Research article

# Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation

## Donglin Liu and Joel H Graber*

Address: The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609, USA

Email: Donglin Liu - liudl@yahoo.com; Joel H Graber* - joel.graber@jax.org

* Corresponding author

## Abstract

**Background:** Publicly accessible EST libraries contain valuable information that can be utilized for studies of tissue-specific gene expression and processing of individual genes. This information is, however, confounded by multiple systematic effects arising from the procedures used to generate these libraries.

**Results:** We used alignment of ESTs against a reference set of transcripts to estimate the size distributions of the cDNA inserts and sampled mRNA transcripts in individual EST libraries and show how these measurements can be used to inform quantitative comparisons of libraries. While significant attention has been paid to the effects of normalization and substraction, we also find significant biases in transcript sampling introduced by the combined procedures of reverse transcription and selection of cDNA clones for sequencing. Using examples drawn from studies of mRNA 3'-processing (cleavage and polyadenylation), we demonstrate effects of the transcript sampling bias, and provide a method for identifying libraries that can be safely compared without bias. All data sets, supplemental data, and software are available at our supplemental web site [1].

**Conclusion:** The biases we characterize in the transcript sampling of EST libraries represent a significant and heretofore under-appreciated source of false positive candidates for tissue-, cell type-, or developmental stage-specific activity or processing of genes. Uncorrected, quantitative comparison of dissimilar EST libraries will likely result in the identification of statistically significant, but biologically meaningless changes.

## Background

Expressed sequence tags(ESTs) are single strand reads of transcribed sequence generated from cDNA clones [2-6]. EST sequencing typically originates in the vector, and can include either 5'- or 3'-terminal sequence of the cDNA clone. ESTs have historically provided data for gene discovery [7-11], tissue- or stage-specific gene expression [12-15], alternative splicing [16,17], and alternative polyadenylation [4,5,18-21].

While EST-based gene discovery can be quite successful, the wide dynamic range of mRNA abundance and the cost of EST creation led to the development of procedures such as normalization and subtraction [13,22], which increase the likelihood of sampling rare or tissue-specific transcripts, at the cost of lost quantitative relationships between different transcripts in a library. Normalization and subtraction utilize a common mechanism, which can be briefly described as heat dehybridization of cDNA, rapid re-hybridization in the presence of a 'driver' sample,
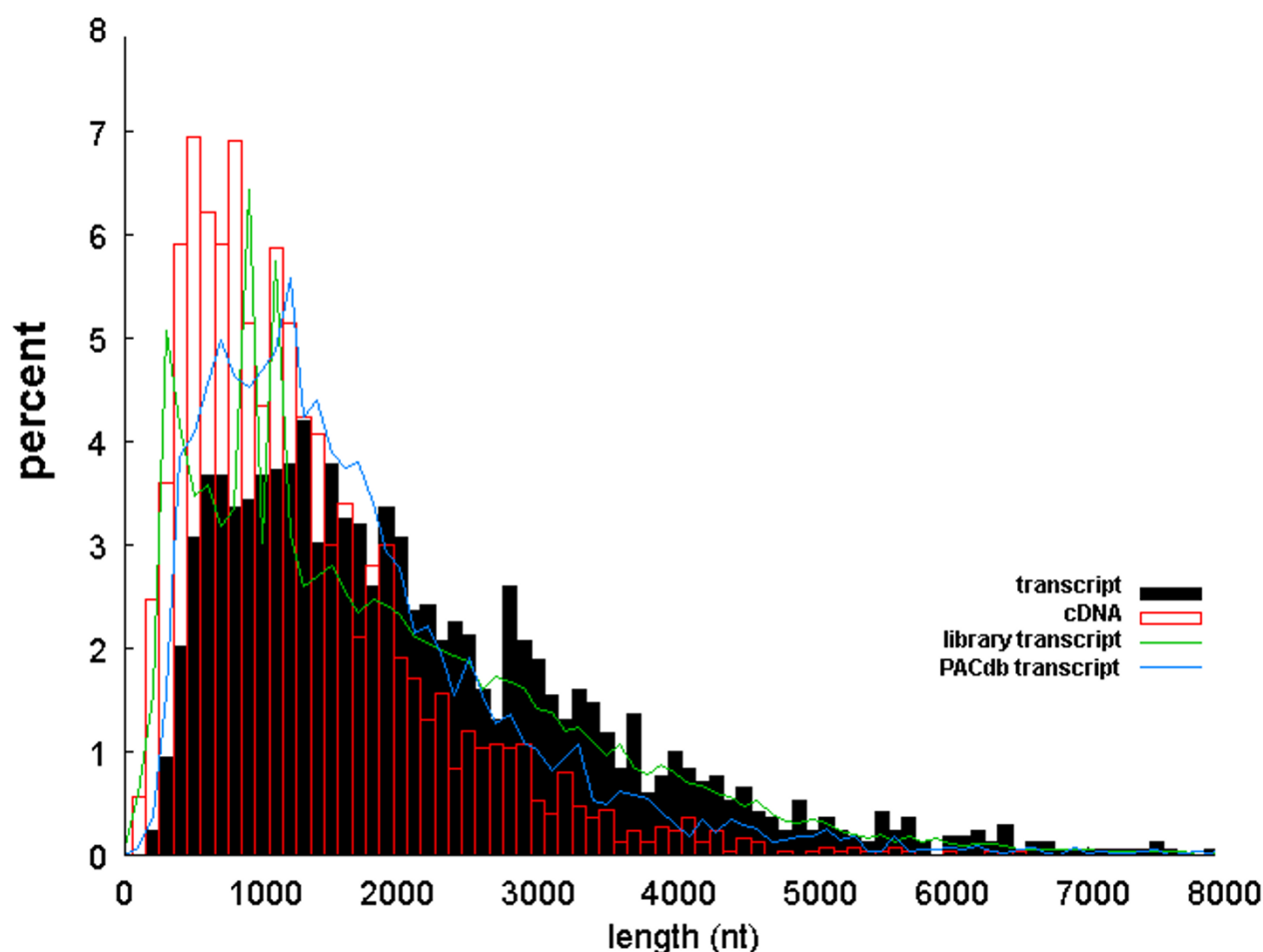
**Figure 1**
**Example estimates of cDNA insert and mRNA transcript length distributions for an EST library**. The estimated length distributions of cDNA inserts (red bars) and sampled transcripts (black bars) in a mouse EST library generated from round spermatids (McCarrey, J., Eddy, M. et al, unpublished data). For comparison, the length distributions of the ENSEMBL [44] and PACdb [21] reference transcripts are plotted in blue and green, respectively.

