’t Hoen *et al*. 2008 (henceforth ’t Hoen) in mouse brain tissue. The very high cutting efficiency of the restriction enzyme in this study makes it possible to resolve counts for isoforms with differing first (3’) tag even when the number of tags sequenced for the gene is moderate. We also show that high cleavage efficiency is not necessarily optimal to resolve isoform expression, particularly if isoforms share the first exon and orientation. Cutting efficiency is controlled by the restriction enzyme digestion, and can be lowered by experimental protocols. In the absence of suitable data, we simulate data with lower cutting efficiency to explore the efficacy of using lower cutting efficiencies in DGE studies used to determine isoform expression. In Section 2 we outline the statistical methodology we use to resolve the isoform data. In Section 3 we analyze 3 genes from the ’t Hoen data to resolve isoform counts. In Section 4 we discuss a simulation study in which we assume that the cutting efficiency is lower than in the ’t Hoen study. In Section 5 we discuss the implications of our results and the directions for future work. The Appendix outlines the bioinformatics work required to prepare the tag and exon databases in a form suitable for the statistical analysis.

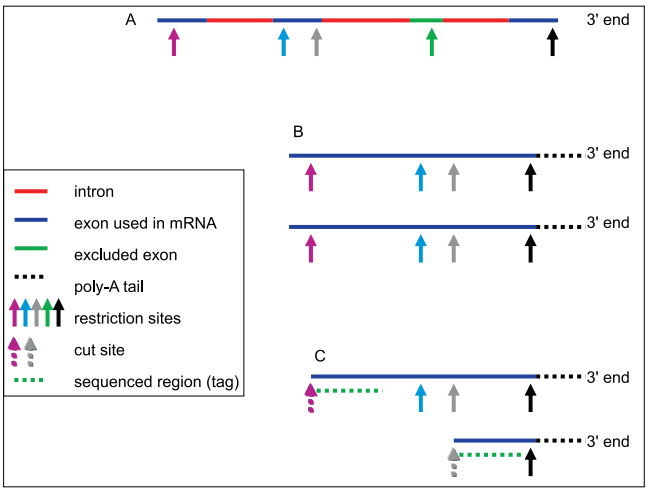

Fig. 1. (A) Genomic region showing exons, introns (regions excluded from mRNA) and potential restriction sites. (B) 2 mRNAs of the same species showing restriction sites prior to fragmentation. Note that the green exon has been excluded. (C) Retrieved fragments prior to sequencing, showing the tag site and sequenced segment.

2. ESTIMATING ISOFORM COUNTS FROM DIGITAL GENE EXPRESSION DATA

Restriction enzyme fragmentation provides data that can be used to assess gene expression by counting the number of reads for each tag. Sequencing proceeds from the cleaved restriction site towards the poly-A tail. Because fragmentation takes place at known sequences, a tag database can be developed from the known exons of the organism yielding a list of possible tags in the

organism, and the exons and genes in which they are found. At this point in time, complete catalogs of exons and mRNAs are not yet available. Our tag and exon database construction for the mouse genome is outlined in the Appendix. While most tags will be represented by zero or very low read counts, tags in some highly expressing genes may yield counts of over 0.1% of the entire sample (for example, Table 3). Data processing after sequencing consists of counting the frequency of each tag found in the sample and resolving these counts into gene or isoform expression. We make the following assumptions:

1. Exons are spliced into an mRNA species sequentially. Exons may be dropped, but their order is fixed.
2. If an exon is in an mRNA species, all restriction sites for that exon are available for the enzyme.
3. All tags can be uniquely mapped to an exon.
4. All retrieved tags are attached to a poly-A tail.
5. The probability that a particular tag is retrieved from an mRNA is a function of the number of restriction sites separating it from the poly-A tail and is homogeneous across genes and isoforms.

Assumptions 3-5 are only approximate. In our analysis, we do not use tags which map to multiple exons. We will use the model of Gilchrist, Qin and Zaretzki (2007) (henceforth abbreviated as GQZ) which satisfies assumption 5 to estimate the retrieval probability of each tag in each isoform.

A tag is observed if it is cleaved by the restriction enzyme, and is the closest cleaved site to the poly-A tail. GQZ postulated a truncated geometric distribution for the detection probabilities. Assuming that all restriction sites are equally likely to be cleaved, site 1 is recovered with probability $\phi_1 = p$ which is the restriction efficiency; site 2 is recovered only if 1 is not cleaved and 2 is cleaved, i.e. with probability $\phi_2 = (1-p)p$ and site z is recovered only if sites 1 through $z-1$ are not cleaved but site z is, with probability $\phi_z = (1-p)^{z-1}p$. GQZ postulated that p should be a function of sample preparation, rather than gene or isoform, and hence should be the same for all RNA species in the sample. Consider 2 adjacent tags on the same exon. Suppose that the tag expression is a mixture of I isoforms, with proportion $\pi_{i\bullet}$ for isoform i. The tag is in position s_i relative to the poly-A tail of isoform i. Then the probability of observing the s_i tag is $\sum_{i=1}^{I} \pi_i \phi_{s_i}$ while the probability of observing the adjacent tag is $\sum_{i=1}^{I} \pi_i \phi_{s_i+1}$. Noting that $\frac{\phi_{k+1}}{\phi_k} = 1-p$ the ratio of the probability of observing adjacent tags on the same exon is $1-p$. Thus, regardless of the isoform expression by the gene, the ratio of counts from adjacent tags on the same exon should be the same.

Since most genes consist of several exons and most exons have multiple tags, there is potentially a large amount of data available to estimate p. Current mRNA annotation libraries such as Aceview (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html; Thierry-Mieg and Thierry-Mieg, 2006) and the Alternative Splicing Database Project (ASD, http://www.ebi.ac.uk/asd/; Stamm *et al.* 2006. Thanaraj *et al*. 2004; Clark and Thanaraj 2002) annotate several possible sites for the poly-A tail in the last exon of some splice variants. Because of this, it is not uncommon for several isoforms to have their first (and most highly detected) tag annotated to the same exon. When these tags are adjacent, the ratio of counts will reflect both p and the isoform abundance. We use the ratio of very high count tags to adjacent tags in the same exon to estimate p in preference to methods such as maximum likelihood because the ratio yields an estimate of p from each exon with sufficiently high tag counts. This makes it possible both to check the GQZ model and to robustify the estimate against annotation of multiple 3' tags to the same exon as well as errors in the exon and tag databases. In the remainder of the analysis, we consider p to be known. When it is not known, it can be estimated along with other parameters of the model.

