In-depth cDNA Library Sequencing Provides Quantitative Gene Expression Profiling in Cancer Biomarker Discovery

Wanling Yang*, Dingge Ying, and Yu-Lung Lau
Department of Paediatrics and Adolescent Medicine, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China.
*Corresponding author. E-mail: yangwl@hkucc.hku.hk
DOI: 10.1016/S1672-0229(08)60028-5

Quantitative gene expression analysis plays an important role in identifying differentially expressed genes in various pathological states, gene expression regulation and co-regulation, shedding light on gene functions. Although microarray is widely used as a powerful tool in this regard, it is suboptimal quantitatively and unable to detect unknown gene variants. Here we demonstrated effective detection of differential expression and co-regulation of certain genes by expressed sequence tag analysis using a selected subset of cDNA libraries. We discussed the issues of sequencing depth and library preparation, and propose that increased sequencing depth and improved preparation procedures may allow detection of many expression features for less abundant gene variants. With the reduction of sequencing cost and the emerging of new generation sequencing technology, in-depth sequencing of cDNA pools or libraries may represent a better and powerful tool in gene expression profiling and cancer biomarker detection. We also propose using sequence-specific subtraction to remove hundreds of the most abundant housekeeping genes to increase sequencing depth without affecting relative expression ratio of other genes, as transcripts from as few as 300 most abundantly expressed genes constitute about 20% of the total transcriptome. In-depth sequencing also represents a unique advantage of detecting unknown forms of transcripts, such as alternative splicing variants, fusion genes, and regulatory RNAs, as well as detecting mutations and polymorphisms that may play important roles in disease pathogenesis.

Key words: cDNA sequencing, sequencing depth, expressed sequence tag, sequence-specific subtraction, biomarker, mutation detection

Introduction

High-throughput cDNA sequencing involves selecting clones from a cDNA library at random and performing automated sequencing read of their insert libraries (*1*). In 1992, a database called dbEST was established to serve as a collection of expressed sequence tags (ESTs) (*2*). In the last one and half decade, millions of clones derived from various sources have been sequenced and deposited into dbEST. As of May 2008, there are more than eight million EST entries of human origin, and 50 million total entries in dbEST from various species (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

The sources of ESTs include tissues of different developmental stages and different physiological and pathological states, which have the potential to be used to answer many important biological questions. For example, for the human libraries, about half of them were generated from normal tissues and another half from tumor tissues or cultured tumor cell lines. So they provide opportunities for detecting cancer-related differential gene expression, alternative splicing events, and other rare forms of transcripts. Indeed, EST sequences have been used to estimate mRNA abundance in various tissues (*3*, *4*), to detect tissue-specific genes (*5*), and to draw inference on differential gene expression (*4*, *6*).

In principle, the number of mRNAs of a particular gene in a library is roughly proportional to the abundance of this mRNA in the transcriptome of the tissue used to prepare the library, given that the library was not treated in any way to significantly change the relative abundance of different transcripts, and

the libraries have certain sequencing depth relative to the size of the transcriptome of the source tissue. However, there are problems with the way that ESTs were generated, which prevented them from being fully used for gene expression analysis. A major purpose of the early stages of EST sequencing was to discover new genes. Various methods were invented to reduce repeated sequencing of known abundant genes and to increase the probability of gene discovery. Normalization and subtraction of different libraries, or libraries undergone different treatment, were used to increase the chance of detecting new genes or differentially expressed genes (*7–9*). Certain libraries underwent PCR amplification process, which could change relative gene abundance due to uneven amplification. Some libraries were generated by random priming (*10*), which may amplify certain genes several magnitudes more efficiently than others due to primer annealing preference, amplification efficiency, or potential secondary structure of some mRNAs. Fractionation of the cDNA pool and other preparation processes may also change the ratio of different transcripts. The number of sequenced clones of various cDNA libraries ranges from a few hundreds to tens of thousands. Relative to the size of the transcriptome, libraries are seriously under-sampled and even for the bigger libraries, it is estimated that only about 60% of expressed genes in an tissue or cell line get represented (*8*).