extraction of the double-stranded portion of the sample, and finally sequencing of a sampling of the remaining single-stranded sequences. The rapid re-hybridization step favors duplex formation of abundant species, therefore the remaining single-stranded sample is enriched for rare transcripts. The difference between normalization and subtraction lies in the choice of driver sequences. In normalization, the drivers come from the same sample, whereas in subtraction, the driver comes from a separate sample (or even pool of samples).

Since cDNA clones are created from mRNA sequences, the distribution of ESTs in a non-normalized library is presumably reflective of the population of mRNA sequences in the originating tissue. Audic and Claverie [12] demonstrated how non-normalized EST libraries could be analyzed to generate "transcript profiles" or "digital Northerns," and further developed rigorous statistical tests for significant variation between tissue or cell types. Several methods have been subsequently developed to enable studies of cDNA libraries to elucidate targets and mechanisms of tissue- and/or stage-specific gene expression [23-25]. Bioinformatic tools, such as Tissuelnfo [26], BodyMap [27] and ExQuest [28] utilize counts of ESTs in libraries for high-throughput identification of tissue expression profiles and specificity, in spite of the known limitations on their fidelity of representation of gene expression on the originating tissue [7].

**Table 1: Analysis of a group of EST libraries from a common sample. Each of these EST libraries was generated from a common tissue sample. Distance is calculated as L-divergence (Equation 3) between distributions of cDNA and transcript lengths for each library. The targeted size range of the cDNA inserts for each library ranges from 0.5 k to 7 k as described previously [30]**

| LibName | Total ESTs | 5' ESTs | L-Divergence | Targeted size(kb) |
|---|---|---|---|---|
| NIH-BMAP-FA0 | 2519 | 2352 | 0.308 | 0.5–1 |
| NIH-BMAP-FB0 | 1092 | 994 | 0.199 | 1–2 |
| NIH-BMAP-FD0 | 6594 | 5812 | 0.191 | 2–3 |
| NIH-BMAP-FC0 | 5969 | 3458 | 0.154 | 3–4 |
| NIH-BMAP-FI0 | 6769 | 5778 | 0.109 | 4–5 |
| NIH-BMAP-FO0 | 6135 | 5109 | 0.077 | 5–7 |

cDNA library generation is dependent on several steps, including selection and preparation of tissue, RNA purification, RNA to cDNA conversion, and cloning and transformation of the cDNA [29]. Procedures such as reverse transcription and selection of cDNA clones for sequencing can introduce systematic biases in any quantitative analysis using cDNA libraries. Confounding matters further, the depth of annotation among EST libraries is not uniform and often incomplete. It is worth noting that differences in transcript sampling between disparate laboratories and research groups are not unexpected, given the varied motivations and resources of the library creators.

We present here a method and means to identify and compensate for the biases in transcript sampling and enable quantitative comparisons between libraries. We analyzed and quantized over 900 mouse EST libraries with at least 100 entries, estimating the length distributions of cDNA inserts and their originating mRNA transcripts through alignment to a reference set of transcripts. (Similar analyses are being prepared for other organisms and will be made available on our web site [1].) The cDNA insert length distribution provides information about the efficiency and/or goals of the reverse transcription reaction, while the transcript length distribution evaluates the variety of transcripts sampled in the library. Our results show that the combined steps of reverse transcription and selection of clones (including optional restriction by insert size) for sequencing can introduce significant biases into transcript sampling in EST libraries, but that through identification and characterization of these biases, quantitative comparisons can be enabled. We identified the systematic biases as a part of a study of alternative 3'-processing in mouse spermatogenesis [Liu et al., in preparation], and therefore present our analysis with examples related to 3'-processing. The biases we describe are equally applicable, however, to any quantitative comparative analysis between distinct EST libraries, including assessment of tissue specificity of gene expression or processing.

## Results and discussion

Comparative studies of EST libraries, such as those to estimate tissue- or stage-specificity of expression, typically avoid the use of normalized or subtractive libraries. We find, however, that even non-normalized libraries are subject to systematic biases (arising from the procedures used in their generation) that can distort quantitative studies. To quantify these biases, we aligned all ESTs from each library to a reference transcript set, and used the results to estimate the length distributions of the cDNA inserts and sampled transcripts. A typical result is shown in Figure 1. A significant fraction of the libraries examined show the roughly lognormal distribution of both cDNA insert and transcript lengths, as shown in Figure 1. Comparison of the cDNA insert and transcript length distributions for a given library can give insights into the goals or parameters of the library generation. Similar distributions are indicative of an attempt to generate full length transcripts. A tight (low-variance), short, distribution of cDNA insert lengths indicates a fixed, relatively short reverse transcription reaction that results in cDNA insert lengths essentially independent of the transcript length. (Interestingly, such libraries are typically relatively unbiased in transcript sampling when compared with the reference sets). A tight distribution of cDNA insert lengths longer than a few hundred nucleotides indicates a size selection step of either mRNA or cDNA sequences prior to sequencing.

