BIOINFORMATICS

Vol. 20 no. 17 2004, pages 2973–2984
doi:10.1093/bioinformatics/bth342

EST clustering error evaluation and correction

Ji-Ping Z. Wang[1,2,*], Bruce G. Lindsay[2], James Leebens-Mack[3], Liying Cui[3], Kerr Wall[3], Webb C. Miller[3] and Claude W. dePamphilis[3]

[1]Department of Statistics, Northwestern University, Evanston, IL 60208, USA, [2]Department of Statistics and [3]Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

Received on February 23, 2004; revised on May 13, 2004; accepted on May 18, 2004
Advance Access publication June 9, 2004

ABSTRACT

Motivation: The gene expression intensity information conveyed by (EST) Expressed Sequence Tag data can be used to infer important cDNA library properties, such as gene number and expression patterns. However, EST clustering errors, which often lead to greatly inflated estimates of obtained unique genes, have become a major obstacle in the analyses. The EST clustering error structure, the relationship between clustering error and clustering criteria, and possible error correction methods need to be systematically investigated.

Results: We identify and quantify two types of EST clustering error, namely, Type I and II in EST clustering using CAP3 assembling program. A Type I error occurs when ESTs from the same gene do not form a cluster whereas a Type II error occurs when ESTs from distinct genes are falsely clustered together. While the Type II error rate is $<$1.5% for both 5′ and 3′ EST clustering, the Type I error in the 5′ EST case is $\sim$10 times higher than the 3′ EST case (30% versus 3%). An over-stringent identity rule, e.g., $P \geq 95\%$, may even inflate the Type I error in both cases. We demonstrate that $\sim$80% of the Type I error is due to insufficient overlap among sibling ESTs (ISO error) in 5′ EST clustering. A novel statistical approach is proposed to correct ISO error to provide more accurate estimates of the true gene cluster profile.

Availability: We have automated the methods developed in this paper in a web-based software ESTstat at http://cwdg5.bio.psu.edu/eststat.

Contact: jzwang@northwestern.edu

Supplementary information: http://cwdg5.bio.psu.edu/eststat

1 INTRODUCTION

Expressed sequence tag (EST) sequencing is a cost-effective way to survey the expressed portions of the genome. The rapidly growing EST database has become an invaluable tool for novel gene discovery (Adams *et al*., 1992, 1993), gene mapping (Khan *et al*., 1992), genome annotation, single nucleotide polymorphism (SNP) discovery (Hu *et al*., 2002; Picoult-Newberg *et al*., 1999) and alternative splicing detection (Lee, 2003; Heber *et al*., 2002; Xu *et al*., 2002; Modrek and Lee, 2002; Modrek *et al*., 2001). Major efforts have been made in clustering of EST data from one or multiple cDNA libraries in many species, including UniGene (Boguski and Schuler, 1995; Schuler *et al*., 1996), the TIGR Gene Index (Liang *et al*., 2000), the Sequence Tag Alignment and Consensus Knowledgebase (STACK) (Miller *et al*., 1999; Christoffels *et al*., 2001) (http://www.sanbi.ac.za/) and IMAGEne (http://image.llnl.gov/). These index systems provide convenient Web interfaces to search for genes or gene families of interest and to investigate gene expression patterns by tissue types.

In contrast to the diverse applications of the sequence information from ESTs, the information on gene expression level from EST clustering has yet to be fully exploited. Let X_j be the number of ESTs from the j-th gene (to be called siblings), then X_j directly reflects the relative expression level of the underlying gene in the cDNA library. Therefore, X_js can be used to detect differential gene expression if the cDNA library is non-normalized (Audic and Claverie, 1997; Stekel *et al*., 2000). If we further define $n_i = \sum_j I(X_j = i)$ as the total number of genes with i ESTs in the sample, then $\mathbf{n} = (n_1, n_2, \ldots)$ is a sufficient statistic for the transcript abundance distribution in the cDNA library (note: if the cDNA library is normalized, then the **n** data will not reflect the true gene expression level in the underlying tissue). Here we call **n** the *true gene cluster profile*.

Our research is motivated by a desire to address a series of questions about the cDNA library that require an accurate estimate of the true gene cluster profile, **n**, for legitimate statistical inferences. For example, given an EST set from a specific tissue at a specific developmental stage, we would like to estimate the total number of expressed genes in this tissue at the developmental stage. We would like to know how many

*To whom correspondence should be addressed.

genes have been sampled and the sequencing redundancy rate. While generating EST sequences, we would like to estimate the transcript redundancy in a cDNA library, and use this estimate to determine the most cost-efficient point to stop random EST sequencing from that library. All of these considerations require an accurate estimate of the true gene cluster profile, **n**. For these applications, the EST data are not restricted to be from a non-normalized cDNA library as long as the cDNA clones are randomly sampled and sequenced.

EST clustering usually refers to the entire process of identifying and assembling sibling ESTs, which can generate the gene cluster profile data as needed. Both UniGene and IMAGEne use BLAST-based procedures with low stringency to assemble sets of ESTs from closely related genes, and are therefore not suitable for our purpose. Here, we used CAP3 assembling program (Huang and Madan, 1999) to provide a quick and simplified clustering to illustrate the study of EST clustering error and error correction. CAP3 assembles ESTs from the same gene under more stringent criteria than the BLAST-based approaches, and was shown (Liang *et al*., 2000) superior to TIGR assembler (Sutton *et al*., 1995), and Phrap (Green, 1996 http://bozeman.mbt.washington.edu/phrap.docs/phrap.html) in its ability to distinguish gene family members while tolerating sequencing error. The efficiency and reliability of CAP3 can also be improved by incorporating basecall quality data in the clustering algorithm. Although our systematic investigation of clustering error in this paper is based on CAP3 assembling alone, this analysis could be extended to other clustering pipelines such as STACK_pack (Miller *et al*., 1999; Christoffels *et al*., 2001) (see Discussion section).

The accuracy of EST clustering is affected by various error sources, such as sequencing error, contaminant sequences and the products of chimeric splicing. Regarding the clustering outcome, EST clustering error can be simply classified into two types, which we will call Type I and Type II through analogy with statistical hypothesis testing theory (Burke *et al*., 1999). The Type I error is a mis-separation error where ESTs from the same gene are falsely separated into two or more clusters (including singletons). The Type II error is a mis-joining error where two or more non-sibling ESTs are clustered together. Burke *et al*. (1999) first described EST clustering error structure in terms of Type I and II errors and claimed that the *d2_cluster* program has upper bound error rates of 0.4 and 0.8% for Type I and II errors, respectively. However, we question these error rate estimates, because regardless of clustering algorithm, the error rate is jointly determined by the quality of the EST data and the clustering stringency. Type I and II errors, as we will show in this paper, are correlated; minimizing one may inflate the other.

There exist substantial EST data for many organisms that lack full genome sequences or genome annotations. For these EST sets, generation of the gene cluster profile data **n** mainly relies on EST clustering programs such as CAP3. For convenience, a cluster or contig from the assembly program here will be called a 'unigene'. Because of the two types of errors, multiple unigenes can represent the same gene (Type I), and a single unigene can involve non-sibling ESTs (Type II). The sibling ESTs are identified if high similarity exists between them, thereby the clustering result is closely related to the stringency of parameter setting in the clustering or assembling algorithms. Furthermore, different types of clustering errors will depend on the parameter setting in different fashions as to be demonstrated. For CAP3, the two main parameters are the overlap length O and percentage identity P in the overlapped region. We found that the clustering error rate is relatively insensitive to the 'overlap length' threshold when it is set within the range from $O = 25$–45 bp (see Discussion section). Hence, one goal here is to investigate the relationship between the stringency of the identity rule in CAP3 and the magnitude of Type I and Type II errors. The results provide insights to optimal choice of the stringency rule. We show that ISO error accounts for the majority of Type I error rate in the 5′ EST clustering case. A novel statistical method is proposed to correct for the ISO error, so as to generate a better estimate of the true gene cluster profile data. Extensions to other clustering procedures and applications of our methods are discussed.