Here we have chosen a subset of cDNA libraries (non-normalized libraries with at least 5,000 sequenced clones) that better reflect gene expression profiles, and have detected significant patterns of differential gene expression, expression correlation, and rare forms of transcripts of certain abundant genes. We propose that with increased sequencing depth, and using technologies such as sequence-specific subtraction of a small fraction of the most abundantly expressed housekeeping genes, cDNA library sequencing could provide a more accurate and quantitative method for extracting gene expression information for less abundant genes, thus shed light on the function and regulation of the genes and their roles in disease pathogenesis.

Results

Differential gene expression related to cancer

Various comparisons have been done to detect genes that are differentially expressed in tumors of a particular tissue origin (*4*, *11*). However, due to the lack of quality libraries for a given tissue, any comparison this way may be underpowered. On the other hand, there are common features to tumors of different tissue origins, such as lack of differentiation, dysregulation in cell cycle control, genome instability, and evasion of apoptosis. These commonalities could very well be reflected in gene expression patterns. Here using a subset of non-normalized libraries with at least 5,000 sequenced clones, which can be divided into groups of normal tissues, tumor tissues, and cultured tumor cells (Table 1, see Materials and Methods for details), we have analyzed the expression of some abundantly expressed genes as well as genes that are known to be involved in cancer (Figure 1). The analysis results were also compared with results from libraries of uncharacterized preparation, normalization and subtraction.

In Figure 1, we demonstrated differential expression of certain genes either using data directly downloaded from UniGene or by EST sequence analysis of our own. It is worth noting that for most of these genes, there is an apparent up-regulation in tumor cell lines, and an intermediate up-regulation in tumor tissues. It is consistent with the notion that the cancer cell lines have undergone selection processes and are probably clones from the fastest growing cancer cells. As shown in Figure 1B, the differential expression of *BIRC5* among the three groups is much more significant when analyzed using non-normalized li-

Table 1 Non-normalized libraries used in the analysis of this study

Tissue origin of libraries	No. of libraries	Total sequence entries	Library size (total entries)		
			Minimum	Maximum	Median
Normal bulk tissue	43	429,781	5,235	23,703	9,180
Normal cell line	12	124,308	6,462	18,479	9,541
Tumor bulk tissue	23	261,446	5,385	25,235	8,430
Tumor cell line	64	908,856	5,178	41,936	11,945
Total	142	1,724,391	5,178	41,936	10,583