We examined several groups of libraries generated from a common tissue sample [30], described in Table 1. These specific libraries were generated in the NIH Brain Molecular Anatomy Project (BMAP) [30], the goal of which was generation of a large number of full-length transcripts from the mouse nervous system. Together these libraries represent a broad sampling of mRNA transcripts between 500 and 7000 nucleotides in length, with different libraries putatively selected to contain a specific size range. Figure 2 displays our estimates of the length distributions for the transcripts (top panel) and cDNA inserts (bottom panel) for these libraries. As expected, the modes of the
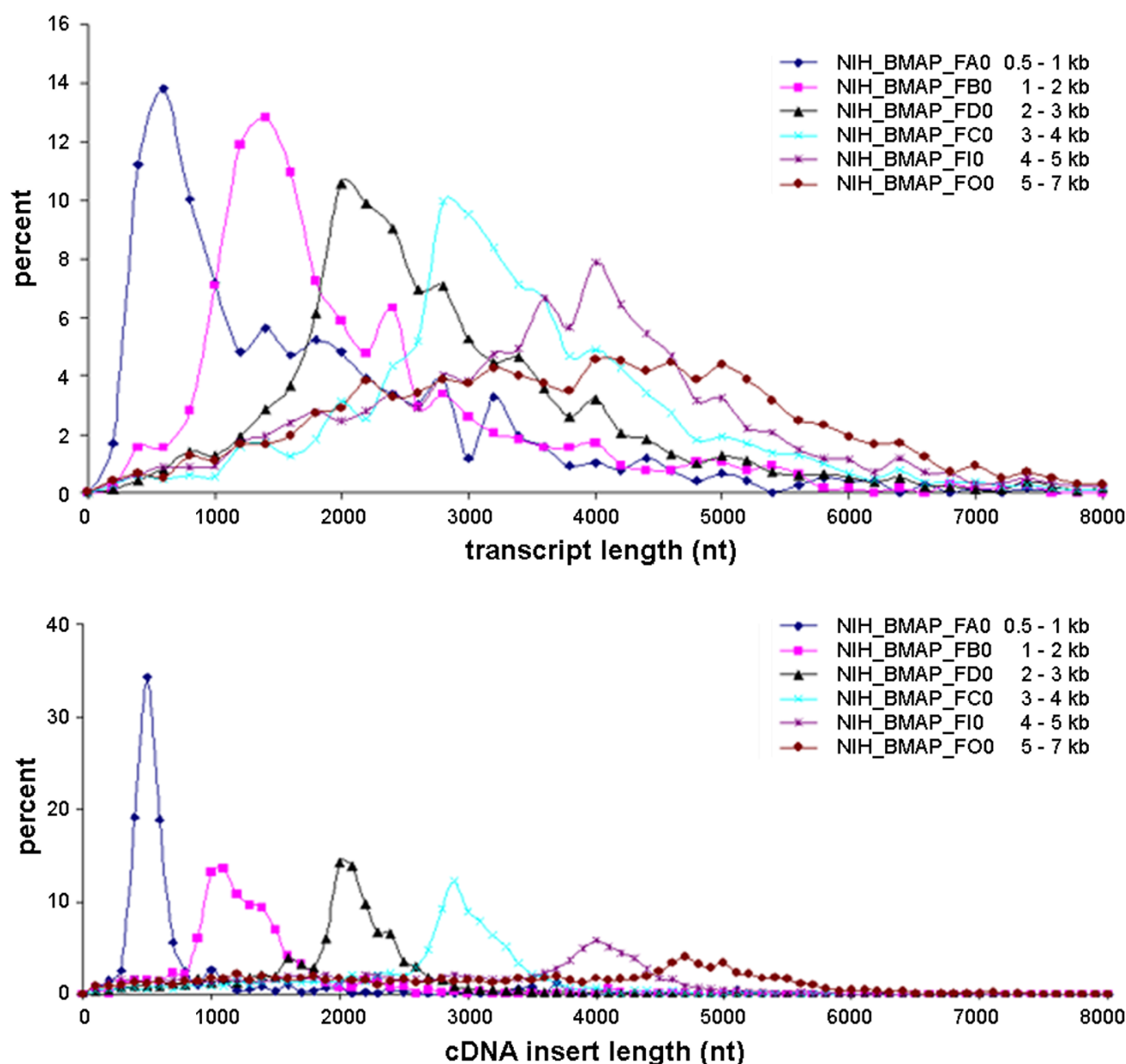
**Figure 2**
**Variation in cDNA insert and transcript length distributions for EST libraries from a common sample**. Estimated transcript (top panel) and cDNA insert (bottom panel) length distributions for several EST libraries derived from a common tissue sample [30]. These libraries were explicitly created to generate full-or near full-length cDNA sequences, and then size selected into the approximate ranges shown in the legends.

cDNA insert and transcript length distributions correlate quite well with the targeted ranges of the EST libraries. The attempt at creation of full-length transcripts in these libraries is apparent in comparison of the two panels in Figure 2, and also in the fourth column of Table 1, which shows the L-Divergence measurement between cDNA insert and transcript distributions for each library.