Let J be the total number of tags that can be observed in any isoform for the gene and let $\pi_{j|i}$ be the probability of observing tag j from a random mRNA of isoform i. The observed tag for any mRNA is the cleaved site closest to the poly-A tail. If the enzyme does not cleave the mRNA at any site, then the mRNA cannot produce a read, so we have to consider unobserved mRNAs to obtain the total count. The probability that an mRNA of isoform i is not cleaved and therefore does not produce a read is $\pi_{(J+1)|i} = 1 - \sum_{j=1}^{J} \pi_{j|i}$. To infer isoform frequency, the probability of nonretrieval must be accounted for. However, the GQZ model suggests that even when the cleavage efficiency is as low as 70%, mRNAs from isoforms with 4 or more restriction sites are cleaved with

probability exceeding 99%. Hence biases introduced by failure to cleave are typically low for most genes.

To illustrate, consider the simple case of a gene which has 3 exons with r, s and t restriction sites respectively and 2 isoforms, one of which includes exons 1 and 2, and the other of which includes exons 2 and 3. The total number of sites is $J = r + s + t$. The observation probabilities of a particular tag in each isoform depends only on the distance to the poly-A tail, which is assumed to be to the left of the first column of the table. Table 1 displays the observation probabilities in tabular form. For this example $\pi_{1|1} = \phi_1 = \pi_{r+1|2}$ and $\pi_{1|2} = 0$.

In general, to estimate isoform abundance from the tag counts, we consider a two-way table with I isoforms in the rows and $J + 1$ tags in the columns, where the $J + 1$ column captures information for mRNAs which were unobserved due to lack of cleavage. We assume that the isoforms are known. Hence, we know which tags can be observed in each isoform, and their probabilities of being observed according to the GQZ model. This yields an unknown $I \times (J + 1)$ joint table of cross-classification of the number of tags n_{ij} for isoform i whose underlying true joint probabilities are $\pi_{IJ} = (\pi_{ij})$. The hypothetical joint table is shown below in Table 2. The marginal probability distributions for isoforms and tags are $\pi_{I\bullet} = (\pi_{i\bullet})$ and $\pi_{\bullet J} = (\pi_{\bullet j})$, respectively. We observe the marginal column counts $T_{\bullet j}$, i.e., tag counts, for the first J columns and wish to infer the marginal isoform counts, $n_{i\bullet}$, which are the row counts. Based on the GQZ model we are also given the $I \times (J + 1)$ matrix of conditional probabilities, $\pi_{J|I} = (\pi_{j|i})$, $i = 1, \ldots, I$, and $j = 1, \ldots, J + 1$.

Arnold *et al*. (1999) and more recently Slavkovic and Fienberg (2004, 2009) show that under certain conditions compatible conditional distributions and one marginal (e.g., $\pi_{\mathbf{J|I}}$ and $\pi_{\bullet\mathbf{J}}$) uniquely identify the joint distribution. Theorem 1 below which is a special case of a more general result due to Slavkovic (2004) shows that in many cases of interest, the row margins of a contingency table are uniquely determined by the column margins, i.e. if the conditional distributions are known within each row and the column marginals are also known, then the row marginal percentages can be inferred. If $J \geq I$ and the rows of $\pi_{\mathbf{J|I}}$ are linearly independent, the row margin of population percentages are uniquely determined by the column margin of population percentages.

Theorem 1. (Slavkovic, 2004) Consider a two-way contingency table and a pair of matrices $T = \{\pi_{J|I}, \pi_{\bullet J})$,

Table 1. A Set of Isoforms Displayed as an Isoform by Site Table

Isoform / tag	exon 1			exon 2			exon 3			not cleaved	Isoform count
	1	. . .	r	$r+1$	. . .	$r+s$	$r+s+1$	. . .	$r+s+t$		
1	ϕ_1	. . .	ϕ_r	ϕ_{r+1}	. . .	ϕ_{r+s}	0	. . .	0	$1-\sum_{j=1}^{r+s}\phi_j$	$n_{1\bullet}$
1	0	. . .	0	ϕ_1	. . .	ϕ_s	ϕ_{s+1}	. . .	ϕ_{s+t}	$1-\sum_{j=1}^{s+t}\phi_j$	$n_{2\bullet}$
tag count	$T_{\bullet 1}$	. . .	$T_{\bullet r}$	$T_{\bullet r+1}$	. . .	$T_{\bullet r+s}$	. . .	. . .	$T_{\bullet r+s+t}$	0	N

Table 2. Contingency Table of Tag by Isoform Counts

Tag / Isoform	1	2	3	...	J	$J+1$	Total
i_1	n_{12}	n_{12}	n_{13}	...	n_{1J}	n_{1J+1}	$n_{1\bullet}$
i_2	n_{21}	n_{22}	n_{23}	...	n_{2J}	n_{2J+1}	$n_{2\bullet}$
⋮	⋮	⋮	⋮	...	⋮	⋮	⋮
i_I	n_{I1}	n_{I2}	n_{I3}	...	n_{IJ}	n_{IJ+1}	$n_{I\bullet}$
Total	$T_{\bullet 1}$	$T_{\bullet 2}$	$T_{\bullet 3}$	...	$T_{\bullet J}$	0	N

If the matrix of conditional probabilities has full rank, and $J > I$, then T uniquely identifies the table of probabilities π_{IJ}, and thus $\pi_{I\bullet}$.

For the current problem, we infer the conditional row probabilities $\pi_{J|I}$ from the GQZ model. The true column marginal probabilities $\pi_{\bullet J}$ are not known, but estimated from the observed marginal counts, $T_{\bullet J} = (T_{\bullet 1} \ldots T_{\bullet \mathbf{J}+1})$ that are distributed Multinomial $(N, \pi_{\bullet \mathbf{J}})$ where N is the number of reads generated by the gene. Notice that we observe $T_{\bullet 1} \ldots T_{\bullet J}$ but not $T_{\bullet J+1}$. The latter is often replaced by count of zero in which case assume that N is known; otherwise we can also estimate N from solving equation (1).