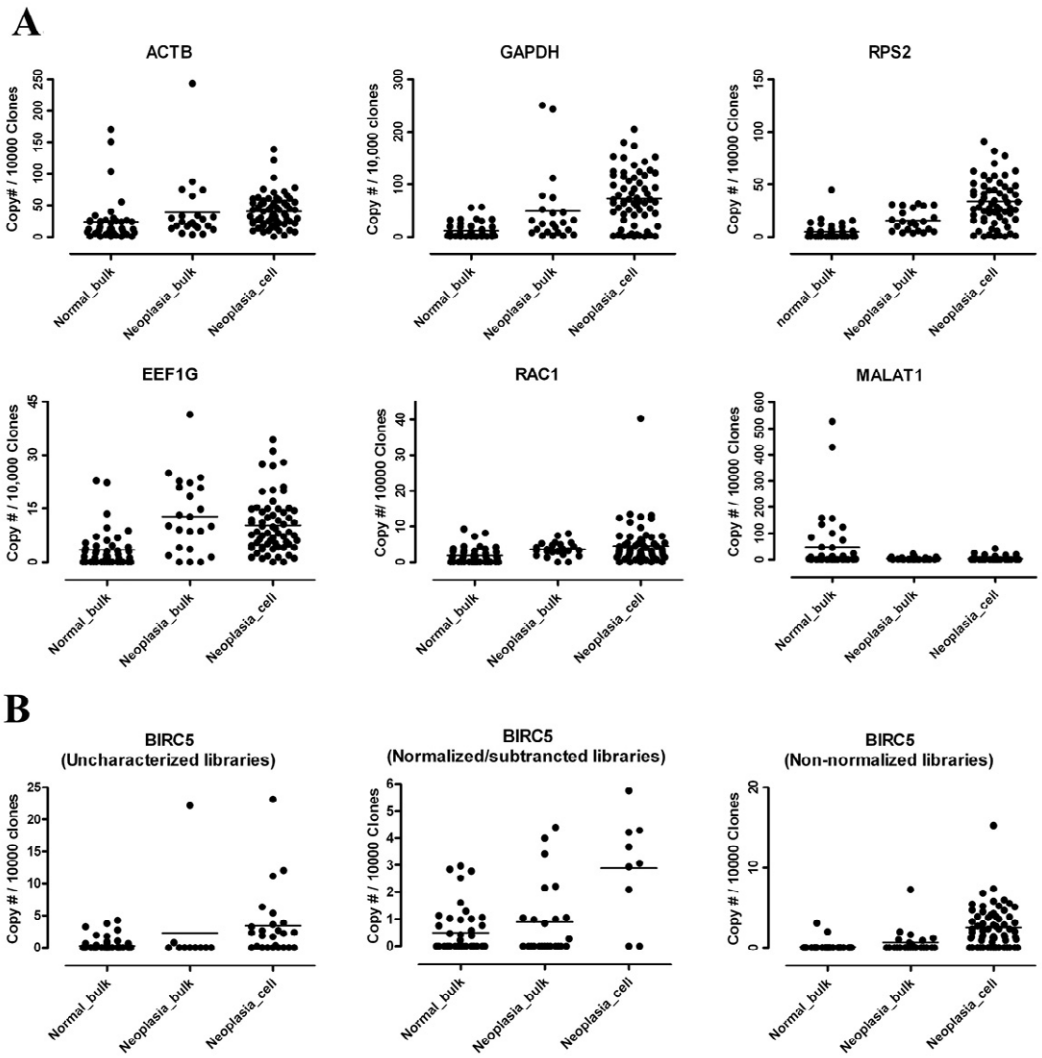

Figure 1 Differentially expressed genes detected by EST analysis using 130 non-normalized libraries. The libraries were grouped according to their tissue origin as "normal tissue (normal_bulk)", "cancer tissue (neoplasia_bulk)", and "cultured cancer cell line (neoplasia_cell)". Differential gene expression was analyzed by their detected copy numbers per 10,000 sequenced clones in each library. **A**. Differential expression for *ACTB, GAPDH, RPS2, EEF1G, RAC1*, and *MALAT1*. Comparisons showed significant P values calculated by unpaired nonparametric t-test with Welch's correction: non-muscle beta actin (*ACTB*) between normal_bulk and neoplasia_bulk, P=0.04; between normal_bulk and neoplasia_cell, P=0.011; glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*) between normal_bulk and neoplasia_bulk, P=0.0016; between normal_bulk and neoplasia_cell, P<0.0001; between neoplasia_bulk and neoplasia_cell, P=0.0002; ribosomal protein S2 (*RPS2*) between normal_bulk and neoplasia_bulk, P=0.0022; P<0.0001 between normal_bulk and neoplasia_cell as well as between neoplasia_bulk and neoplasia_cell; eukaryotic translation elongation factor 1 gamma (*EEF1G*) between normal_bulk and neoplasia_bulk, P=0.0004; between neoplasia_bulk and neoplasia_cell, P<0.0001; ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1, *RAC1*) between normal_bulk and neoplasia_bulk, P=0.0034; between normal_bulk and neoplasia_cell, P=0.0013; between neoplasia_bulk and neoplasia_cell, P=0.043; metastasis associated lung adenocarcinoma transcript 1 (non-protein coding) (*MALAT1*) between normal_bulk and neoplasia_bulk, P=0.0084; and between normal_bulk and neoplasia_cell, P=0.0098. **B**. Differential expression of *BIRC5* analyzed by libraries of different preparations. For baculoviral IAP repeat-containing 5 (survivin) (*BIRC5*) in uncharacterized libraries, between normal_bulk and neoplasia_cell, P=0.0067; between neoplasia_bulk and neoplasia_cell, P=0.017. For normalized/subtracted libraries, between normal_bulk and neoplasia_cell, P=0.0065; between neoplasia_bulk and neoplasia_cell, P=0.0090. For non-normalized libraries between normal_bulk and neoplasia_cell, P<0.0001; between neoplasia_bulk and neoplasia_cell, P<0.0001.