Interestingly, for the libraries with the shorter targeted lengths (*e.g.*, libraries NIH-BMAP-FAO and NIH-BMAP-FBO), our analysis indicates that the majority of the cDNA inserts and sampled transcripts come from the targeted range, however, as the targeted range increases, this specificity decreases, and a significant number of transcripts with length apparently outside the targeted range are
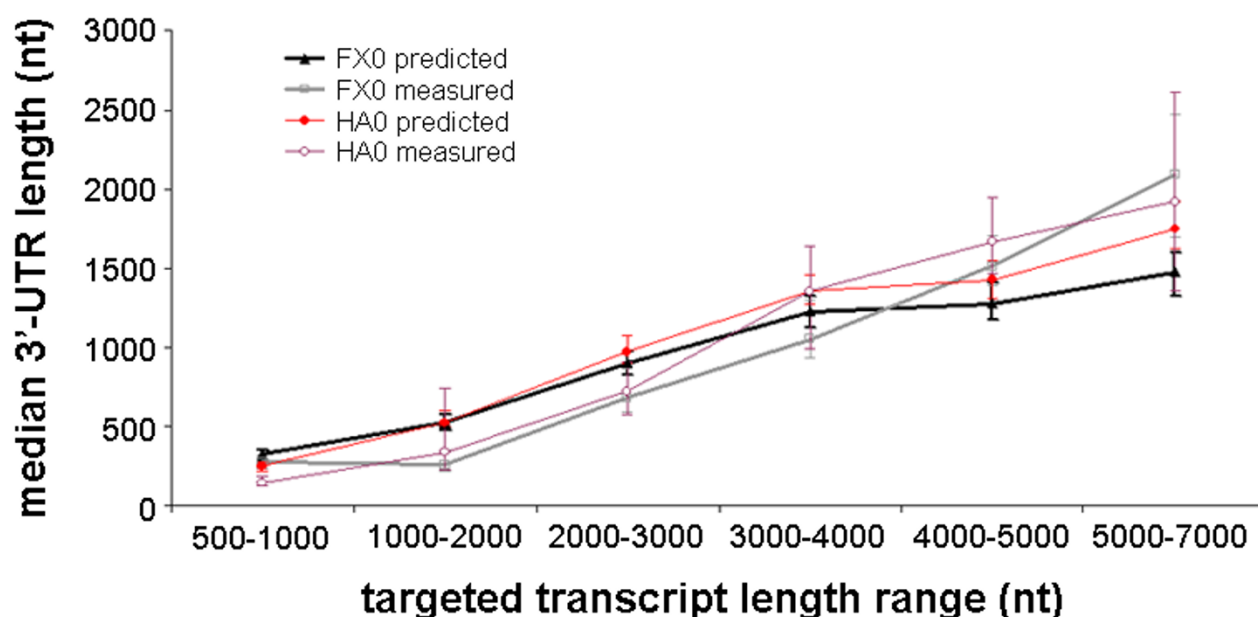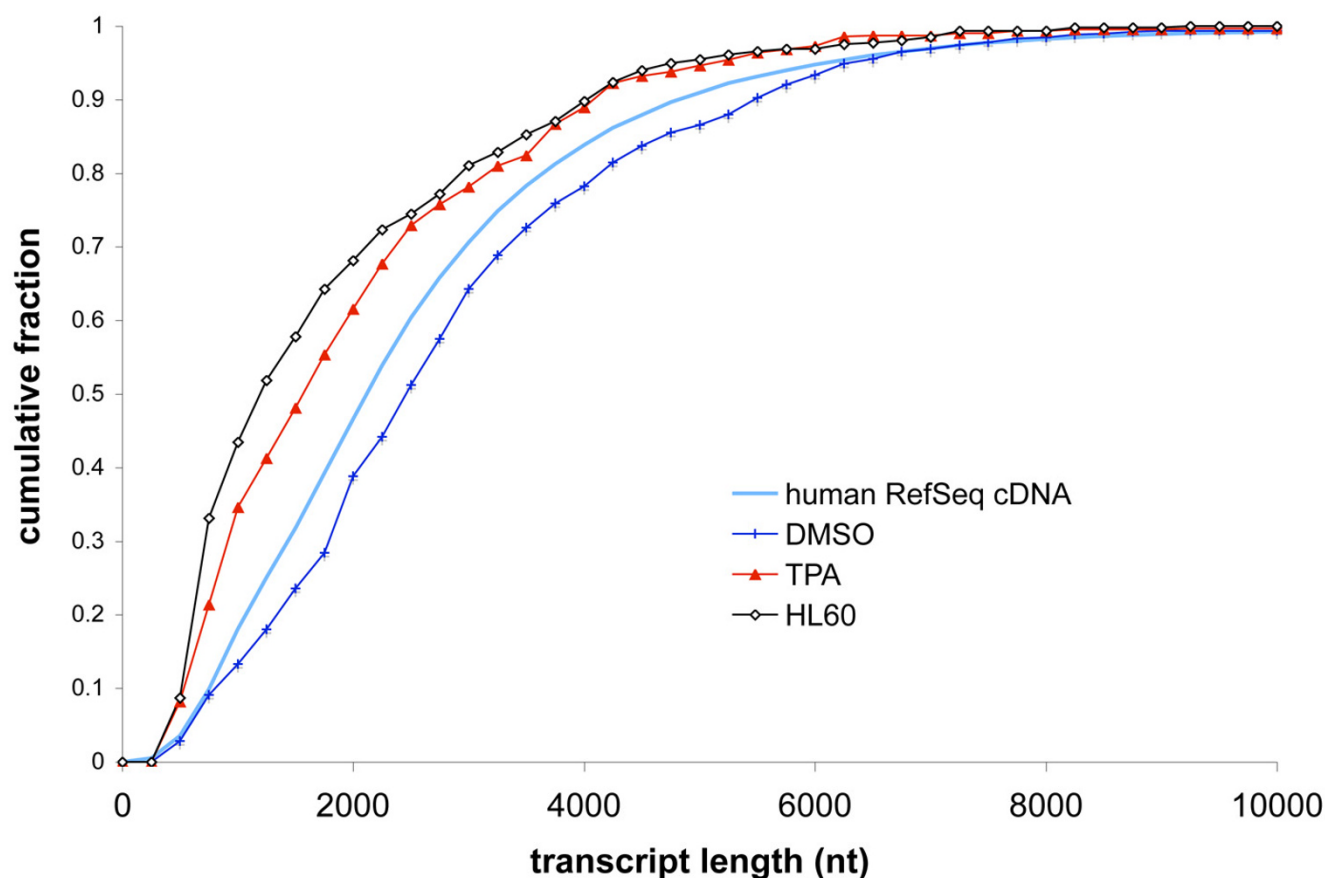
**Figure 3**
**Comparison of measured and expected median 3'-UTR lengths in EST libraries from a common sample**. A comparison of empirically derived and expected median 3'-UTR length for two groups of related EST libraries. Each set contains 6 libraries from a common tissue sample and targeted transcripts range from 0.5–1 k, 1–2 k, 2–3 k, 3–4 k, 4–5 k and 5–7 k. Expected 3'-UTR length distributions were calculated as described in Methods. The FX0 and HA0 library sets represent EST libraries from pools of tissues from brain and eye, respectively [30]. The correlations between the measured and predicted median 3'-UTR lengths for the FX0 and HA0 sets are 0.935 and 0.982, respectively.