Assuming that the given probabilities are compatible, Theorem 1 ensures there is one solution for the missing marginals. The solution can be found simply by solving an over-determined set of linear equations $A\pi_{\mathbf{I}+} = \pi_{+\mathbf{J}}$ where the matrix A picks out the appropriate conditional probabilities. Alternatively, if we had the expected counts $E(T_{\bullet J})$ for each column of the matrix, we could solve, $AE(\mathbf{n_{I\bullet}}) = \mathbf{E(T_{\bullet J})}$ where $\mathbf{n_{I\bullet}}$ are the marginal isoform counts. However, the variation from the GQZ model, the variation in the marginal counts and the missing marginal count for the RNAs that were not cleaved, introduce error into the system, and we approximate the exact solution by the least squares solution

$$\hat{\mathbf{n}}_{\mathbf{I}\bullet} = (A'A)^{-1}A'\mathbf{T}_{\bullet \mathbf{J}} \qquad (1)$$

to obtain a point estimate of the expected isoform counts. We can estimate the standard deviation of the estimated isoform counts by the sandwich variance estimator

$$\hat{Var}(\hat{\mathbf{n}}_{\mathbf{I}\bullet}) = (A'A)^{-1} A'\hat{Var}\,(T_{\bullet J})\mathrm{A}(A'A)^{-1}$$

where $\hat{Var}\,(\mathbf{T}_{\bullet \mathbf{J}})$ is estimated from the estimated Multinomial distribution for $\mathbf{T}_{\bullet \mathbf{J}}$.

It is illustrative to understand ways in which Theorem 1 can fail for isoform data. Firstly, there can be more isoforms than tags, so that $J < I$. Secondly, if some exons do not have tags, two or more isoforms can differ only in exons without tags. Finally, if there are insufficient data, some tags will have zero counts. Practical implementation of the algorithm uses only the columns with non-zero marginal counts. Let J_n be the number of columns (tags) with non-zero marginal counts. Then to resolve isoform counts, we actually need $J_n \geq I$. Lower restriction efficiency (which leads higher probability of observing tags far from the poly-A tail) and higher total reads sequenced both improve our ability to resolve isoform counts.

3. ANALYSIS OF THE 't HOEN MOUSE DATA

't Hoen *et al.* (2008) extracted mRNA from brain tissue from 4 wild-type and 4 transgenic male mice. Fragmentation was done with the N1aIII enzyme, which cleaves at sites with sequence CATG. 17 bp tags were sequenced so that the total sequence length for the tag is 21 bp. Technical details can be found in the 2008 paper. We downloaded the sequences from GEO (/www.ncbi.nlm. nih.gov/geo) data set GSE10782. As shown in Table 3 the numbers of reads per sample

Table 3. Summary of 't Hoen tag counts. Sample is the GEO sample label. Name is the label of the sample in the Figures and Tables in this paper indicating wild type (WT) or transgenic (Mut).

Sample	Name	Total Reads	Reads Matching known exons rule 1	Reads Matching known exons rule 2	Max reads per tag	Max reads per annotated tag
GSM272105	WT 1	2685418	701828 (26.13%)	720404 (26.83%)	40816	10495
GSM272106	Mut 1	3517977	1108543 (31.51%)	1138058 (32.35%)	52287	14528
GSM272318	WT 2	3202246	915442 (28.59%)	938025 (29.29%)	33081	12434
GSM272319	Mut 2	3558260	1073210 (30.16%)	1098022 (30.86%)	49386	16597
GSM272320	WT 3	2460753	660168 (26.83%)	77294 (27.52%)	19422	12250
GSM272321	–	294909	67952 (23.04%)	69507 (23.57%)	35096	1026
GSM272322	–	651172	179881 (27.62%)	184287 (28.30%)	22653	1931
GSM272323	Mut 3	3142280	889837 (28.32%)	914787 (29.11%)	49532	11562