braries. While the same trend is observed, the data are less consistent when the expression of this gene is compared among three tissue origins in uncharacterized libraries (with unknown preparation process) and normalized/subtracted libraries. This reflected the effect of library preparation on gene expression profiling and comparison. There are reports that *RAC1b*, the splicing variant form of *RAC1*, is involved in various cancers (*12*, *13*). Using the 57-nucleotide sequence from the insertion exon of *RAC1b*, we searched dbEST for entries corresponding to this variant and examined its expression in the 142 selected non-normalized libraries. From Table 2, we can clearly see that the six *Rac1b* ESTs detected in these libraries were all neoplasia origin, a result consistent with the reports showing that the variant form is involved in important tumor growth processes.

Detection of co-regulation of gene expression and rare forms of gene expression

Since ESTs are capable of expression profiling for abundant genes, it should be able to detect genes that are co-expressed or co-regulated for these genes. Here we examined whether there is an expression correlation between *ACTB* and *ACTG1*, the two non-muscle cytoskeletal actins. The two actins co-exist in most cell types as components of the cytoskeleton, and as mediators of internal cell motility. They form dimers to be functional, and therefore an expression correlation is assumed. As shown in Figure 2, there is very good expression correlation between the two genes as analyzed by ESTs in the 142 libraries. Not surprisingly, the expression correlation between the two genes is most convincing when analyzed by the selected non-normalized libraries, with a less correlation observed in the normalized/subtracted libraries while no correlation is found in the uncharacterized libraries. Similarly, high expression correlation is also observed between *KRT8/18*, two genes known to interact and are involved in many cancer types (Figure 2D) (*11*).

Compared with microarrays and SAGE (serial analysis of gene expression), EST sequencing has the unique advantage of detecting transcripts of rare and unknown forms, such as alternative splicing, fusion genes, and other RNA species, given that the preparation process of cDNA library was not excluding those forms. We examined alternative splicing forms for *CD44*, a gene that is known to be present in many different splicing forms and involved in metastasis of cancer. Indeed, ESTs are capable of detecting various *CD44* alternative splicing forms (data not shown). With increased sequencing depth of cDNA libraries, more rare forms of transcripts of different genes could be detected and have the potential to be used as diagnosis and prognosis markers.

Evaluation of sequencing depth and false negative detection of expressed genes

The possibility of using ESTs to analyze gene expression profile has been proposed before (*3*, *10*). However, variations in library preparations and more importantly, serious under-sampling regarding the sequenced clones per library relative to the size of total transcriptome, prevented usage of ESTs as a real quantitative measure for gene expression profiling. The question then is what would be the targeted sequencing depth to make cDNA sequencing a desirable method for accurately measuring of gene expression in a quantitative manner. Let's assume the process of picking clones for sequencing is truly random. Then representation of a gene of given expression level in a library of certain size follows a binomial distribution with parameters P and N, where P is the probability of observing x number of tags for a gene, randomly sampling from total N cDNA clones. The expression

Table 2 Expression of *RAC1b* variant form detected in the 142 non-normalized libraries

Library name	Tissue	Histology	Total sequenced clones	*RAC1b* detection (copies)	Library origin
NIH_MGC_70	pancreas	neoplasia	16,633	1	cell line
NIH_MGC_15	colon	neoplasia	14,224	1	cell line
NIH_MGC_98	brain	neoplasia	12,808	1	cell line
NIH_MGC_42	pancreas	neoplasia	10,751	1	cell line
NIH_MGC_101	lung	neoplasia	9,166	1	cell line
NCI_CGAP_GU1	uncharacterized tissue	neoplasia	5,550	1	bulk