found. This increasing discrepancy could arise from a number of effects. Electrophoretic separation is imprecise, including imperfections in the gels used for size selection and anomalous migration times for specific transcripts. Such unpredictable behavior will more significantly affect measurements of large molecules. Large molecules are also comparatively more prone to breakage. Our method for assignment of EST to transcripts is also a likely source of some ambiguity, as it is designed to give a reasonable picture of the characteristics of the complete library, but is certainly not guaranteed to give exact results for any specific EST or transcript.

mRNA 3'-processing (cleavage and polyadenylation) and the corresponding specification of 3'-UTR in eukaryotes play an important role for mRNA localization, translational efficiency, and stability [31-35]. We use EST-to-genome alignments to identify and characterize 3'-processing sites and 3'-UTRs [21]. Our studies of alternative 3'-processing during mouse spermatogenesis (Liu et al., in preparation) indicated that the distribution of 3'-UTR lengths can systematically change as a function of tissue type or growth conditions. Our analysis of PACdb [21] makes it clear that 3'-UTR length is dependent on

transcript length (see supplemental materials), therefore accurate characterization of these changes requires that we identify and compensate for the variation that arises due to differences in transcript sampling, as described above. As described in Methods, we generate an expected distribution of 3'-UTR lengths for any given library based on random sampling of (1) the transcript length distribution of the library, and (2) the distribution of 3'-UTR lengths as a function of transcript length. In Figure 3, we compare the measured and expected median 3'-UTR lengths (including ninety-five percent confidence intervals) for two additional groups of EST libraries, generated as size selected ranges from a common tissue in the Brain Molecular Anatomy Project [30]. The FX0 set, generated in a pool of brain tissues, includes libraries NIH-BMAP-FV0, NIH-BMAP-FX0, NIH-BMAP-FR0, NIH-BMAP-FW0, NIH-BMAP-FY0 and NIH-BMAP-GI0, while the HA0 set, generated from eye tissues, consists of libraries NIH-BMAP-HC0, NIH-BMAP-HA0, NIH-BMAP-GZ0, NIH-BMAP-HD0, NIH-BMAP-HB0 and NIH-BMAP-HE0. The agreement is quite good, with measured-expected correlations calculated at 0.935 and 0.982 for the FX0 and HA0 sets, respectively. Interestingly, the difference between measured and expected 3'-UTR lengths is greatest in library

**Figure 4**
**Estimated cumulative length distributions of the transcripts sampled by the HL60, DMSO-induced granulocy-toid, and TPA-induced monocytoid cell lines [36]**. For comparison, the cumulative length distribution of the human Ref-Seq cDNA collection [37] is also shown.

NIH-BMAP-GI0, which represents the longest transcripts in the sample of pooled brain tissues. In preliminary studies, we have found (data not shown) that brain tissues may, in fact, be subject to a selection for longer 3'-UTR sequences.

From the analysis presented in Figure 3, we conclude that with care, quantitative comparisons between EST libraries can be used for the identification of tissue- or stage-specific phenomena, however, systematic variations in transcript sampling must be controlled. We explicitly identify EST libraries that can be safely compared without this bias by using the L-Divergence (Equation 3) to compare the estimated transcript length distributions of all pairs of libraries. Quantitative comparison can be made between pairs of EST libraries whose divergence is less than an empirically determined threshold value. To facilitate such analyses by other investigators, we have made available

our tools, data, and results on a web server [1]. Included in this package is a tool that will take as input a single EST library identifier and return a list of other libraries for which quantitative comparisons can safely be made.

While our examples have come from studies of mRNA 3'-processing, the systematic issues we have identified can bias similar analyses of phenomena such as tissue-specific changes in either global transcription patterns of many genes or processing of a specific gene. As long as the measured quantity has a dependence on (or causes a variation in) the transcript size, the effects we describe here can result in a false positive identification of a statistically significant, but biologically meaningless change. As a simple example, consider a gene with a 5000-nucleotide long transcript. If the relative expression level for this gene was compared between two EST libraries whose transcript size sampling resembled libraries NIH-BMAP-FA0 and NIH-

BMAP-FO0, respectively, we would expect a large discrepancy favoring the latter library, even if the true expression level in the two tissues was identical.

To further investigate the effect on measurement of gene expression, we examined the EST libraries originally generated by Okubo *et al.* [36] and later used as examples by Audic and Claverie [12]. These libraries consist only of 3'-ESTs, so we can assess the sampled transcript length distribution but not the insert length distribution. Briefly, these include three libraries, one generated from an HL60 human promyelocytic cell line, and two derived cell lines, a DMSO-induced granulocytoid cell line and a TPA-induced monocytoid cell line. In both the original and subsequent analyses, a number of genes were identified as differentially expressed between the cell types. Using the human RefSeq cDNA collection [37] as a reference, we generated transcript length distributions (Figure 4). Intriguingly, the DMSO-induced granulocytoid library appears to have sampled transcripts with a significantly longer length distribution than either of the other two libraries, which, in contrast, follow distributions quite similar to each other. Our analysis indicates the possibility that a portion of the genes identified as differentially expressed are false positives. Okubo *et al.* experimentally verified a number of genes as differentially expressed [36], however their measurements were on double-stranded cDNA saved at an intermediate stage of library construction. This brings up two distinct possibilities: either the list of differentially expressed genes includes a non-zero number of false positives due to differences in sampling of the transcripts during library construction, or the change in expression pattern upon DMSO-induced differentiation includes a systematic shift to significantly longer transcripts. Our analysis does not question the computational approaches of either the original [36] or subsequent manuscript [12], but rather the required assumption that the two originating samples are *identical* in preparation.