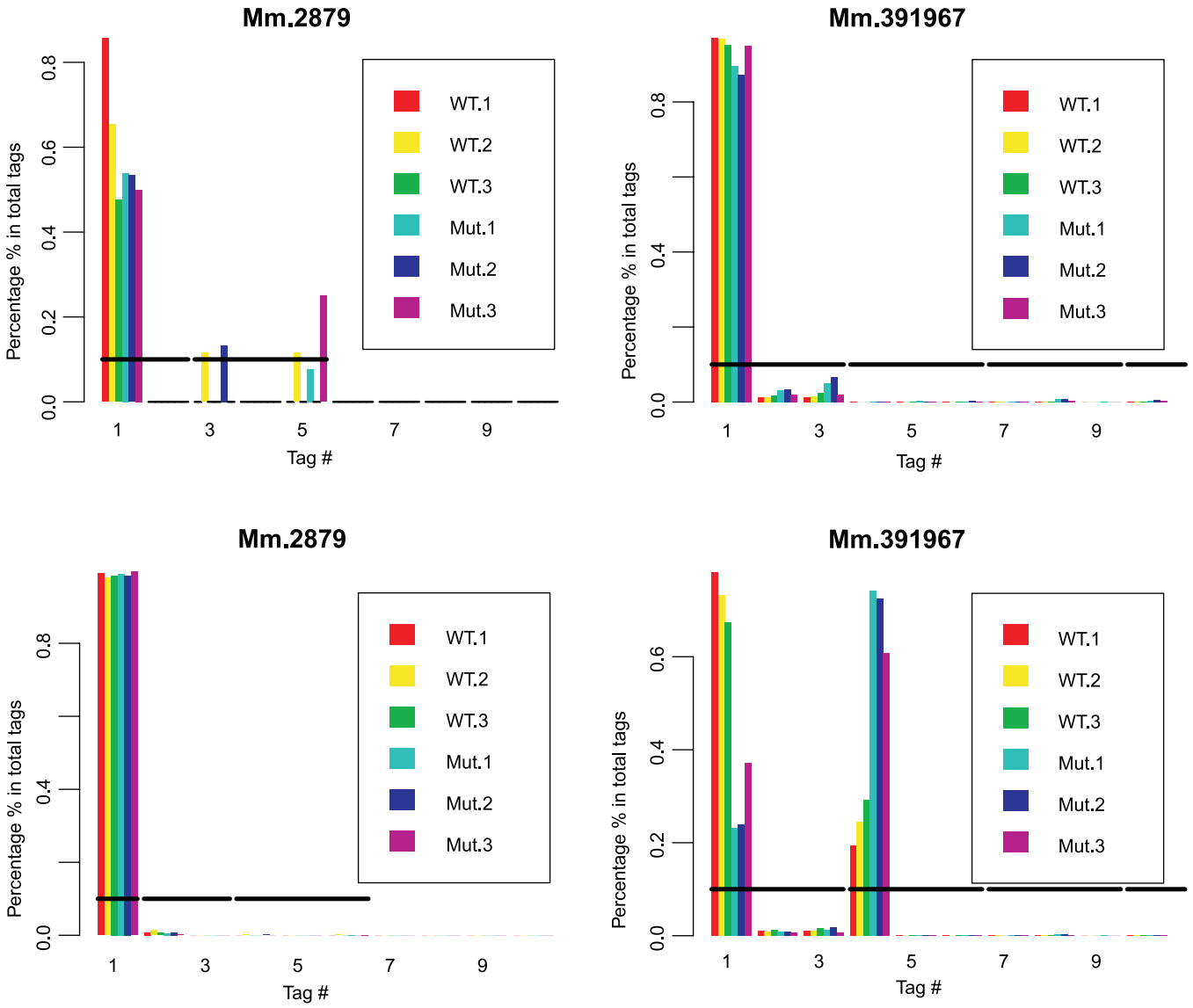

Fig. 2. Percentage of reads for each gene assigned to the tags in various samples showing the first 15 tags per gene. The top row uses the preferred exon database to assign tags to exons. The bottom row uses the maximum count to assign tags to exons. Left column: The maximum count rule provides a more concise summary. Right column: The preferred exon database rule provides a more precise summary.

varied from 2.9×10^5 to 3.5×10^6. We did not use samples GSM272321 and GSM272322 which have substantially fewer reads than the other six samples. Because CATG matches the same sequence on the noncoding strand of the gene, each restriction site potentially produces 2 possible tags: one in the coding (sense) direction and the other in the non-coding (anti-sense) direction. 't Hoen reported that 51% of the detected genes transcribed in both directions. In our study, we used tags only in the sense direction of the gene as annotated in mouse genome mm8 from the UCSC site.

We prepared a database of restriction tags and exons as described in the Appendix. Because of conflicts in the gene and exon annotations from the sources used, we sometimes found multiple tags mapped to the same site. We used 2 different rules to resolve inconsistent location information. These are described in the Appendix. In the 350 most highly expressed genes, the two rules produce very similar counts for all but 12 genes. For 3 of these genes, rule 1 appears to provide a more concise summary; for 7 rule 2 appears to provide a more concise summary and for 2, the summaries differed but neither rule appears

to be concise. Fig. 2 shows bar charts for the 3'-most tags of two of these genes illustrating the difference between the rules.

Table 3 shows the total number of reads sequenced, the total number of reads matched to exons in the direction of gene expression, and the maximum count for a single tag in each sample. We used only tags in the coding direction of the gene, yielding somewhat less than 50% of the usable reads reported by 't Hoen.

The model postulates that for each exon the highest tag count should be the 3' tag. When p is close to 100%, only the first few tags in each isoform will be observed. We used all genes with total tag counts over 1000, and selected exons with at least two tags in which the maximum tag count was the first tag in the exon in the coding direction and was over 500. There were from 385 to 588 usable ratios in each sample. We found multiple instances when peaks occurred in other tags, indicating multiple poly-A sites within an exon or annotation errors. We used the median of the ratio of adjacent tags to estimate p to avoid the need to manually delete these errors. For the 6 samples, the estimate of p was .996 in each sample. Each sample also had several outlying ratios, ranging down to $p = .09$. For example, Fig. 3 shows the estimates of $1 - p$ on the $\log_{10}$ scale from each exon in Sample 2 which had the maximum number of usable ratios, 588. This includes 115 exons for which the estimate of p is exactly 1, due to a zero count for the tag adjacent to the tag with the maximum count.

Isoform databases for mouse genes can be found at Aceview (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html; Thierry-Mieg and Thierry-Mieg, 2006) and the AlternativeSplicing Database Project (ASD, http://www.ebi.ac.uk/asd/; Stamm *et al*. 2006; Thanaraj *et al.* 2004; Clark and Thanaraj 2002). The databases both indicate multiple isoforms for many genes, and possible alternative stop sites within exons. As well, the databases do not necessarily agree. For example, for one gene highly expressed in the 't Hoen study, Mm.155896 (Hnrpa2b1, Heterogeneous Nuclear Ribonucleoprotein A2/B1) Aceview indicates 18 isoforms and 22 exons. ASD indicates 8 isoforms and 18 exons. In both annotations, several isoforms share first and/or last exons and many have a substantial number of shared exons.

A number of features of the data make it difficult to check the fit of the GQZ model. Firstly, the very high

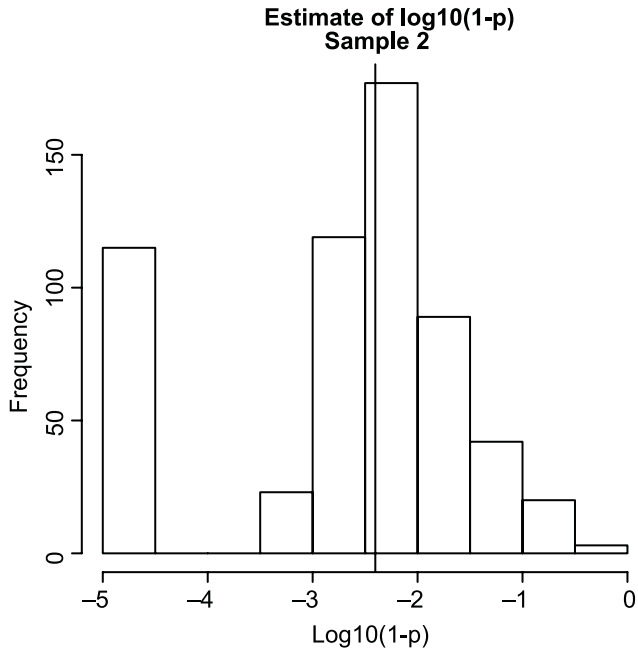

Fig. 3. Histogram of estimates of $\log_{10}(1 - p)$ using exons in Sample 2. The bar at –5 contains the 0 estimates. The vertical line is at the median, which is $1 - \hat{p} = :0039$.

restriction efficiency implies that for each isoform most of the reads are derived from the site closest to the poly-A tail. This makes it difficult to infer which isoforms are present in the data if they have the same 3' tag. Secondly, many genes appear to have multiple isoforms, some of which are expressed at quite low levels. Finally, there appear to be many errors and/or omissions in the exon and isoform databases, leading to peak counts in poorly annotated "putative" exons, peak sites in the middle of exons and possibly detection of anti-sense transcription.