2 METHODS

2.1 Evaluation of Type I and II errors

We performed EST preprocessing before clustering to reduce the errors due to contaminant sequences and sequencing errors. The SEQCLEAN program from TIGR (http://www.tigr.org/tdb/tgi/software) was used to trim vector, poly(A) tail and low-quality bases for each EST at the default settings. EST sequences shorter than 100 bp after trimming were discarded. After preprocessing, we used CAP3 to cluster ESTs with various stringency criteria (sequence identity $P = 75$, 80, 85, 90, 95, 97.5). A range of overlap lengths was initially examined ($O = 25, 30, 35, 40, 45$), but the findings were not sensitive to these choices (see Discussion section), so $O = 40$ was used for clustering experiments.

We have defined the true gene cluster profile, $\mathbf{n} = (n_1, \ldots, n_i)$, where n_i is the actual number of genes with i ESTs in the sample, and $n_+ = \sum_i n_i$ is the total number of genes. Owing to clustering error, the n_is are not directly observed. However, **n**, and the accuracy of estimates of **n**, can be directly obtained for organisms that have well-annotated genomes. Here we took advantage of the annotated *Arabidopsis thaliana* genome, using BLASTn to align the ESTs to the genome and cluster the ESTs that matched the same locus at the E-value threshold 10^{-10}. If one EST has multiple matches on the genome, the most significant locus (least E-value) was recorded. In most cases where multiple matches existed in this analysis, the E-value of the top match was essentially 0 ($\ll 10^{-10}$), much smaller in scale than the rest. Therefore we

felt confident that the top match recorded was the true locus, and this issue should not be problematic to the clustering error analysis and correction method proposed in this paper. The gene cluster profile generated by this method will be regarded as **n**.

Correspondingly, we define the *observed gene cluster profile* as $\mathbf{c} = (c_1, \ldots, c_i, \ldots)$ where c_i counts the clusters with i ESTs that is produced by a clustering program such as CAP3 without additional correction. The discrepancy between **n** and **c** is a direct quantitative measure of the clustering error. We also compared the effect of different stringency parameters on the error rates.

2.2 ISO error and simulation

An ISO error occurs if the sibling ESTs are separated into different clusters because they do not meet the specified overlap threshold , e.g. $O \geq 40$ bp. Figure 1 illustrates a situation where only two of four 5′ sibling ESTs overlap. Consequently, one gene would be interpreted as three unigenes, each representing a different portion of the complete cDNA. ISO error may occur in 3′ EST clustering, but is especially problematic in 5′ EST clustering because transcripts found in cDNA libraries are usually truncated at their 5′ ends to different extents, whereas transcripts with truncated 3′ ends are typically removed in the cDNA library building process. Therefore ISO error discussion here is focused on the 5′ EST case.

Suppose one gene is represented by X ESTs in a sample. After clustering, these ESTs could end up in one or more clusters due to ISO error. Let $Y_1 \geq Y_2 \geq \cdots \geq Y_k$ be the EST counts in these clusters. For example, in the situation shown in Figure 1, we have $Y_1 = 2$, $Y_2 = 1$, $Y_3 = 1$. In our analyses of many EST data sets, we rarely observed a true EST cluster that was separated into more than four subclusters due to insufficient overlap. We therefore write the clustering outcome in a four-dimensional vector $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$ with $Y_1 \geq Y_2 \geq Y_3 \geq Y_4$. The ISO error distribution is defined in a conditional form as $P(\mathbf{Y} = \mathbf{y}|X = x)$ where X denotes the true number of ESTs in the sample that represent a particular gene, i.e. $x = \sum_i y_i$, and $\mathbf{y} = (y_1, y_2, y_3, y_4)$ represents the clustering outcome. Back to the above example, given $X = 4$, we seek the conditional probability of observing the following outcomes: one cluster with all four ESTs (no error): $\mathbf{y} = (4, 0, 0, 0)$; two subclusters with counts $\mathbf{y} = (3, 1, 0, 0)$ or $\mathbf{y} = (2, 2, 0, 0)$; three subclusters with counts $\mathbf{y} = (2, 1, 1, 0)$ and four subclusters with counts $\mathbf{y} = (1, 1, 1, 1)$. Note that under this definition, $P[\mathbf{y} = (1, 0, 0, 0)|x = 1] = 1$ since a singleton cannot be broken into subclusters.

The ISO error distribution $P(\mathbf{Y}|X)$ is related to three factors as seen in Figure 1: (1) the complete cDNA (mRNA) length L_m; (2) the EST length L_E; and (3) the EST 5′ end location S. The marginal distribution of EST length, denoted as $F(L_E)$, is usually determined by sequencing technology. The marginal distribution of cDNA length $F(L_m)$ may be organism-specific. The distribution of the EST 5′ end location given a cDNA length will be written as $F(S|L_m)$; this distribution is determined by experimental conditions or biochemical mechanisms, and defines the quality of the cDNA library. If these three distributions were known, then the ISO error distribution $P(\mathbf{Y} = \mathbf{y}|X = x)$ could be determined under some reasonable assumptions about the sampling process. Here, we used training data to estimate these three distributions.

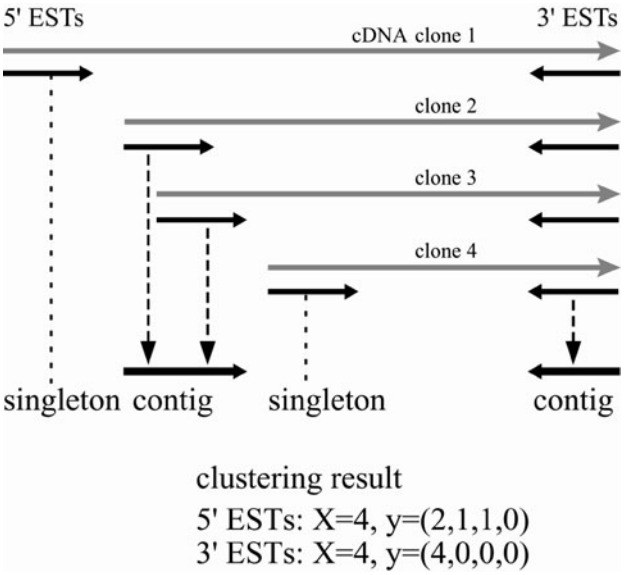

Fig. 1. A hypothetical gene that is represented by $X = 4$ clones in an EST set. **y** is the clustering result given X as defined in the text. The length of the cloned DNAs will vary due to transcript fragmentation in the RNA extraction and library building process. In the most commonly used library construction protocols, only transcripts with intact polyadenylated 3′ ends will be included in the cDNA library. As a result 3′ ESTs usually align to the 3′ end of the cDNA, but 5′ ESTs will often align to different regions of the complete cDNA.

The training data consisted of 8095 complete cDNAs of *A.thaliana* from Riken at http://www.gsc.riken.go.jp (Seki *et al.*, 2002), and 48 827 5′ ESTs of *A.thaliana* from NCBI dbEST (May 2002). The 8095 complete cDNAs were first clustered using CAP3 (at $P = 95\%$, $O = 40$) and 6759 contigs were produced. Some 'complete cDNAs' were probably incomplete because they were much shorter than the corresponding contigs. Therefore, the contig length was expected to give a better approximation of the mRNA length distribution $F(L_m)$. We then clustered the complete cDNA contigs together with the 5′ ESTs. Only the clusters containing complete cDNAs were selected. The distribution of the contig length of selected clusters can be regarded as an empirical version of $F(L_m)$. Similarly, the EST 5′ end locations along these contigs can be used to approximate $F(S|L_m)$. These empirical distributions are shown in the Supplementary materials.