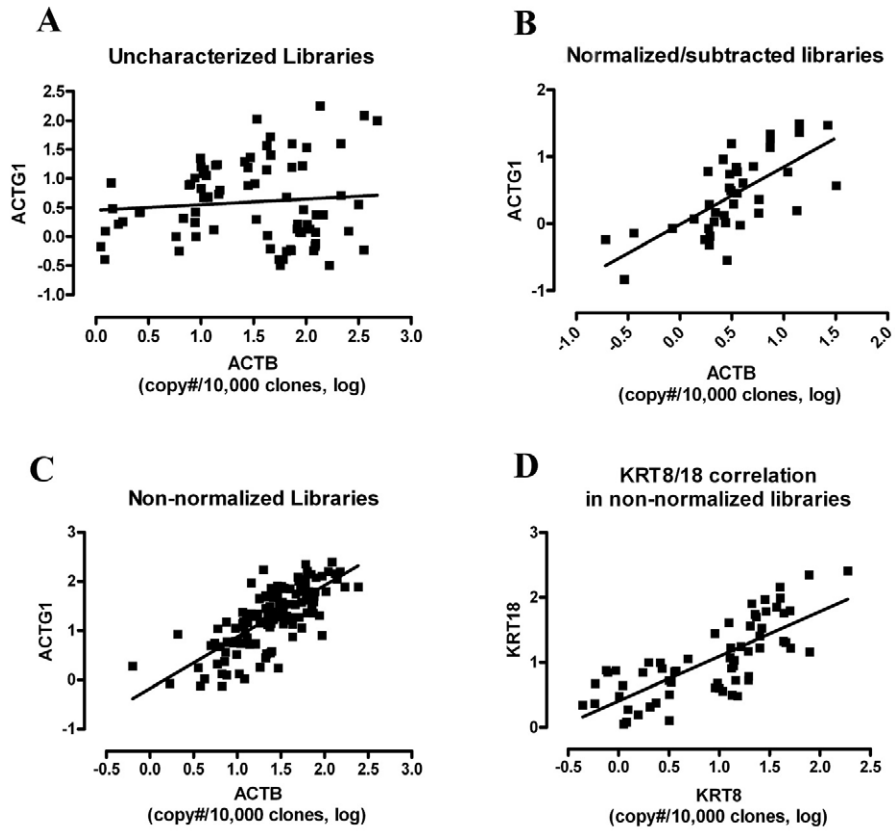

Figure 2 Expression correlation between *ACTB* and *ACTG1* and between *KRT8/18*. **A**. Correlation between *ACTB* and *ACTG1* in the uncharacterized libraries. There is no significant expression correlation between the two genes detected (correlation coefficient r^2=0.0086, P=0.43). **B**. Correlation between *ACTB* and *ACTG1* among the normalized/subtracted libraries. The correlation is significant with r^2=0.47 and P<0.0001. **C**. Correlation between *ACTB* and *ACTG1* among the non-normalized libraries. The correlation is significant with r^2=0.56 and P<0.0001. **D**. Expression correlation between *KRT8* and *KRT18* in non-normalized libraries. The correlation is significant with r^2=0.57 and P<0.0001.

level for a gene is t, and t assumes 0.01 when the gene constitutes one percent of all the transcripts in a transcriptome. Since the expression level t is considerably small compared with N, the binomial distribution can be approximated by Poisson distribution. Therefore, the probability of having x tags for a given gene in a library of total size N, with gene expression level of t, where $\lambda = Nt$, is:

$P(x) = (e^{-\lambda}\lambda^x)/x!$

For negative representation, $x = 0$, then $P(0) = 1/e^{\lambda}$. In Figure 3, we showed the relationship of false negative detection (expressed genes not being represented) with the total number of sequenced clones in a library for different gene expression levels. It is obvious that when expression levels are low, such as lower than 1 copy in 50,000 transcripts, the probability of its being detected in libraries of current sizes is extremely low. Similarly, for libraries with less than 10,000 sequenced clones, the probability of false negative is high for probably majority of expressed genes.

In Table 3, we showed the targeted library sizes when a certain false negative detection level is considered acceptable. With the sequencing cost goes down rapidly, sequencing of larger number of cDNAs from a certain tissue could become reality in the near future. If hundreds of thousands, or even millions of clones can be sequenced for a given tissue or cell line, extremely valuable expression information can be extracted and analyzed, thus shed light on gene functions and cellular activities.