It is clear from our analysis that the common practice of computationally pooling EST libraries from a common tissue type or developmental stage, while increasing the gene coverage, is not likely to accurately reproduce relative expression levels of the original tissue sample. To see this, consider that our results indicate that in nearly all unbiased samples, the underlying transcript length distribution follows a roughly lognormal distribution, similar to the PACdb transcript, in Figure 1. As we have shown, distinct EST libraries will sample from this underlying distribution in a manner determined by preparation of the library. Simply combining disparate samples together will almost certainly distort the relative contributions of the different size transcripts. For example, the three groups of Brain Map libraries shown in Table 1 and Figure 3 represent roughly equal size samples from six different size

ranges of transcripts, which if pooled, would oversample long (>3000 nucleotides) transcripts at the expense of shorter ones.

## Conclusion
EST libraries contain a wealth of data regarding gene expression and specifics of transcript processing across a broad range of tissue- and cell-types and developmental stages. These data have been collected, however, by a wide range of researchers with varied procedures and goals, making large-scale comparative studies using these sequences problematic. The level of detail in the annotation regarding preparation techniques is not uniform, and in many cases incomplete (especially with regard to whether or not clones were size-selected). Even in libraries prepared without the two best understood systematics, normalization and subtraction, we find systematic variations in the sampling of transcripts. We have shown that estimates of the cDNA insert and originating transcript length distributions can be used to assess and compensate for systematics of library generation and enable quantitative analysis. Our tools and analyses are available via a web server [1] to help other researchers separate truly biologically meaningful changes in gene expression or processing from those that arise due to systematic biases.

## Methods
We downloaded the NCBI dbEST (release 030405, 03/04/2005) [38], and extracted 935 mouse libraries from a variety of tissues and organs, for a total of approximately 4.3 million ESTs. We used several sources for transcript reference sets, including 26,000 sequences from NCBI's mouse RefSeq transcript set [37], 20,515 non-redundant transcript sequences from ENSEMBL, version 27.33c [39], and the set of 39,000 transcripts implied by all putative mouse 3'-processing sites in PACdb [21]. Since EST libraries can contain contaminant sequences [40], we filtered and eliminated ESTs with evidence of vector/linker, *E.coli*, mitochondrial, or non-protein-coding RNAs. The filtered EST sequences were subsequently aligned to the reference transcripts using BLAT [41]. BLAT was chosen based on execution and ease of use, however, since we are looking for very high quality alignments of ESTs to mRNA sequences, the alignment problem is conceptually quite easy, and the choice of alignment tool is essentially immaterial. Each EST-transcript alignment was ranked in terms of quality and EST coverage, as defined in Equation 1 and 2.

$$alignmentScore = m - n - g \quad (1)$$

$$coverageOnEST = \frac{alignedLengthOnEST}{ESTLength} \quad (2)$$
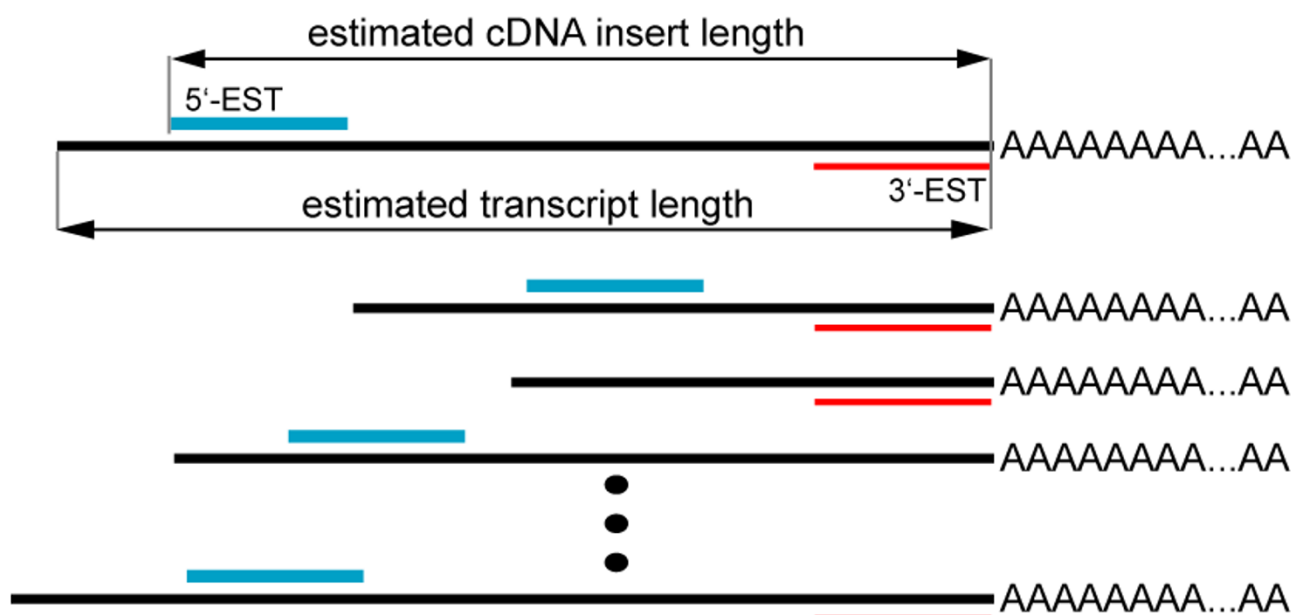
**Figure 5**
**Estimation of transcript and cDNA insert lengths through EST to reference transcript alignments**. Using EST-to-transcript alignments to estimate cDNA insert and originating transcript length distributions for EST libraries. 5'-EST alignment positions are used to determine the size of cDNA inserts. Reference transcripts matched by either 5'- or 3'- ESTs are used to estimate the sampled transcript lengths. Black lines represent polyadenylated mRNA transcripts, and 5'-ESTs and 3'-ESTs are represented in blue and red, respectively.

where $m$, $n$, and $g$ denote the count of matched, mismatched, and gapped positions, respectively, in the alignment. We retained only alignments for which *coverageOnEST* was greater than or equal to 0.9. If more than one reference transcript was aligned above this threshold, the alignment with the best *alignmentScore* was selected. In case of a tie, one transcript was selected at random.