To find $P(\mathbf{Y}|X = x)$, the probability of clustering outcome **Y** given that the specific cluster consists of x ESTs for

$x = 2, 3, \ldots$, one can simulate the true sampling process as follows: for each $x = 2, 3, 4, 5, \ldots$,

(1) Sample a complete cDNA length L_m from $F(L_m)$.
(2) Sample x independent 5′ end locations from the EST clusters in the training data with contig length $L_m \pm 50$ [*i.e.* sample from $F(S|L_m)$].
(3) Sample x independent EST lengths from $F(L_E)$ [we are assuming L_E is independent of (S, L_m)].
(4) Align the x ESTs along the cDNA according to the sampled 5′ end locations and EST lengths. If two ESTs overlap more than 40 bp, cluster them together. Record the clustering outcome as $\mathbf{y}|x$.
(5) Repeat the above four steps 5000 times, assigning the outcomes as $\mathbf{y}_i$, i=1,2, … ,5000. Calculate the empirical probability $P(\mathbf{Y} = \mathbf{y}|x) = \sum_{i=1}^{5000} I(\mathbf{y}_i = \mathbf{y}|x)/5000$ for every distinct $\mathbf{y}|x$ observed.

In step 2, we sample from $F(S|L_m \pm 50)$ rather than $F(S|L_m)$ in order to avoid small sample effects. By assuming that the conditional distribution $F(S|L_m)$ is similar in the neighborhood of the chosen L_m, we can sample the start positions from those clusters with contig length in the neighborhood of the sampled L_m and avoid the possibility of re-sampling the same EST over and over in a small cluster.

3 RESULTS

Arabidopsis thaliana EST sets were used for this investigation because we are able to take advantage of a well-annotated genome and a large set of full-length cDNAs (Seki *et al*., 2002). In addition, *Arabidopsis* and perhaps all plant species exhibit a low rate of alternative-splicing compared to mammalian transcripts (Haas *et al*., 2002). As we discuss below, inferring the representation of genes in EST sets is problematic when alternative-splicing is common. In addition to the typical sources of Type I and II error, ESTs from distinct splice variants may be incorrectly joined (Type II error) when genes are defined as unique transcripts, and ESTs from distinct splice variants may be incorrectly separated (Type I error) if genes are defined as locations in the genome.

3.1 Type I and Type II errors

Two libraries were analyzed to evaluate Type I and II error rates: (1) a 3′ flower bud EST set and (2) a 5′ EST set derived from all above-ground organs, 2–6 weeks after planting (Asamizu *et al*., 2000). To obtain the true gene cluster profile **n**, we aligned each EST to the *A.thaliana* genome (TIGR Version 4/17/2003) using BLASTn. Each EST that matched a specific location on the *Arabidopsis* genome was identified from the corresponding annotation file. The **n** elements were obtained from the exact count of loci that had i ESTs. The clustering errors for these two data sets are summarized in Tables 1, 2 and 3 and discussed below.

3.1.1 A.thaliana 3′ Flower bud EST set The flower bud EST set obtained from GenBank dbEST (UniLib # = 17697) included 5827 3′ ESTs of which 5710 were retained after sequence cleaning. Among these, only 5499 ESTs matched annotated loci on the genome. The direction for 451 of 5499 (8.2%) ESTs contradicted the genome annotation, implying that 8.2% of the cDNA inserts were inverted if the genome annotation was correct. Seven ESTs had no significant match on the genome and 204 matched loci where no gene model had been predicted. To evaluate the clustering stringency, we compared the true expression profile, **n**, for the remaining 5048 (= 5499 − 451) verified 3′ ESTs with the observed expression profile, **c**, inferred from CAP3 clustering with a range of identity parameters.

Overall, the total number of inferred unigenes, c_+, was closest to the true number of genes represented in the EST set, n_+, when the identity parameter was set at 90% (Table 1). The singleton count c_1 is usually a sensitive indicator of the overall error. For example, when P was changed from 75 to 85%, c_1 only increased by 5 (= 1464 − 1459) and remained much smaller than the true number of singletons $n_1 = 1488$. This implies that for the identity rule within this range, the Type II error was relatively larger than the Type I error, and resulted in an under-count of genes. At $P = 90\%$, however, the difference between c_1 and n_1 was minimized as was the difference between c_+ and n_+. At 95%, c_1 started to increase dramatically, indicating that $P \geq 95\%$ is too stringent to tolerate sequencing error, thus inflating Type I errors.

The behavior of **c** (Table 1) is explained by the frequency of the two types of error incidents (Table 2). For example, at $P = 75\%$, there were 20 genes with ESTs separated into two sub-clusters. The Type I error frequency remained virtually constant as P ranged from 75 to 90%. But at $P = 95\%$, the number of Type I error occurrences jumped to 62 and doubled at $P = 97.5\%$. On the other hand, the Type II error rate decreased as the identity percentage P increased, just as expected. We found that some ESTs from neighboring loci on the genome formed false clusters, which probably indicated that the two loci were derived from recent tandem duplications, or the EST was chimeric. This error was present even at $P = 97.5\%$. Apparently $P = 97.5\%$ was stringent enough to prevent most Type II errors, but at the cost of more Type I errors.

We further define several statistics to summarize the Type I and II error rates (Table 3). Let EI be the number of genes which have their ESTs split among several clusters (>1) due to Type I error, and let EI_{tot} be the number of clusters that resulted. The net Type I error, $EI_{\text{net}} = EI_{\text{tot}} - EI$, measures the number of additional clusters generated due to Type I error. For example, 22 genes were broken into 44 clusters at $P = 80\%$, giving $EI_{\text{net}} = 22$. For Type II error, let EII be the number of clusters which have ESTs from more than one gene. Let EII_{tot} be the number of genes representing

Table 1. Comparison of true and observed gene cluster profiles for two *A.thaliana* EST sets

Cluster size (i)	Flower bud							ABGR						
	[a]n_i	[b]c_i^{75}	[b]c_i^{80a}	[b]c_i^{85}	[b]c_i^{90}	[b]c_i^{95}	[b]$c_i^{97.5}$	[a]n_i	[b]c_i^{75}	[b]c_i^{80}	[b]c_i^{85}	[b]c_i^{90}	[b]c_i^{95}	[b]$c_i^{97.5}$
1	1488	1459	1460	1464	1496	1564	1658	1600	1918	1923	1933	1969	2018	3083
2	351	332	333	335	337	332	325	439	390	393	391	399	396	390
3	130	128	128	129	130	128	132	171	157	157	160	152	157	153
4	65	66	67	69	67	69	64	75	58	58	57	60	58	58
5	41	36	36	37	39	38	39	54	47	49	49	46	44	44
6	24	25	25	26	24	23	19	30	26	25	24	25	26	26
7	19	27	26	22	20	17	23	18	16	14	15	14	14	12
8	11	11	11	11	10	13	13	17	17	17	18	17	19	20
9	17	14	14	14	15	15	11	12	11	11	9	8	10	8
10	8	11	11	11	11	9	10	13	13	13	13	13	9	12
>10	7	61	63	63	62	60	58	50	48	47	47	48	47	45
n_+ or c_+	2216	2172	2174	2181	2211	2269	2352	2479	2701	2708	2717	2751	2798	2851
EST_{tot}	5048	5048	5048	5048	5048	5048	5048	5287	5287	5287	5287	5287	5287	5287

The 3′ Flower Bud EST set shows significantly less overall clustering error than the 5′ above-ground-organ (ABGR) EST set.
[a]n_is are gene frequencies based on genome annotation.
[b]c_is are gene frequencies based on CAP3 clustering with overlap length = 40 bp, and identity criterion in the superscripts.

these clusters. The net Type II error, $EII_{net} = EII_{tot} - EII$, is a measure of the underestimate of unigene number due to Type II error. For example, at $P = 80\%$, each of 61 (EII) clusters contained ESTs at least from 2 of 125 different genes (EII_{tot}). A simple calculation gives a loss of 64 ($= 125 - 61$) (EII_{net}) in the cluster count. Difference between EI_{net} and EII_{net} gives the deviation of c_+ from n_+, that is,

$$c_+ - n_+ = EI_{net} - EII_{net}.$$

If $c_+ > n_+$, then Type I error is dominant, while Type II error dominates if $c_+ < n_+$. We note that one could have $c_+ = n_+$ but not have $\mathbf{c} = \mathbf{n}$. So c_+ and n_+ merely summarize the overall errors.