Increasing sequencing depth by sequence-specific subtraction

NCBI grouped most EST entries to UniGene clusters according to their chromosomal locations (*14*). In UniGene build 210, there are 123,687 clusters for human (http://www.ncbi.nlm.nih.gov/UniGene/

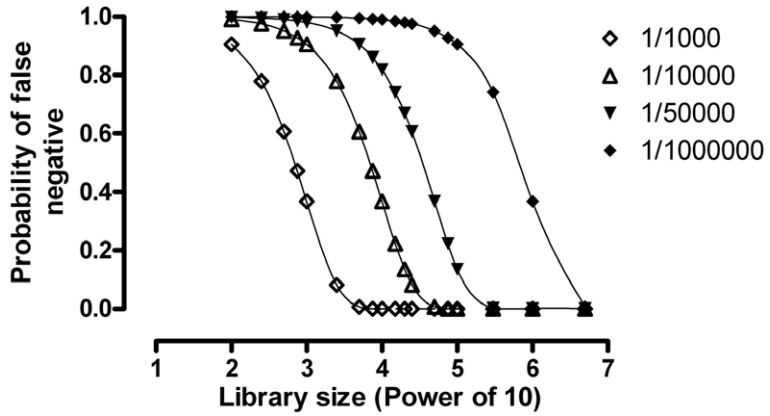

Figure 3 Probability analysis of false negative detection in relationship with library sizes and gene expression levels. The analysis is based on library sizes (x-axis, to the power of 10) and gene expression levels at 1 copy of mRNA in 1,000, 10,000, 50,000, and 1 million total transcripts, using Poisson distribution as described in Results. The y-axis stands for the probability of false negative detection—undetected when the gene is really expressed.

Table 3 Targeted library sizes for certain false negative rates for various expression levels

Expression level	Targeted library sizes			
(copy number in transcriptome)	$P(0)$=30%	$P(0)$=20%	$P(0)$=10%	$P(0)$=5%
1/100	120	160	230	300
1/1,000	1,200	1,600	2,300	3,000
1/10,000	12,000	16,000	23,000	30,000
1/100,000	120,000	160,000	230,000	300,000
1/1,000,000	1,200,000	1,600,000	2,300,000	3,000,000

UGOrg.cgi?TAXID=9606). The clusters have a wide range of different numbers of sequence entries from one entry to nearly 50,000 entries. It is easily seen that a group of most abundantly expressed genes constitute a significant portion of the transcriptome. We ranked the UniGene clusters according to the number of sequence entries they have and plotted the portion of UniGene clusters from the most abundant ones with the portion of their entries in the total UniGene sequences (Figure 4). It can be seen that sequences from 2.3% of most abundant UniGene clusters constituted 53% of all the sequence entries. Similarly, entries from 300 most abundant genes constituted about 20% of the total entries. Some most abundant genes such as *EEF1A1*, *GAPDH*, and *ACTB* were each sequenced more than 25,000 times. Thus here we propose that, if a process of sequence-specific subtraction is used to subtract the most abundant transcripts, sequencing depth can be significantly increased without increasing the cost or affecting the relative expression ratio of other genes. The sequence-specific subtraction could also be done in a tissue-specific fashion. For libraries of liver origin, we found that transcripts from albumin constituted around 10% of all the transcripts (Table 4). Thus sequence-specific subtraction for albumin alone could increase sequencing depth significantly for their libraries.

Discussion

A deep understanding of the genes expressed in different tissues, developmental stages, and pathological states is a vital step towards understanding of their biological functions. cDNA library sequencing (EST), SAGE, and microarrays have all been widely used to help the understanding of gene expression with different advantages and disadvantages. There are many issues for microarrays and the information extracted from microarray experiments is probably not proportional to the scale of experiments performed so far using this technology. SAGE usually has more sequencing depth than EST libraries. However, the accuracy of using a SAGE tag to reflect expression of the corresponding gene is compromised by tag-sharing by different genes, and generation of more tags by a given gene due to internal priming, alternative polyadenylation or alternative splicing.