To estimate the cDNA insert and originating transcript length distributions, we collected information as shown in Figure 5. We identified 5'- and 3'-ESTs as those that aligned in sense and anti-sense, respectively with the reference transcript, with care to identify libraries in which the reported sense of the EST was conceptually reverse-transcribed. The distance from 5'-EST starting points to the 3'-end of the matched transcripts was used to estimate the cDNA insert size, while the transcript length distribution was estimated by collecting the lengths of all reference transcripts matched by either 5'- or 3'-ESTs. An example of transcript and cDNA insert length distributions is presented in Figure 1. This analysis explicitly ignores randomly primed libraries. However, such libraries represent a relatively small fraction of the total (less than ten mouse libraries in dbEST are explicitly labeled as

randomly primed). In addition, even if a library was randomly primed, it would affect our estimate of the cDNA insert size, but not the transcript size, on which our comparative analysis is based.

For comparison of distributions, many metrics are available, including Euclidean distance, and Pearson or Spearman Correlation. We used a normalized L-Divergence [42], which is based on Shannon's entropy and is defined in Equation 3:

$$L = \frac{1}{2} \sum_i p_1(i) log \frac{p_1(i)}{\frac{p_1(i) + p_2(i)}{2}} + p_2(i) log \frac{p_2(i)}{\frac{p_1(i) + p_2(i)}{2}} \qquad (3)$$

where $p_1$ and $p_2$ denote the two density distributions being compared. This metric is bounded between 0 and 1, and furthermore has shown to be robust under a wide variability in probability distributions [42,43]. To characterize library-specific changes in 3'-processing, we used the ESTs from each library to identify probable 3'-processing sites as described previously [21]. Projected 3'-UTR lengths were calculated by measuring the genomic separation between the 3'-processing site and the stop codon of the assigned gene. Putative 3'-processing sites for a given EST

library are computationally normalized, such that all statistical analysis (*e.g.*, 3'-UTR length distributions) is performed on the set of unique sites, with no weight given to EST copy numbers.

We used an empirically determined distribution of 3'-UTR lengths as a function of transcript length (available in the supplement) as part of a two-step sampling process to generate an expected 3'-UTR length distribution for each mouse EST library. We first randomly sampled transcript lengths according to the library's estimated distribution, which was determined as shown in Figure 5. For each transcript length drawn, a second random draw was made from the 3'-UTR length distribution indicated from the PACdb data [21].