Based on these statistics, we define Type I and Type II error rates as follows:

$$\alpha = \frac{EI}{n_+ - n_1},$$

and

$$\beta = \frac{EII}{n_+}.$$

Note the difference in the denominators. Type I error involves the separation of a true cluster which requires at least two ESTs in that cluster. Genes only represented by singletons never contribute to Type I error. Therefore, the denominator of α does not include true singletons. However, true singletons can be mis-clustered with other ESTs if sequence similarity is sufficiently high. Therefore, these genes were counted in the the Type II error rate, β.

These error rates clearly summarize the changing pattern of the two types of errors with P (Table 3). The Type I error rate doubled as P changed from 90 to 95%, and doubled again from $P = 95$ to 97.5%. The Type II error rate kept decreasing as P increased. These results are intuitive and suggest that the choice of identity parameter depends on whether one is interested in minimizing Type I, Type II or overall error rates. Our analysis of the *Arabidopsis* flower bud EST set suggests that 90% was the optimal identity criterion in terms of overall error.

3.1.2 A.thaliana above ground organ 5′ cDNA library The second data set we examined was *A.thaliana* 2–6 weeks above-ground organ ESTs (UniLib # = 17695, to be called ABGR hereafter). In this data set, 5522 ESTs of 5894 had significantly matching regions on the genome sequence. Among these, 5284 matched the annotated loci with concordant coding direction, and 238 ESTs with opposite coding direction. If we regard the 238 sequences as 3′ ESTs mislabeled to be 5′ ESTs, this will give a mislabeling rate estimate of 4.3%, about half as large as the 3′ EST case (8.2%). Among the remaining 327 (5894 − 5522) ESTs, 279 matched regions that were not annotated as genes, and 48 generated no significant match to the genome.

The Type II error rate was similar to that of the 3′ EST set (Table 3). It decreased from 2.9% at $P = 75\%$ to 0.5% at $P = 97.5\%$. At $P = 95$ and 97.5%, Type II errors were mainly due to instances where ESTs within each cluster matched the neighboring loci on the genome, similar to the case of the 3′ EST set.

A slight increase in Type I error rate was observed when P was increased from 75% to 95%. However, there was a 2.9% jump in Type I error when P was increased from 95 to 97.5%. This suggests that the sequencing error problem was not as severe in this 5′ EST set as in the 3′ EST

Table 2. Type I and Type II error decomposition in *Arabidopsis thaliana* Flower Bud tissue and above-ground-organs (ABGR) EST sets

Error		Flower bud												ABGR																	
		75%			80%			85%			90%			95%			97.5%			75%			80%			85%			90%			95%			97.5%		
Type I	[a]Clusters/Gene	2	—		2	—		2	—		2	2	3	2	3	4+				2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4
	[b]Gene Freq.	20	—		22	—		22	—		23	60	2	103	12	3				234	23	17	235	22	8	236	22	8	244	24	7	252	34	7	265	43	11
Type II	[c]Genes/Cluster	2	3		2	3		2	3		2	2	—	2	—	—				2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	—
	[d]Cluster Freq.	58	3		58	3		55	1		28	11	—	6	—	—				62	43	1	59	6	1	51	6	1	28	5	1	17	1	1	10	1	—

[a]Clusters/Gene is the number of subclusters formed in the sample by the ESTs from one gene when a Type I error occurs. [b]Gene Freq. is the frequency of such genes. For example, at $P = 80\%$ in Flower Bud data, 22 genes were separated into two subclusters, while there were 60 such genes at 95%.

[c]Genes/Cluster is the number of genes that the ESTs in one cluster belong to when a Type II error occurs and

[d]Cluster Freq. is the frequency of such clusters. For example, at $P = 80\%$ in the Flower Bud data, there were 58 clusters that contained two distinct genes due to Type II error.

Table 3. Type I and Type II clustering error rates over a range of clustering identity parameters[a]

Library	Identity[a] (%)	EI[b]	EI_{tot}[c]	EI_{net}[d]	α[e] (%)	EII[f]	EII_{tot}[g]	EII_{net}[h]	β[i] (%)
Flower bud	75	20	40	20	2.7	61	125	−64	2.8
	80	22	44	22	3.0	61	125	−64	2.8
	85	22	44	22	3.0	56	113	−57	2.5
	90	23	46	23	3.2	28	56	−28	1.3
	95	62	126	64	8.5	11	22	−11	0.5
	97.5	118	258	140	16.2	6	12	−6	0.3
ABGR	75	265	565	300	29.8	70	149	−79	2.9
	80	265	568	303	30.1	66	140	−74	2.7
	85	266	570	304	30.3	58	124	−66	2.3
	90	275	588	313	31.3	34	75	−41	1.4
	95	293	634	341	33.3	19	41	−22	0.8
	97.5	319	703	384	36.2	11	23	−12	0.5

Owing to insufficient overlap of EST sequences, the Type I error rates are quite high for the 5′ above-ground-organ EST set relative to the 3′ flower bud EST.
[a]Identity rule P in CAP3 with overlap length $O = 40$ bp.
[b]The number of genes that have Type I errors.
[c]The total number of subclusters generated from genes in EI due to Type I error.
[d]$EI_{net} = EI_{tot} - EI$ is the net inflation of EST cluster count due to Type I error.
[e]Type I error rate: $\alpha = EI/(n_+ - n_1)$. n_+ is the total number of genes from genome annotation result in Table 1.
[f]The number of EST clusters that contain ESTs from at least two different genes.
[g]The total number of genes represented by EST clusters in *EII*.
[h]$EII_{net} = EII_{tot} - EII$ measures the net reduction of EST cluster count due to Type II error.
[i]Type II error rate: $\beta = EII/n_+$.

example. However, as expected, the Type I error rate was ∼10 times higher when clustering 5′ ESTs relative to the error rate observed in the 3′ EST example (Table 3). The cause of this substantial Type I error for 5′ EST assembly is the ISO error.

Despite the large Type I error rate, we can still determine the optimal identity rule P by examining the change of net Type I and Type II errors $\Delta E = \Delta EI + \Delta EII$ across different identity criteria. The optimal P will lead to $\Delta E < 0$. For example, from $P = 80$ to 85%, Type I error frequency increased by 1 (266 − 265), while Type II error decreased by 8 (66 − 58). Since $\Delta E = -7 < 0$, 85% was better than 80%. We found that $P = 90\%$ was the optimal identity criterion and $P = 95\%$ only slightly worse ($\Delta E_{85–90} = -15$, $\Delta E_{90–95} = +3$).