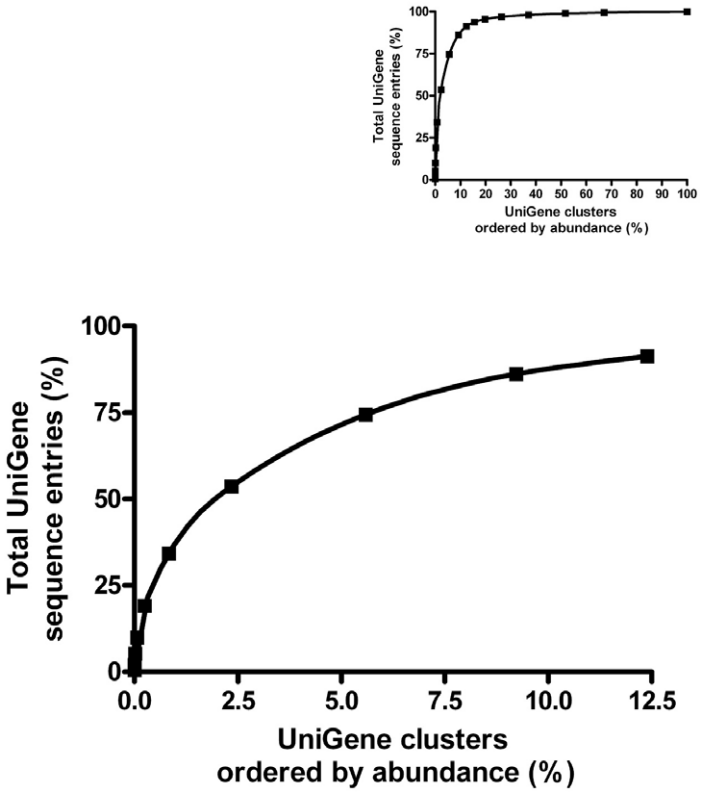

Figure 4 Percentage of sequence entries in the total transcriptome by the most abundant genes. The x-axis stands for the portion of UniGene clusters ranked by abundance from the most abundant to the least abundant clusters. The y-axis stands for the cumulative portion of total UniGene entries constituted by the most abundant UniGene clusters. The inset is the full scale of the same figure.

Table 4 Albumin sequences in libraries of liver origin*

Library name	Total sequenced clones	Detected albumin sequences	Portion of albumin (%)	Library histology	Library protocol
LIVER2	6,715	2,336	34.79	normal	uncharacterized treatment
TLIVE2	8,656	2,202	25.44	neoplasia	uncharacterized treatment
human hepatoblastoma cDNA	7,898	280	3.55	neoplasia	uncharacterized treatment
Stratagene liver (#937224)	8,417	1,022	12.14	normal	non-normalized
Homo sapiens FETAL LIVER	10,027	907	9.05	normal	non-normalized
NIH_MGC_76	11,960	1,046	8.75	normal	non-normalized
GLC	19,285	1,272	6.60	normal	uncharacterized treatment
GKC	17,736	1,146	6.46	neoplasia	uncharacterized treatment
779 (synonym: hncc1)	10,690	682	6.38	normal	uncharacterized treatment

*Including libraries with 5,000 sequenced clones or more, and tissues only (no libraries from cell lines).

cDNA library sequencing is a simple method that has the best potential in many aspects. With the prospect of dramatically reduced sequencing cost, it becomes realistic to sequence hundreds of thousands or even millions of clones or just cDNA molecules for a give tissue or cell line. So if the issues in library preparation and sequencing depth are solved, cDNA pool/library sequencing could be the most accurate and most quantitative method of gene expression profiling. It also has the unique advantage of detecting rare forms of transcripts, such as rare alternative splicing forms, fusion gene transcripts, and non-coding RNA species such as microRNAs. It can also be used to detect mutations, polymorphisms, and RNA editing events if relevant process can be adopted to ensure sequencing quality (*15*).

From the current dbEST, irrespective of library preparation, there are 370 libraries sequenced so far

with more than 5,000 sequenced clones. This magnitude is not nearly comparable to the scale of microarrays that have been performed. In this study we focused our analysis on the 142 non-normalized libraries supposedly better reflecting relative gene expression levels. Even with the limited scale, analysis of differential expression and expression correlation still revealed significant information as we have demonstrated here. This probably only reflected a tip of the iceberg that cDNA sequencing is capable of detecting in terms of gene expression information.