For illustrative purposes, we selected 3 genes that have 3 - 4 dominant count peaks and resolve the tag counts into isoform counts for each of the 6 samples. Fig. 4 shows the read counts for each tag as a percentage of the total counts for the gene. Table 4 has summary information about these genes. This includes the gene name, the number of isoforms in the isoform databases, the number of tags and exons we were able to annotate, the number of obvious "peaks" in tag expression and the number of isoforms that we postulate are present in the data. A "peak" is a tag with over 5% of the reads in at least one sample. Since p appears to be close to 99.6% for all of the samples we postulated an isoform 3' tag for any tag with more than 5 reads in a single sample.

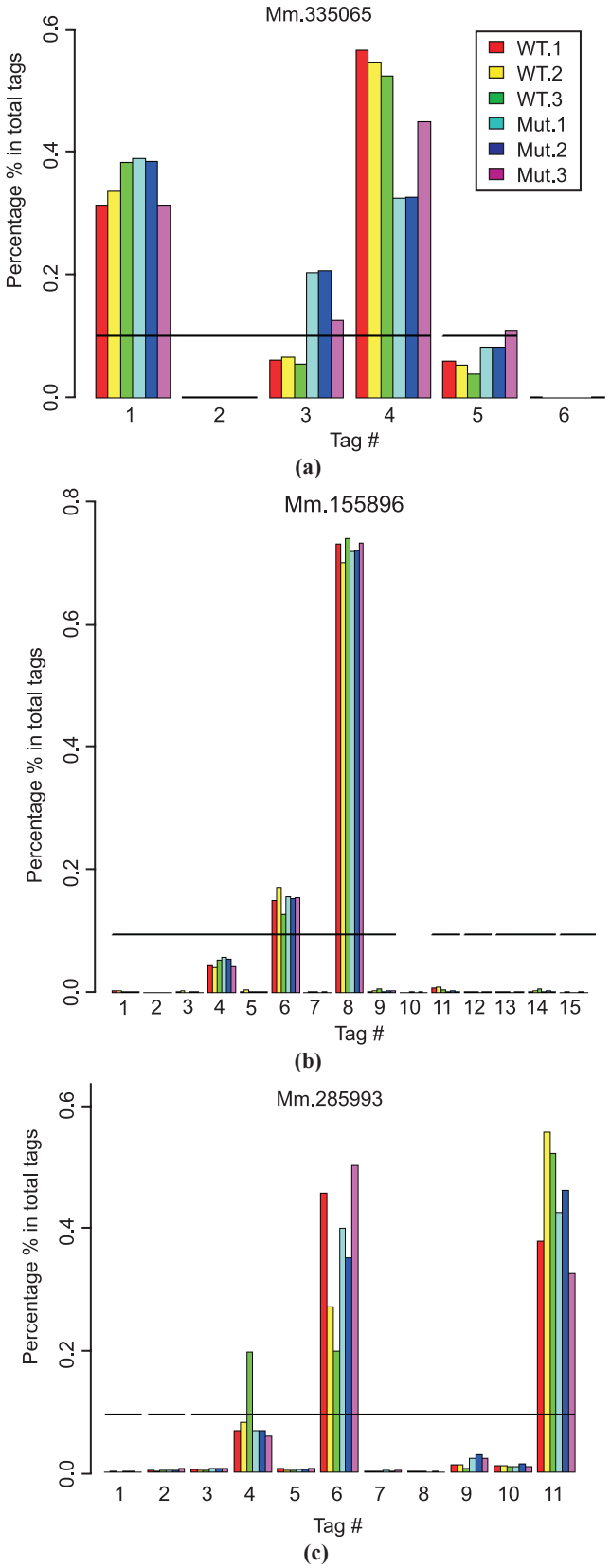

Fig. 4. Bar plots of the percentage of reads at each tag for each sample for 3 genes. The horizontal bar marks annotated exon boundaries.

Table 4. Genes selected for isoform analysis

Gene ID	Mm. 155896	Mm. 285993	Mm. 335065
Gene Name	Hnrpa2b1	Calm1	Nrgn
Aceview Isoforms	18	6	5
ASD Isoforms	8	2	NA
# Tags (J)	16	11	6
# Exons with tags	7	3	2
# Peaks	3	3	4
# Isoforms (I)	9	9	4
Total reads	15244	21921	30404

Because of the high cutting efficiency of the restriction enzyme and information that differed among the 2 isoform databases and the observed data, we were not able to determine which of the known isoforms are likely to be the ones observed in the data. Accordingly for each postulated start tag, we inferred the table of "known" conditional probabilities by selecting the 4 adjacent tags in the direction of exon expression. With $p = .996$, the probability of observing a read from any tag past location 2 in an isoform is about 1.6×10^{-5}. 't Hoen *et al*. postulated that due to the high cutting efficiency in their study, any tags observed for a gene besides the "canonical" tag closest to the poly-A tail must be due to isoform expression. However for these very high throughput studies, in which some tags have counts in the order of 10^4 we can expect to have non-negligible counts for the second tag in the highest expressing isoforms. As well, sample preparation and sequencing errors can introduce a small number of counts of one or two reads.