3.2 ISO error

3.2.1 ISO error distribution The results shown in Table 3 suggest a Type I error rate as large as 36.2% in the 5′ EST clustering, substantially larger than the 3′ EST example. The significant increase in Type I error was caused by ISO error. In the present study, the empirical ISO error distribution simulated based on the training data is partially listed in Table 4 (a complete version through $X = 30$ can be found in the supplementary materials). Inspection of $P(\mathbf{Y}|X)$ suggests, e.g. that 20.8% of the genes with two ESTs in a sample are expected to be observed as two singletons due to ISO error.

Table 4. The empirical ISO error distribution

X[a]	Y_1[b]	Y_2[b]	Y_3[b]	Y_4[b]	Prob.[c]
2	1	1	0	0	0.208
	2	0	0	0	0.792
3	1	1	1	0	0.027
	2	1	0	0	0.234
	3	0	0	0	0.739
4	1	1	1	1	0.001
	2	1	1	0	0.041
	2	2	0	0	0.057
	3	1	0	0	0.187
	4	0	0	0	0.713
5	2	1	1	1	0.004
	2	2	1	0	0.016
	3	1	1	0	0.030
	3	2	0	0	0.077
	4	1	0	0	0.168
	5	0	0	0	0.706
6	2	2	1	1	0.002
	2	2	2	0	0.001
	3	1	1	1	0.003
	3	2	1	0	0.022
	3	3	0	0	0.026
	4	1	1	0	0.027
	4	2	0	0	0.065
	5	1	0	0	0.145
	6	0	0	0	0.707
7	2	2	2	1	0.001
	3	2	1	1	0.003
	3	2	2	0	0.002
	3	3	1	0	0.002
	4	1	1	1	0.003
	4	2	1	0	0.017
	4	3	0	0	0.040
	5	1	1	0	0.022
	5	2	0	0	0.059
	6	1	0	0	0.140
	7	0	0	0	0.712
8	3	2	2	1	0.001
	3	3	1	1	0.001
	3	3	2	0	0.002
	4	2	1	1	0.002
	4	2	2	0	0.002
	4	3	1	0	0.008
	4	4	0	0	0.016
	5	1	1	1	0.002
	5	2	1	0	0.015
	5	3	0	0	0.038
	6	1	1	0	0.022
	6	2	0	0	0.058
	7	1	0	0	0.119
	8	0	0	0	0.714

[a]X is the true number of ESTs for one gene.
[b]$\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)$ is the clustering outcome given X (so $X = Y_1 + Y_2 + Y_3 + Y_4$), each subcluster with Y_1, Y_2, Y_3, Y_4 ESTs respectively. If $Y_2 = Y_3 = Y_4 = 0$, then the cluster is complete.
[c]Prob. is the empirical probability of observing an outcome $\mathbf{Y}$ given the true number of ESTs that one gene has, namely, Prob($\mathbf{Y}|X = x$).

The ISO error can be further summarized in a conditional expectation matrix **P** (1), with entries

$$P_{ij} \equiv \sum_{t=1}^{4} P(y_t = j | X = i), \quad j = 1, \ldots, \leq i.$$

$\mathbf{P_{10}} =$

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .416 & .792 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .315 & .234 & .739 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ .275 & .155 & .187 & .713 & 0 & 0 & 0 & 0 & 0 & 0 \\ .256 & .112 & .106 & .168 & .706 & 0 & 0 & 0 & 0 & 0 \\ .235 & .096 & .076 & .092 & .147 & .707 & 0 & 0 & 0 & 0 \\ .219 & .085 & .048 & .059 & .081 & .140 & .712 & 0 & 0 & 0 \\ .200 & .083 & .052 & .044 & .055 & .080 & .119 & .714 & 0 & 0 \\ .197 & .079 & .052 & .034 & .036 & .054 & .071 & .112 & .717 & 0 \\ .187 & .073 & .040 & .028 & .032 & .033 & .047 & .058 & .110 & .730 \end{bmatrix} \quad (1)$$

Here, P_{ij} is the expected frequency of clusters with j ESTs that would be generated from a true cluster with $X = i$ ESTs due to ISO error. The 10×10 triangular matrix $\mathbf{P}_{10} = \{P_{ij} : i = 1, \ldots, 10, j = 1, \ldots, 10\}$ summarizes the simulation results for $X \leq 10$. For example, for every gene with exactly two ESTs in the sample, the expected number of singletons generated due to insufficient overlap of ESTs is 0.416. In other words, for every 100 genes with two ESTs, 79 are expected to form contigs, and 21 to be split into two singletons creating 42 'clusters'. The i-th diagonal elements, P_{ii}, plotted in Figure 2 gives the probability that there is no ISO error, so the ESTs from a gene with i ESTs in the sample are clustered together.

The probability of no ISO error, P_{ii}, as a function of the true number of ESTs per gene, shows an interesting convex pattern (Fig. 2). Starting with 0.792 at $X = 2$, the error-free rate keeps descending until $X = 5$, then climbs steadily. Initially the decline is due to the fact that in order to connect two ESTs, only one region of overlap is needed, whereas more regions of overlap are required to join additional ESTs. As a consequence, there is a higher probability of ISO error when $X = 3$ than $X = 2$. However, since the cDNA length is limited, eventually more ESTs result in a larger chance of overlap and the probability of ISO error decreases. Of course this pattern is dependent upon the three afore-mentioned distributions. For example, increasing the EST length may result in more reduction of ISO errors at $X = 3$ than $X = 2$. As a consequence, this convex pattern may vanish as sequencing technology improves.

3.2.2 ISO error correction The simulated ISO error distribution can be used to correct for the ISO error based on its probabilistic definition, thereby improving the estimates of the true gene clustering profile data **n**. Usually in an EST set from one cDNA library, cluster counts n_i for $i \geq 20$ are relatively much smaller than those of smaller clusters. Furthermore, the ISO error rate for $X \geq 20$ is smaller than small Xs

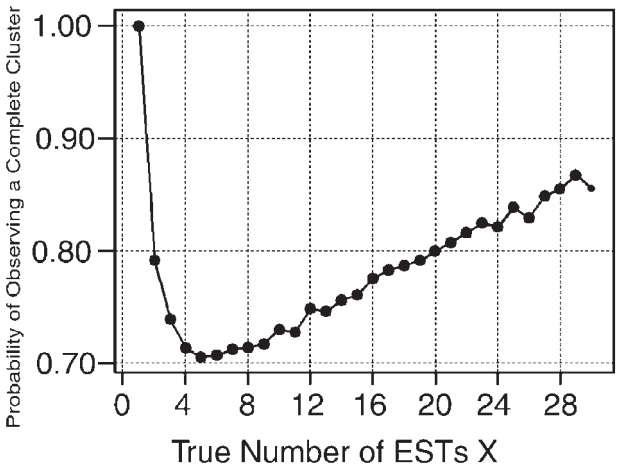

Fig. 2. Probability of being a complete cluster (no ISO error), given the true number of ESTs for a gene. The x-axis is the true number of ESTs that one gene has in the sample. The y-axis is the probability of observing that all the ESTs from that gene with X ESTs are clustered together in the sample. Genes with lower expression contribute most of the ISO error.