## Acknowledgements

## References

1. **EST library analysis web supplement** [http://harlequin.jax.org/estlib/]
2. Adams M, Kelly J, Gocayne J, Dubnick M, Polymeropoulos M, Xiao H, Merril C, Wu A, Olde B, Moreno R, Kerlavage A, McCombie W, Venter J: **Complementary DNA Sequencing: Expressed Sequence Tags and Human Genome Project.** *Science* 1991, **252:**1651-1656.
3. Adams M, Dubnick M, Kerlavage A, Moreno R, Kelley J, Utterback T, Nagle J, Fields C, Venter J: **Sequence identification of 2,375 human brain genes.** *Nature* 1992, **355:**632-634.
4. Gautheret D, Poirot O, Lopez F, Audic S, Claverie J: **Alternative Polyadenylation in Human mRNAs: A Large-Scale Analysis by EST Clustering.** *Genome Research* 1998, **8:**524-530.
5. Burke J, Wang H, Hide W, Davison D: **Alternative Gene Form Discovery and Candidate Gene Selection from Gene Indexing Projects.** *Genome Research* 1998, **8:**276-290.
6. Marra M: **Expressed sequence tags – ESTablishing bridges between genomes.** *Trends in Genetics* 1998, **14:**4-7.
7. Schmitt A, Specht T, Beckmann G, Dahl E, Pilarsky C, Hiznmann B, Rosenthal A: **Exhaustive mining of EST libraries for genes differentially expressed in normal and tumor tissues.** *Nucleic Acids Research* 1999, **27:**4251-4260.
8. Ewing B, Green P: **Analysis of expressed sequence tags indicates 35,000 human genes.** *Nature Genetics* 2000, **25:**232-234.
9. Takasuga A, Hirotsune S, Itoh R, Jitohzono A, Suzuki H, Aso H, Sugimoto Y: **Establishment of a high throughput EST sequencing system using poly(A) tail-removed cDNA libraries and determination of 36 000 bovine ESTs.** *Nucleic Acids Research* 2001, **29:**e108. 1–7
10. Zhu Y, King B, Parvizi B, Brunk B, Stoeckert C Jr, Quackenbush J, Richardson J, Bult C: **Integrating computationally assembled mouse transcript sequences with the mouse Genome Informatics(MGI) database.** *Genome Biology* 2003, **4(2):**R16.1-R16.8.
11. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J: **The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes.** *Nucleic Acids Research* 2005, **33:**D71-D74.
12. Audic S, Claverie J: **The significance of digital gene expression profiles.** *Genome Research* 1997, **7:**986-995.
13. Bonaldo M, Lennon G, Soares M: **Normalization and Subtraction: Two Approaches to Facilitate Gene Discovery.** *Genome Research* 1996, **6:**791-806.
14. Claverie J: **Computational methods for the identification of differential and coordinate gene expression.** *Human Molecular Genetics* 1999, **8(21):**1821-1932.
15. Megy K, Audic S, Claverie J: **Heart-specific genes revealed by expressed sequence tag(EST) sampling.** *Genome Biology* 2002, **3(12):**research0074.1-0074.11.
16. Wolfsberg T: **A comparison of expressed sequence tags (ESTs) to human genomic sequences.** *Nucleic Acids Research* 1997, **25:**1626-1632.
17. Gupta S, Zink D, Kom B, Vingron M, Haas S: **Strengths and -weaknesses of EST-based prediction of tissue-specific alternative splicing.** *BMC Genomic* 2004, **5(72):**1-8.
18. Kan Z, Rouchka E, Gish W, States D: **Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs.** *Genome Research* 2001, **11:**889-900.
19. Beaudoing E, Gautheret D: **Identification of Alternate Polyadenylation Sites and Analysis of their Tissue Distribution Using EST Data.** *Genome Research* 2001, **11:**1520-1526.
20. Yan J, Marr T: **Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat.** *Genome Research* 2005, **15:**369-375.
21. Brockman J, Singh P, Liu D, Quinlan S, Salisbury J, Graber J: **PACDB: PolyA cleavage site and 3'UTR database.** *Bioinformatics* 2005, **21:**3691-3693.
22. Soares M, Bonaldo M, Jelene P, Su L, Lawton L: **Construction and characterization of a normalization cDNA library.** *Proc Natl Acad Sci USA* 1994, **91:**9228-9232.
23. Schena M, Shalon D, Davis R, Brown P: **Quantitative monitoring of gene expression patterns with a complimentary DNA microarray.** *Science* 1995, **270:**467-470.
24. Nguyen C, Rocha D, Granjeaud S, Baldit M, Bernard K, Naquet P, Jordan B: **Different gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones.** *Genomics* 1995, **29:**207-216.
25. Zhao N, Hashida H, Takahashi N, Misumi Y, Sakaki Y: **High-density cDNA filter analysis: a novel approach for large-scale, quantitative analysis of gene expression.** *Gene* 1995, **156:**207-213.
26. Skrabanek L, Campagne F: **TissueInfo: high-throughput identification of tissue expression profiles and specificity.** *Nucleic Acids Research* 2001, **29:**el02.
27. Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K: **Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression.** *Nature Genetics* 1992, **2:**172-179.
28. Brown A, Kai K, May M, Brown D, Roopenian D: **ExQuest, a novel method for displaying quantitative gene expression from ESTs.** *Genomics* 2004, **83:**528-539.
29. Lennon G, Lehrach H: **Hybridization analysis of arrayed cDNA libraries.** *Trends in Genetics* 1991, **7:**314-317.
30. Bonaldo M, Bair T, Scheetz T, Snir E, Akabogu I, Bair J, Berger B, Crouch K, Davis A, Eyestone M, Keppel C, Kucaba T, Lebeck M, Lin J, de Melo A, Rehmann J, Reiter R, Schaefer K, Smith C, Tack D, Trout K, Sheffield V, Lin J, Casavant T, Soares M: **1274 Full-Open reading frames of transcripts expressed in the developing mouse nervous system.** *Genome Research* 2004, **14:**2053-2063.
31. Colgan D, Manley J: **Mechanism and regulation of mRNA polyadenylation.** *Development* 1997, **11:**2755-2766.
32. Gray N, Wickens M: **Control of translation initiation in animals.** *Annual Review of Cell and Developmental Biology* 1998, **14:**399-458.
33. Zhao J, Hyman L, Moore C: **Formation of mRNA 3' Ends in Eukaryotes: Mechanism, Regulation, and Interrelationships with Other Steps in mRNA Synthesis.** *Microbiology And Molecular Biology Reviews* 1999, **63(2):**405-445.
34. Mignone F, Gissi C, Liuni S, Pesole G: **Untranslated regions of mRNAs.** *Genome Biology* 2002, **3:**reviews0004.1-0004.10.
35. Kuersten S, Goodwin B: **The power of the 3'UTR: translational control and development.** *Genetics* 2003, **4:**626-637.
36. Okubo K, Itoh K, Fukushima A, Yoshii J, Matsubara K: **Monitoring cell physiology by expression profiles and discovering cell**

**type-specific genes by compiled expression profiles.** *Genomics* 1995, **30:**178-186.

37. Pruitt K, Tatusova T, Maglott D: **NCBI Reference sequence (Ref-Seq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Research* 2005, **33:**D501-D504.

38. Boguski M, Lowe T, Tolstoshev C: **dbEST-database for "expressed sequence tags".** *Nature Genetics* 1993, **4:4:**332-333.

39. **Ensembl cDNA collection, v27.33c** [ftp://ftp.ensembl.org/pub/release-27/mouse-27.33c/data/fasta/cdna/]

40. Sorek R, Safer H: **A novel algorithm for computational identification of contaminated EST libraries.** *Nucleic Acids Research* 2003, **31:**1067-1074.

41. Kent J: **BLAT – the BLAST-like alignment tool.** *Genome Research* 2002, **12:**656-664.

42. Lin J: **Divergence Measures based on the Shannon Entropy.** *IEEE Transaction on Information Thoery* 1991, **37:**145-151.

43. Liu D, Singh G: **Entropy based Clustering for High Dimensional Genomic Data Sets.** In *Proc of 2002 SIAM International Conference on Data Mining, Volume Workshop on Clustering High Dimensional Data Sets* 2002:27-36.

44. Curwen V, Eyras E, Andrews T, Mongin E, Searle S, Clamp M: **The Ensembl Automatic Gene Annotation System.** *Genome Research* 2004, **14:**942-950.