The tables of counts with the imputed isoform totals for the most abundant isoform and a lower expressing isoform for each gene are shown in Table 5. When available, the lower expressing isoform was chosen to demonstrate discordant up or down-regulation compared to the dominant isoform. To make it simpler to compare counts across samples, which have a differing number of total reads, all counts have been normalized to a pseudo-total of 3×10^6. Notice that for each gene there appears to be significant differential total expression of the genes across the two genotypes (for example, using a Wilcoxon test).

Table 5. Estimated isoform expression for 3 genes showing total gene expression (total), the estimated expression of the most highly expressed isoform (Iso Hi) and another selected isoform (Iso other).

	Mm.155896			Mm.285993			Mm.335065		
Sample	Total	Iso Hi	Iso other	Total	Iso Hi	Iso other	Total	Iso Hi	Iso other
WT.1	1977.0	1545.6	315.2	3701.3	1501.9	41.0	6677.0	3791.7	406.0
WT.2	1815.9	1364.8	333.9	4732.3	2828.4	60.0	5780.8	3165.5	377.2
WT.3	1898.1	1510.5	258.3	4871.8	2732.0	25.0	4972.7	2613.4	274.1
Mut.1	2901.7	2228.6	482.9	3008.9	1370.8	82.2	4248.8	1378.4	868.1
Mut.2	3094.0	2388.8	507.9	2484.1	1230.8	91.2	4123.7	1346.7	854.1
Mut.3	2761.1	2167.3	458.2	2936.1	1024.7	71.2	4137.1	1861.3	524.3

However, in some cases, there is differential expression in the isoforms that is discordant with the total count. None of the isoforms of Mm.155896 are discordant. However, for Mm.285993 while all 3 wild type samples have higher total count than any of the mutant samples, only one of the 9 isoforms has this pattern, while 3 isoforms have the opposite pattern. The other 5 isoforms show a random pattern. Mm.335065 has one concordant isoform and one discordant isoform. Assuming that each isoform codes a different protein, these patterns could have biological significance.

4. SIMULATION OF TAG DATA

For isoform detection and identification it is clearly preferable to have tag counts from several exons from the same gene. For this purpose, it is preferable to have cutting efficiencies much less than 1.0 so that there is a reasonable probability of observing tags other than the tag closest to the poly-A tail. Following the GQZ model, the probability of retrieving the 6^{th} closest tag is 1% when $p = 0.55$. Hence for studies in which it is desirable to observe many sites per isoform, for example, to confirm which exons are in the isoform, very low restriction efficiencies are desirable.

We simulated data for an imaginary gene with 4 exons and 2 restriction sites per exon, assuming median restriction efficiency 70%. To consider the utility of the GQZ model when the cleavage probabilities differ slightly for each restriction site, we generated a "cleavage probability" for each site from Beta(700,700*28/72) which was selected to have mean 0.7 and small variance. This gives 8 probabilities $p_1 \ldots p_8$.

We then generate an isoform by selecting exons. If the isoform includes tags $T_{(1)} \ldots T_{(k)}$ then the probability of observing the tags is $p_{(1)}$, $p_{(2)}(1 - p_{(1)})$, $\ldots p_k \prod_{j=1}^{k-1}(1 - p_{(j)})$ where the subscripts (j) are relative to the position within the isoform, not the gene. For example, the isoform consisting of exons 2 and 4 (denoted isoe2e4) has 4 tags with observation probabilities $\{p_3, p_4(1 - p_3), p_7(1 - p_3)(1 - p_4), p_8(1 - p_3)(1 - p_4)(1 - p_7)\}$ where p_j are the cleavage probabilities. Notice also that the observation probabilities do not sum to 1. There is a positive probability that no tag can be observed.

We used 2 sets of 2 isoforms. The first set consisted of isoe1e2 using exons 1 and 2, and isoe1 using exon 1 only. The second set considered of isoe1e2 using exons 1 and 2 and isoe2e3 using exons 2 and 3. We expect the first set to be more difficult to resolve, as most of the information is in the first few tags, which are the same for isoe1e2 and isoe1. Two sample sizes were used: $n_{1\bullet} = 1000$, $n_{2\bullet} = 500$ and $n_{1\bullet} = 500$, $n_{2\bullet} = 50$. For each sample size 10,000 samples were generated.

Although the data are generated with differing cleavage probabilities, we use the GQZ model to resolve the counts using the empiricial mean cleavage probability $\bar{p}$ in place of p. We used the least squares estimator to obtain estimated isoform margins as described in Section 2. The results are displayed in Fig. 5. Table 6 displays the mean estimated count, the SD of the histogram of estimates from the simulation study and the square root of the mean sandwich variance estimate. The estimated isoform counts appear to be approximately normally distributed about the true mean, even for relatively low expression levels of the