(see Supplementary materials and Fig. 2). Therefore, the ISO error from clusters with more than 20 ESTs (n_i for $i > 20$) is trivial. It is usually adequate to estimate n_is, for $i \leq 20$ and accept the observed expression profile values, c_i, for $i > 20$. Suppose we observe

$$\mathbf{c} = (c_1, c_2, \ldots, c_{20}, \ldots, c_t).$$

Let

$$\mathbf{P}_{20.t} = \begin{bmatrix} \mathbf{P}_{20}^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{t-20} \end{bmatrix},$$

where $\mathbf{I}_{t-20}$ is a $(t - 20) \times (t - 20)$ identity matrix. Then under certain assumptions, i.e. the other error sources can be ignored compared with the ISO error in 5′ EST clustering, and the simulated error distribution represents the true one, we approximately have

$$E(\mathbf{c}|\mathbf{n}) = \mathbf{P}_{20.t}\mathbf{n}.$$

where 'E' means expectation. This immediately gives an unbiased estimate of **n** as

$$\hat{\mathbf{n}} = \mathbf{P}_{20.t}^{-1}\mathbf{c}. \quad (2)$$

$\mathbf{P}_{20.t}$ will be called the ISO Error Correction Matrix. We can decrease $\mathbf{P}_{20}$ to $\mathbf{P}_{10}$ if the EST sample size is relatively small, e.g. $c_i \leq 5$ for $i \geq 10$. The estimate $\hat{\mathbf{n}}$ can be regarded as a much better estimate of **n** than the observed expression profile, **c**.

To illustrate the performance of this method, we now use $\mathbf{P}_{20}$ to correct $\mathbf{c}^{90}$ in *A.thaliana* ABGR set, and an additional *A.thaliana* Root 5′ EST set (UniLib # = 17709) obtained and

Table 5. ISO correction for *A.thaliana* above-ground-organ and Root 5′ EST sets

Cluster size (i)	ABGR tissue			ROOT tissue		
	[a]n_i	[b]c_i^{90}	[c]$\hat{n}^{90}$	[a]n_i	[b]c_i^{90}	[c]$\hat{n}^{90}$
1	1600	1969	1684	1985	2304	1997
2	439	399	421	479	435	469
3	171	152	173	177	133	134
4	75	60	63	101	96	114
5	54	46	54	53	38	38
6	30	25	28	35	40	49
7	18	14	14	20	18	21
8	17	17	20	22	15	18
9	12	8	8	9	6	6
10	13	13	16	5	7	8
>10	50	47	52	42	45	49
n_+ or c_+	2479	2708	2533	2928	3137	2903
EST_{tot}	5287	5287	5287	5573	5573	5573

[a]n_is are gene frequencies based on genome annotation.
[b]c_is are gene frequencies based on CAP3 clustering with overlap length = 40 bp, and identity in the superscripts.
[c]$\hat{n}_i$ is the corrected result from $\mathbf{P}_{20}$.

processed in the same way. The superscript in $\mathbf{c}^{90}$ indicates that a 90% identity rule was used in CAP3.

The $\mathbf{n}$, $\mathbf{c}$, $\hat{\mathbf{n}}$ are listed in Table 5. Under this correction, for the ABGR data, the singleton count decreased from 1969 to 1684. The bias decreased from 369 (= 1969−1600) to 84 (= 1684−1600). Furthermore, the bias for the frequency of clusters of size 2 was reduced from −40 to −19. The total number of inferred unigenes is $\hat{n}_+$=2533, much closer to the $n_+ = 2479$ than $c_+ = 2780$. A substantial correction effect is also observed for the root data. In this case, ISO correction resulted in a very small (0.8%) underestimate of the true number of unigenes in the EST set ($\hat{n}_+ = 2903$ versus $n_+ = 2928$), whereas the uncorrected unigene estimate ($c_+ = 3137$) was a substantial 7.1% overestimate. Methods for simulation of ISO error distribution when the genome or complete cDNA sequences are not available are further discussed in Discussion section.

3.2.3 Expectation of Type I error from ISO error From the ISO error distribution, we can calculate the expectation of Type I error due to ISO. Note P_{ii} in the $\mathbf{P}$ matrix is the probability of NOT observing ISO error for $x = i$. If we know $\mathbf{n}$, the expectation of α can be calculated by

$$E(\alpha|\mathbf{n}) = 1 - \frac{\sum_{i=2}^{t} n_i P_{ii}}{n_+ - n_1}. \quad (3)$$

If $\mathbf{n}$ data are not available, one can use $\hat{n}_i$ from (2) in Equation (3) to calculate an approximate expectation. In the ABGR example where $\mathbf{n}$ is available, this gives $E(\alpha|\mathbf{n}) = 25\%$. The expectation here is 6% lower than $\alpha \approx 31\%$ observed without ISO correction (Table 3). The difference could be due to the random variation of the realized ISO error, sequencing error and other sources of random error. Clearly the majority of Type I error in estimates of gene cluster profiles was effectively corrected by the proposed method (25/31 = 81%).

4 DISCUSSION

Accurate estimation of the gene cluster profile $\mathbf{n}$ allows investigators to use EST data sets to make important inferences about the cDNA libraries from which ESTs were sampled. We have provided a means for improving estimates of gene cluster profiles by correcting for ISO error. In the following paragraphs, we further discuss the relationship between clustering criteria and clustering error; we propose two alternative ways for ISO error distribution simulation when the complete cDNA or genome sequences are unavailable; we also discuss the impact of the alternative splicing on the estimation of $\mathbf{n}$ and illustrate an application of ISO error correction to estimate the sampling redundancy in an EST data set.

4.1 Clustering algorithm and clustering criteria

One can clearly see the interaction of Type I and Type II errors as clustering criteria change in CAP3. The optimality of clustering criteria depends on the desired clustering outcome. For example, in both the 5′ and 3′ EST examples, the Type II error rate decreased steadily as the identity criterion increased and almost vanished when using $P = 95\%$. From $P = 95$ to 97.5% the Type I error due to sequencing error increased dramatically in the 3′ case, but the increase was much milder in the 5′ EST clustering. These case studies suggest that for applications that require minimizing Type II errors, 95% is a good rule, but 90% as shown, is better in minimizing the overall error levels. For applications using the gene cluster profile $\mathbf{n}$ or digital gene expression profile X_j (Audic and Claverie, 1997; Stekel *et al*., 2000), we cautiously warn readers against using an excessively stringent identity rule P in EST clustering because it can inflate the Type I error rate. In particular, when using $\mathbf{n}$ data for inference of cDNA library properties, substantial ISO error must be taken into account for legitimate quantitative conclusions, especially for 5′ EST data. We also tried Type I and Type II error decomposition on several other individual EST sets of *A.thaliana* including data derived from a root cDNA library, 3′ ESTs from ABGR and seed cDNA libraries. The distributions of Type I and Type II errors were similar to what we observed in the above examples (results not shown).

The other main criterion in clustering is the overlap length O. Throughout this paper, we used $O = 40$ bp as the cutoff (also used by TIGR, see http://www.tigr.org/tdb/tgi/definitions.html). We did compare the clustering result using different overlap lengths from 25 to 40 bp. The resulting difference in the clustering outcome is negligible for both 3′ and 5′ EST cases (data not shown). Furthermore, we rarely observed the occurrence of false joining of two ESTs (Type II error) due to the short overlap length when we investigated

the Type II error cases at the 90% identity rule. This suggests that a 25 bp overlap rule would be adequate.