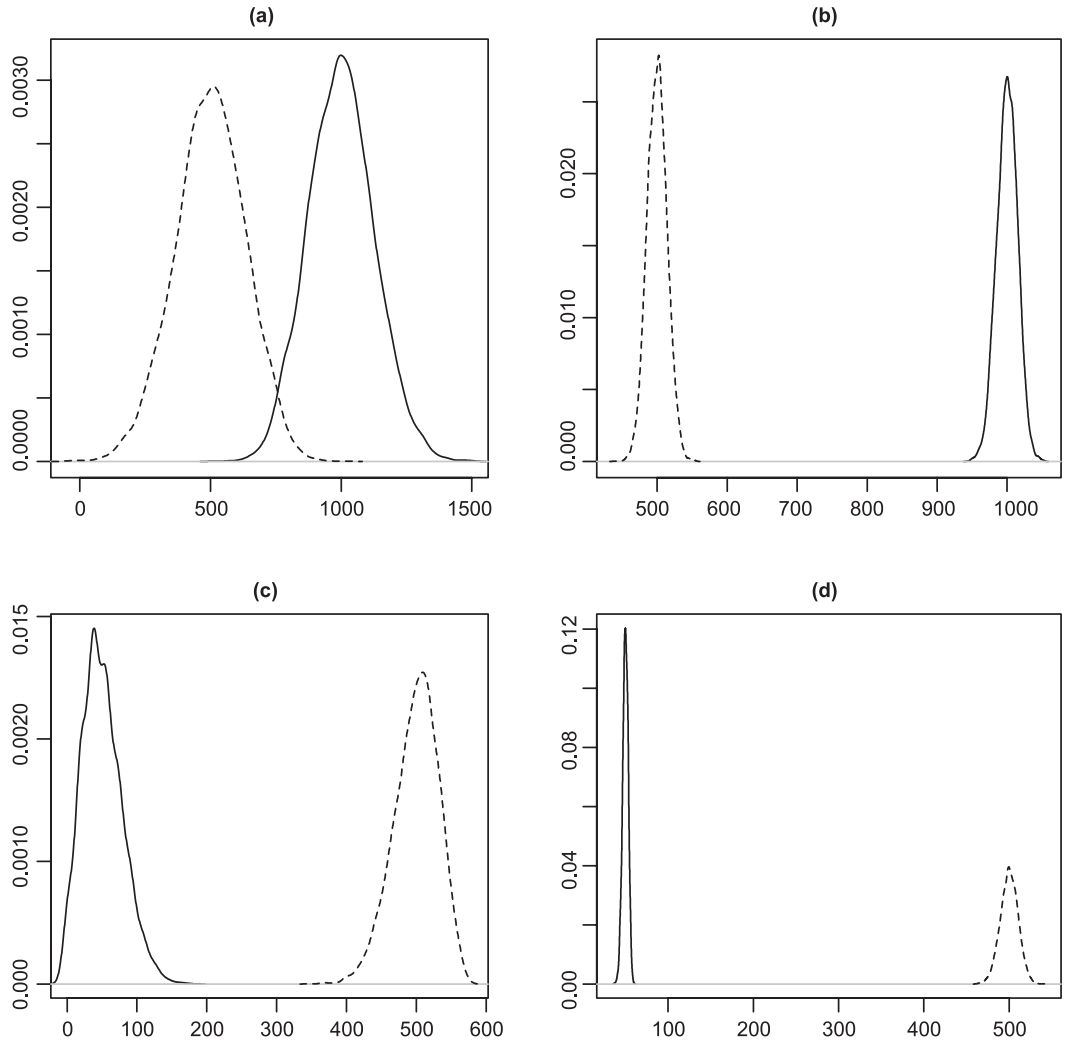

Fig. 5. Estimated counts from simulation study with $p \approx 0.7$. Top row (a. and b.) $n_{1\bullet} = 500$ (left peak) and $n_{2\bullet} = 1000$ (right peak). Bottom row (c. and d.) $n_{1\bullet} = 50$ (left peak) and $n_{2\bullet} = 500$ (right peak). Left hand column (a. and c.) isoforms using exons 1 only (left peak) and exons 1 and 2 (right peak). Right hand column (b. and d.) isoforms using exons 1 and 3 (left peak) and exons 2 and 3 (right peak).

Table 6. Simulation Results including estimated counts, the standard deviation of the estimated count using the simulated counts and the standard deviation of the estimated count using the sandwich estimator.

	1000 & 500 reads		500 & 50 reads	
Isoforms	$\hat{n}_{1\bullet}$ (sim SD) [Sandwich SD]	$\hat{n}_{2\bullet}$ (sim SD) [Sandwich SD]	$\hat{n}_{1\bullet}$ (sim SD) [Sandwich SD]	$\hat{n}_{2\bullet}$ (sim SD) [Sandwich SD]
isoe1e2, isoe1	1000.79	499.47	500.73	49.13
	(126.07)	(135.16)	(28.36)	(32.02)
	[84.14]	[90.18]	[43.31]	[40.74]
isoe1e2, isoe2e3	999.99	499.87	499.92	50.07
	(14.84)	(14.26)	(3.27)	(10.42)
	[23.39]	[23.12]	[7.49]	[12.48]

gene. As expected, the standard deviations of the estimated counts from the study using isoe1e2 and isoe2e3 are smaller than the standard deviations using isoe1e2 and isoe1. The sandwich variance estimates appear to be too high in most cases.

5. DISCUSSION

Very high throughput, very short read technologies provide an open platform for gene expression studies, making it possible to characterize the transcriptome of species for which reference genomic or transcriptomic information is available. Although these technologies are currently more costly than microarrays, they do not require a priori information about sequences expected in the sample. Hence they are ideal for exon and isoform discovery.

Since the first few studies were published using high throughput sequencing for differential expression analysis most investigators have used random fragmentation (shotgun sequencing) for transcriptome sequencing (e.g. Marioni *et al*. 2008, Mortazavi *et al.* 2008, Sultan *et al*. 2008). Random fragmentation of the transcripts allows direct observation of reads that cross exon boundaries, and hence provides direct observation of novel splicing events. As well, random fragmentation can extend transcriptome information by yielding information about mRNAs derived from regions close to the known transcriptome. However, because each transcript can provide multiple fragments, random fragmentation provides an indirect measurement of gene expression. The expression level of different genes in the same sample cannot be compared. Also, due to the possibilities that different mRNA isoforms might express under different biological conditions and these isoforms might generate different mean numbers of fragments, random fragmentation can lead to misleading conclusions about differential expression. Finally in random fragmentation studies, a large percentage of the reads cannot be mapped to the genome. Without more information about how this percentage varies among samples, the accuracy of the measures of gene expression cannot be determined. Jiang andWong (2009) provides a computational method for inferring isoform expression with this type of data.

Digital gene expression analysis, which fragments the transcripts using restriction enzymes combined with retrieval of tags attached to the poly-A tail of the transcript yields less novel information about the transcriptome. Splice junctions cannot be directly observed. On the other hand, tags not annotated to the known exons can still be annotated to putative noncoding regions of the genome, providing information about novel transcription events. For example, 't Hoen found that about 8% of their reads mapped to mitochondrial RNA and another few percent mapped to regions of the genome with no evidence of transcription. A large percentage of the reads mapped to the genome.

We have shown that restriction enzyme sequencing can be used to resolve gene expression at the isoform level for isoforms with moderate expression levels. In the 't Hoen study, enzyme digestion was allowed to continue to near completion, which allows for very accurate detection of isoform 3' tags. Because of the high cutting probability, we are principally able to distinguish among isoforms with different first tags. We demonstrated several examples in which the overall gene expression differed between two mouse populations and differences in isoforms included both up and downregulation in the same gene. However, in these data, differentiating between isoform which share the 3' tag is not feasible.