The results on the clustering criteria choice based on CAP3 are not directly applicable to other clustering approaches, however they provide insights to other clustering procedures. All the clustering algorithms based on sequence similarity comparison must face the Type I and Type II error issue. An ideal setting of criteria would be one that is stringent enough to separate paralogs while is capable of tolerating sequencing error to avoid Type I error. These two types of errors may differ in magnitude but must have similar dependence pattern as illustrated based on CAP3. In addition, the ISO error is the common and unavoidable issue for any approach if genomic or proteomic information is unavailable, even the sequence quality is perfect. We believe that the scope and pattern of ISO error presented here will be similar under different gene-identification-oriented EST clustering programs (rather than gene family). For example, STACK_pack first performs loose pre-clustering using *d2_cluster* (Burke *et al*., 1999) based on sequence similarity, then uses CRAW (Burke *et al*., 1998; Chou and Burke, 1999) and CONTIGPROC (Miller *et al*., 1999) to detect subgroups (e.g. alternative splicing forms, paralogs) present in the pre-clustering result. The loose pre-clustering allows capturing splice variants and true sibling ESTs with poor quality, thereby reducing Type I error, whereas the CRAW and CONTIGPROC step helps to reduce Type II error. Since the 'consistency' criterion in the group partitioning strategy of CRAW is defined based on sequence similarity (Burke *et al*., 1998), we suspect that a good rule for distinguishing the substructure while tolerating sequencing error also depends on the sequence quality. In addition the ISO error problem remains since an EST that does not overlap with its siblings, will be regarded as a singleton in the pre-clustering stage. It would be our great interest to conduct a similar error analysis and compare it with CAP3 results in the near future.

4.2 ISO error correction and simulation

One remarkable feature of the proposed ISO error correction method is its applicability to an EST set of arbitrary size. The distribution of ESTs sampled per gene (X) obviously changes with the EST sample size. Thus the overall Type I error rate due to ISO also changes since $P(\mathbf{Y}|X)$ depends on X. In general, the error rate decreases as the EST sample size increases but at a slow pace as shown in Figure 2. However by design the ISO error correction matrix $\mathbf{P}$ is independent of sample size or the X distribution. For example, if one generates ESTs from the same species under the similar protocol, the ISO error simulated from a subsample can be used to correct for ISO error for larger samples.

ISO error estimation and correction are dependent upon three sources of information: the distribution of EST read lengths $[F(L_E)]$, the mRNA length distribution $[F(L_m)]$, and the conditional distribution of the start position given the mRNA length $[F(S|L_m)]$. The simulation method we described could easily be repeated to find the correction matrix $\mathbf{P}$ that is based on the observed sample EST length distribution $F(L_E)$ for any given library. However, both $F(L_m)$ and $F(S|L_m)$ for a given library may be influenced by among-species variation in mRNA length as well as particulars of mRNA extraction and library building procedures. In our examples above, these parameters were based on analysis of complete cDNAs that may not be available for a given species.

How could one simulate the ISO error distribution for a species that does not have complete cDNA sequences? One simple solution is suggested here. We can first cluster all available sequence data for a particular species and treat contigs of large clusters (e.g. $X \geq 20$ or 30) as complete cDNAs, then simulate the ISO error as we have done in this paper. For the same training *Arabidopsis* ESTs, the average contig length for $X \geq 20$ and $X \geq 30$ were 1263 and 1440 respectively, shorter than 1553 for the complete cDNA sequences in the training data. The simulated ISO error distribution under this strategy, however in both situations were satisfactory (Supplementary materials). The error-free probabilities were presented in Table 6 for $X \leq 10$.

This implies that when X is small, the ISO error rate is not too sensitive to the relatively small change of $F(L_m)$. Since the distribution of X in EST data is concentrated at small values, e.g. $n_1 + n_2 + n_3$ in the ABGR data accounted for $2210/2479 = 89\%$ of the genes. Therefore, the correction effect from the simplified method should be satisfactory.

The above strategy in ISO error simulation has been integrated into the ESTstat software. In addition to using contigs with many ESTs, one alternative solution that ESTstat provides is to utilizes $F(L_m)$ and $F(S|L_m)$ information from a species with a large full-length cDNA set such as *A.thaliana* to simulate $P(\mathbf{Y}|X)$ for the new species. This approach is based on the assumption that the complete cDNA length distribution $F(L_m)$ from the known and unknown species are similar. For example, the average length of rice complete cDNA sequences from http://cdna01.dna.affrc.go.jp/cDNA/ is 1700 bp, ~130 bp longer than that of *A.thaliana* obtained in the training data, suggesting that the complete cDNA length distribution over different plant species can be similar. If the ESTs are sequenced under similar protocols, one can use $F(L_m)$ and $F(S|L_m)$ information from *A.thaliana* or rice to simulate the ISO errors for other plant species.

One assumption we implicitly made in ISO error simulation is that the transcript length does not depend on the expression level. This assumption is supported by observations of no strong relationship between protein length and EST counts in *Caenorhabditis*, *Drosophila* or *Arabidopsis* (Duret and Mouchiroud, 1999). It is also strongly supported by our ongoing examination of the relationship between gene length and expression level using large EST and complete cDNA sets from *A.thaliana* and mouse. This hypothesis will be examined further using results from microarray experiments.

Table 6. Comparing the error-free probability $P[\mathbf{Y} = (x, 0, 0, 0)|X = x]$ for $x = 2$–10 simulated with and without complete cDNAs

X	2	3	4	5	6	7	8	9	10
With cDNAs[a]	0.80	0.74	0.71	0.71	0.69	0.72	0.72	0.71	0.75
No cDNAs[b]	0.82	0.73	0.74	0.76	0.76	0.78	0.79	0.80	0.80
No cDNAs[c]	0.79	0.73	0.71	0.70	0.71	0.71	0.74	0.75	0.75

[a]ISO error was simulated using complete cDNA set in the EST clustering as discussed in the text.
[b]ISO error was simulated without using complete cDNA set. We clustered the same 48 827 5′ *A.thaliana* ESTs as used in the text. The contigs of the large EST clusters with $X \geq 20$ respectively were treated as complete cDNAs and used to simulate the ISO error distribution $P(\mathbf{Y}|X)$.
[c]Same as the previous ore, but used $X \geq 30$.

4.3 Alternative splicing

The existence of alternative splicing forms (Modrek and Lee, 2002) causes some ambiguity in the definition of Type I and Type II errors. We defined Type I and Type II errors in terms of genes (loci) rather than transcripts. This is primarily because ESTs from different splicing forms can be indistinguishable due to the short length of EST sequences. In this case, two sibling ESTs can originate from transcripts with different splicing forms, but if both ESTs span only the common exon regions of a gene, they will not appear to be different. Thus, the observed splicing rate from EST sequences will be a lower bound and underestimate of the true splicing rate in a given library. Additionally, when the alternative splicing rate is low in individual tissues, using the genome annotation-based clustering result as reference counts becomes a sensible way to evaluate Type I and Type II errors. If we had defined these errors in terms of unique transcripts, then the lack of complete transcriptome information, and the inability to distinguish some ESTs derived from different transcripts of the same gene, would make Type I and Type II errors hard to evaluate.

In our method for ISO error correction (also in the ESTstat 1.0 software), alternative splicing is currently ignored. By matching ESTs to genome we found a very low observed alternative splicing rate in individual *Arabidopsis* EST sets. This is consistent with the expectation of a relatively low frequency of alternative splicing phenomena in plants (Haas *et al*., 2002) as compared to mammalian systems (Wright *et al*., 2001). This observation is also supported by the results reported by TIGR at http://www.tigr.org/tdb/tgi/plant.shtml. For example, 2216 alternative splicings were reported for *A.thaliana* based on the gene indices of 227 670 ESTs (January 12, 2004 version). For the the tomato data, only 492 such cases based on 155 317 ESTs were listed at the same website (Gene Index version April 17, 2003). The motivating applications often involve EST data of much smaller size, hence the chance of detecting such alternative splicings must be proportionally smaller. As we increase the sample size or mix ESTs from different tissues, the alternative splicing phenomenon may be observed more frequently when the true rate is high (Harrison *et al*., 2002; Modrek and Lee, 2002). In that situation, the recovered counts $\hat{n}_i$s will deviate upward from the genome annotation-based result $\mathbf{n}$ to an extent depending on the true alternative splicing rate.