We demonstrated using a simulation study that the tag detection model of GQZ provides a useful model for tag detection even if the cutting probability varies somewhat among restriction sites. As well, we showed that the marginal count data can be resolved into isoform counts with high accuracy when the cutting probability is lowered to allow detection of multiple tags per isoform, even for moderately expressing genes with a few hundred reads, although it is still difficult to resolve isoforms which share 3' tags. This means that DGE can be used to determine differential expression at both the gene and isoform level. More work needs to be done to resolve the variation of the estimates. The possibility of simultaneous modeling of isoform counts for many genes along with the cutting probability suggests the use of Bayesian modeling to improve the estimation procedure.

Finally, DGE can be used with very short read lengths, so that larger samples can be processed. Although the 17 bp read length used SAGE sequencing (Velculescu. 1995) and in the 't Hoen study leads to many tags which match multiple sites, Wall *et al.* (2009) found that for many organisms sequences of lengths 25 bp and above are close to unique. Counting the 4 bp restriction site, this means that 21 bp reads are sufficient for studies using restriction enzyme fragmentation.

Because the counts can be resolved into isoform counts, at most one read is recovered per transcript and differences in detection probability are accounted for, DGE can be used both for detecting differential expression among genes and isoforms and for detecting differences in expression between different genes in the same sample. On the other hand, exons which do not have restriction sites cannot be detected and hence not all isoforms can be identified.

Finally, while gene expression via random fragmentation requires matching the reads to the genome or transcriptome and then accumulating read counts over regions, gene expression via restriction enzyme fragmentation requires only matching a much smaller catalog of restriction tags. This reduces the computational complexity of accumulating the tag counts. While random fragmentation RNA-seq has recently been used in many studies, we feel that DGE has much to offer, particularly for species with well-characterized transcriptomes.

6. APPENDIX: CONSTRUCTING THE TAG DATABASE

We used a custom tag database provided by Illumina Incorporated based on mouse genome mm8 from the UCSC site http://genome.ucsc.edu/ and mouse mRNA transcriptome based on refseq, mRNA and ESTs found in GenBank as of November 2006 and exon Unigene version Mm159 to annotate the tags with gene identifiers. Although 't Hoen reported that 51% of the genes transcribed in both forward and reverse direction, we selected only the tags in the direction indicated in the gene annotation. We removed tags with no annotation. Finally, we matched the tags for each gene to the known exon locations using the mm8 exon database.

In all, there are 844316 unique sequence tags combined over the 8 samples and about 20000 unique tags for each sample. A series of custom C routines were written to annotate all tags and identify transcriptomic tags. We annotated all sequenced tags in 8 samples using our annotation database. To resolve conflicts in which multiple tags were annotated in the same position by different sources, we developed two different rules for resolving conflicts. Firstly, we resolve in favor of database reliability – we assumed that transcripts from refseq are more reliable than those from ESTs and tags mapped to transcripts from ESTs are more reliable than those from the mRNA database. Finally, when tags are mapped to the same position within a single database, we chose the tag with the highest count. The second rule is based only on tag counts. For any restriction site, the tag with the highest reads which is annotated to the site is considered the best annotation.

ACKNOWLEDGEMENTS

The authors thank Illumina Incorporated for providing the mouse tag database. Naomi Altman's work is partially supported by NSF Grants OCE- 0825979 and GEPR-0820729. Qingyu Wang's work is supported by NHGRI grant HG02238. Vishesh Karwa and Aleksandra Slavkovic are supported by NSF Grant SES-0532407. The authors would like to thank Dr. Rudolf Schilder for critically reading the manuscript.

REFERENCES

't Hoen, P.A., Ariyurek, Y., Thygesen, H.H., Vreugdenhil, E., Vossen, R.H., de Menezes, R.X., Boer, J.M., van Ommen, J.M. and den Dunnen, J.M. (2008). Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res.*, **36**, e141.

Arnold, B., Castillo, E. and Sarabia, J. (1999). *Conditional Specication of Statistical Models*. Springer Verlag.

Clark, F. and Thanaraj, T.A. (2002). Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Human Mol. Genet.*, **11**, 451-464 .

Gilchrist, M.A., Qin, H. and Zaretzki, R. (2007). Modeling SAGE tag formation and its effects on data interpretation within a Bayesian framework. *BMC Bioinformatics*, **8**, 403-419.

Jiang, H. and Wong, W.H. (2009). Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026-1032.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008). RNA-seq an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509-1517.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5,** 621-628.

Morozova, O. and Marra, M.A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255-264.

Slavkovic, A.B. (2004). Statistical Disclosure Limitation Beyond the Margins: Characterization of Joint Distributions for Contingency Tables. Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University.

Slavkovic, A.B. and Fienberg, S.E. (2004). Bounds for Cell Entries in Two-Way Tables Given Conditional Relative Frequencies. In: *Privacy in Statistical Databases*, Springer, Heidelberg.

Slavkovic, A.B. and Fienberg, S.E. (2009). Algebraic geometry of 2×2 contingency tables. In : Algebraic and Geometric Methods in Statistics, P. Gibilisco, E. Riccomagno, M.P. Rogantin, H.P. Wynn (eds.), pages 63-81, Cambridge University Press, UK.

Stamm, S., Riethoven, J.J.M., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N.L. and Thanaraj, T.A. (2006). ASD: A bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, **34**, D46-D55.

Thanaraj, T.A., Stamm, S., Clark, F., Riethoven, J.J.M., Le Texier, V., and Muilu, J. (2004). ASD: The alternative splicing database. *Nucleic Acids Res.*, **32**, D64-D69.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H. and Yaspo, M.L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956-960.

Thierry-Mieg, D. and Thierry-Mieg, J. (2006). AceView: A comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.*, **7(Suppl 1)**, S12.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995). Serial analysis of gene expression. *Science*, **270**, 484-487.

Wall, K., Leebens-Mack, J.H., Barakat, A., Chanderbali, A.S., Landherr, L.L., Altman, N., Carlson, J.E., Ma, H., Miller, W., Schuster, S., Soltis, D.E., Soltis, P.S. and dePamphilis, C.W. (2009). Comparison of next generation sequencing technologies for de novo transcriptome characterization. *BMC Genomics*, **10**, 347-356.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P. and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470-476.