4.4 Applications

Explicit identification of Type I and Type II error rates, and analytical procedures to minimize errors of interest, can help the interpretation of EST datasets and clustering results. For example, it has been reported that the UniGene cluster count is ∼35% larger than the true number of genes predicted for *Arabidopsis* (The *Arabidopsis* Genome Initiative, 2000; Van der Hoeven *et al*., 2002). One expected reason is that the usual ISO clustering error dominates the error structure of 5′ EST clustering results, as shown in this paper. In addition, 5′ and 3′ ESTs are mixed in the usual UniGene clustering. This increases further the chance of insufficient overlap especially between 5′ ESTs and their 3′ siblings if they are located far away from each other at the two ends of long cDNA clones (Fig. 1). Alternatively, for relatively short cDNA sequences, availability of 5′ as well as 3′ EST sequences should reduce ISO error. For these reasons, the ISO error is hard to evaluate quantitatively. Over-stringent clustering criteria can be another cause of high Type 1 error rate in EST clustering.

A simple application using the ISO error correction is the estimation of the number of genes n_+ that have been sampled. This number is needed in order to evaluate EST sampling redundancy, which can be defined as the average ESTs per gene, i.e. $\#\text{ESTs}/n_+$. For example the true number of genes sampled in the ABGR library was 2479; the observed number of unique sequences was 2708 and the estimate after ISO correction was 2533 (Table 5). Sequencing redundancy for this library is $5287/2479 = 2.13$ whereas it would be estimated as 1.95 without ISO error correction, and 2.08 with correction. In a subsequent paper, we will show how to use $\hat{\mathbf{n}}$ from a single EST set or multiple sets to estimate the number of expressed genes in the underlying tissue(s). This will further illustrate the importance of correcting ISO error.

Although we used plant EST sets for illustration throughout this paper, the methods and software developed here are also applicable to EST data from other organisms including mammals. However, as discussed earlier, if alternative splicing occurs frequently in the given species and library, then the gene cluster profile data from CAP3 can be inflated even after ISO error correction (for 5′ case). The methods developed in this and a subsequent papers are designed for statistical analysis of the properties of a single or multiple cDNA libraries rather than genome-assisted EST clustering. In the current version of ESTstat 1.0, there is no limit to the EST sample size.

ACKNOWLEDGEMENTS

The authors thank three anonymous reviewers for suggestions to help clarify important concepts; Drs Hong Ma and Francesca Chiaromonte for helpful comments and suggestions; Dr Xiangqiu Huang for providing the CAP3 program and Drs John Quackenbush and Geo Pertea for help in EST cleaning. The research was jointly supported by NSF Grant DMS0104443 to B.G.L. and NSF Grant DBI0115684 to C.W.D. at the Pennsylvania State University. This is paper #19 from The Floral Genome Project.

REFERENCES

Adams,M.D., Kerlavage,A.R., Fields,C. and Venter,J.C. (1993) 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.*, **4**, 256–267.

Adams,M.D., Dubnick,M., Kerlavage,A.R., Moreno,R., Kelley,J.M., Utterback,T.R., Nagle,J.W., Fields,C. and Venter,J.C. (1992) Sequence identification of 2,375 human brain genes. *Nature*, **355**, 632–634.

Asamizu,E., Nakamura,Y., Sato,S. and Tabata,S. (2000) A large scale analysis of cDNA in *Arabidopsis thaliana*: generation of 12,028 non-redundant expressed sequence tags from normalized and size-selected cDNA libraries. *DNA Res.*, **7**, 175–180.

Audic,S. and Claverie,J.M. (1997) Computational methods for the identification of differential and coordinated gene expression. *Human Mol. Genet.*, **8**, 1821–1832.

Boguski,M.S. and Schuler,G.D. (1995) ESTablishing a human transcript map. *Nat. Genet.*, **10**, 369–371.

Burke,J., Davison,D. and Hide,W. (1999) d2_cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.*, **9**, 1135–1142.

Burke,J., Wang,H., Hide,W. and Davison,D. (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.*, **8**, 276–290.

Chou,A. and Burke,J. (1999) CRAWview: for viewing splicing variation, gene families, and polymorphism in clusters of ESTs and full-length sequences. *Bioinformatics*, **15**, 376–381.

Christoffels,A., Van Gelder,A., Greyling,G., Miller,R., Hide,T. and Hide,W. (2001) STACK: Sequence Tag Alignment and Consensus Knowledgebase. *Nucleic Acids Res.*, **29**, 234–238.

Duret,L. and Mouchiroud,D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis. Proc. Nat. Acad. Sci., USA*, **96**, 4482–4487.

Green,P. (1996) Phrap. documentation. University of Washington, Seattle.

Haas,B.J., Volfovsky,N., Town,C.D., Troukhan,M., Alexandrov,N., Feldmann,K.A., Flavell,R.B., White,O. and Salzberg,S.L. (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.*, **3**, Epub 2002 May 30.

Harrison,P.M., Kumar,A., Lang,N., Snyder,M. and Gerstein,M. (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.*, **30**, 1083–1090.

Heber,S., Alekseyev,M., Sze,S.H., Tang,H. and Pevzner,P.A. (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, **18**, 181–188.

Hu,G., Modrek,B., Stensland,R.H.M., Saarela,J., Pajukanta,P., Kustanovich,V., Peltonen,L., Nelson,S.F. and Lee,C. (2002) Efficient discovery of single-nucleotide polymorphisms in coding regions of human genes. *Pharmacogenomics J.*, **2**, 236–242.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **6**, 829–845.

Khan,A.S., Wileox,A.S., Polymeropoulos,M.H., Hopkins,J.A., Stevens,T.J., Robinson,M.R., Orpana,A.k. and Sikela,J.M. (1992) Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nat. Genet.*, **2**, 180–185.

Lee,C. (2003) Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, **19**, 999–1008.

Liang,F., Holt,I., Pertea,G., Karamycheva,S., Salzberg,S.L. and Quackenbush,J. (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acids Res.*, **28**, 3657–3665.

Miller,R.T., Christoffels,A.G., Gopalakrishnan,C., Burke,J., Ptitsyn,A.A., Broveak,T.R. and Hide,W.A. (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.*, **9**, 1143–1155.

Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.

Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.

Picoult-Newberg,L., Ideker,T.E., Pohl,M.G., Taylor,S.L., Donaldson,M.A., Nickerson,D.A. and Boyce-Jacino,M. (1999) Mining SNPs from EST databases. *Genome Res.*, **9**, 167–174.

Schuler,G.D., Boquski,M.S., Stewart,E.A., Stein,L.D., Gyapay,G., Rice,K., White,R.E., Rodriquez-Tome,P., Aggarwal,A., Bajorek,E. *et al*. (1996) A gene map of the human genome. *Science*, **274**, 540–546.

Seki,M., Narusaka,M., Kamiya,A., Ishida,J., Satou,M., Sakurai,T., Nakajima,M., Enju,A., Akiyama,K., Oono,Y. *et al*. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.

Stekel,D.J., Git,Y. and Falciani,F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res.*, **10**, 2055–2061.

Sutton,G., White,O., Adams,M.D. and Kerlavage,A.R. (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci. Technol.*, **1**, 9–18.

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature*, **408**, 796–815.

Van der Hoeven,R., Ronning,C., Giovannoni,J., Martin,G. and Tanksley,S. (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell*, **14**, 1441–1456.

Wright,F.A., Lemon,W.J., Zhao,W.D, Sears,R., Zhuo,D., Wang,J.P., Yang,H.Y., Baer,T., Stredney,D., Spitzner,J. *et al*. (2001) A draft annotation and overview of the human genome. *Genome Biol.*, **2**, 1–18.

Xu,Q., Modrek,B. and Lee,C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.*, **30**, 3754–3766.