Transcription profiles are used for molecular classification of cancers, as well as prediction of prognosis and tumor progression (*16*, *17*). In this study we have chosen some genes known to be involved in various cancers and tested whether EST data can reflect differential expression of these genes in cancer. *BIRC5* (*Survivin*) is a member of the inhibitor of apoptosis protein family, and is known to be elevated in many human cancer tissues. It is involved in inhibition of apoptosis and regulation of mitosis in many tumor types (*18*, *19*). In a study on all the UniGene clusters, we have identified *BIRC5* as a gene that is on the top of the list of up-regulated genes in tumor and tumor cell lines (unpublished data). This is consistent with the result here using EST analysis, which found that this gene is up-regulated in many tumor cell lines regardless of their origin, and probably up-regulated in tumor tissues as well. Similarly, *RAC1b* is known as a tumor-specific splicing variant form of *RAC1*, with an insertion of 19 amino acids next to the switch II region of *RAC1*. It was found to be elevated in colorectal cancer at various stages of tumor progression (*12*). It is involved in promoting cellular transformation (*13*) and proliferation (*20*). Quite convincingly, *RAC1b* was only detected in 6 libraries from the 142 libraries used and all of them were neoplasia origin.

Metastasis associated lung adenocarcinoma transcript 1 (*MALAT1*, GeneID: 378938, UniGene cluster: Hs.642877), also known as nuclear enriched abundant transcript 2 (*NEAT2*) or Pro1073, is a non-coding RNA with undefined function. The UniGene cluster corresponding to this gene (Hs.642877) has a total of 16,287 EST entries according to UniGene build 210, which put it among the top 15 most abundantly expressed genes in human genome. The gene was proposed to be functioning in mRNA metabolism and was demonstrated to have an intimate association with SC35 nuclear speckles in both human and mouse cells (*21*). However, EST data suggest a potential down-regulation of this gene in tumor and tumor cells, something worth further investigation for a gene with such an abundant expression but unclear function.

In this study we only analyzed differential gene expression between normal tissue and tumor tissue or cells, irrespective of tissue type or other biological information due to lack of power for detailed analysis. If more libraries of enough sequencing depth can be generated, more detailed comparison can be performed to gain information on the genes expressed from different tumor types and stages. On the other hand, certain changes in gene expression could be common to most of the tumors or tumor cell lines. From this analysis, it seems that *RAC1*, *RAC1b*, and *BIRC5* are up-regulated in a variety of tumor types. The intermediate change of some of the genes analyzed in tumor tissues, with higher expression than in normal tissues but lower expression level than in tumor cell lines, probably reflected the heterogeneity of tumor tissues.

An interesting finding of our analysis is the variability of expression level of the most abundant and ubiquitously expressed genes, such as *GAPDH*, *ACTB*, and *ACTG1* (Figure 1). mRNAs of these genes have been used as internal controls when comparing message levels between different tissues or different treatment. The fact that messages of those genes vary greatly both between tissues and by other biological processes (such as malignancy) calls for caution when using these genes as controls.

Regulation and co-regulation of genes often has functional implications (*22*) and can be used to identify unknown members of the same signal transduction pathways (*23*). Although much efforts have been spent to detect co-expressed genes (*24*, *25*), we still do not fully understand gene expression correlation patterns in various physiological and pathological states. cDNA sequencing is capable of providing an accurate and quantitative way of finding gene orders. Detection of co-expressed genes by more in-depth cDNA library sequencing may shed new light on protein interaction, signal transduction pathways, and transcriptional regulation.

Using ESTs for expression analysis has been proposed before (*3*, *24*). However, how well can ESTs reflect gene expression information is very controversial. For example, an earlier study analyzed 1,573 libraries, and even the known most abundant, ubiquitous housekeeping genes were only detected in less than one-third of the libraries used (*24*). This demonstrates the importance of sequencing depth in faith-

fully reflecting gene expression information, as well as the methods used in library preparation. In addition to the 142 libraries we used in most analysis, there are 61 normalized libraries, 10 subtracted libraries, and 152 libraries with uncharacterized treatment containing more than 5,000 sequenced clones. We found that for the non-normalized libraries, there is a correlation coefficient of 0.56 between the expression of *ACTB* and *ACTG1* (Figure 2). The trend is similar for the normalized and subtracted libraries, with a correlation coefficient of 0.47. However, there is no significant correlation between the two genes when analyzed by the uncharacterized libraries. It is unclear how normalization process would change the relative expression level of less abundant transcripts. The maintained correlation of *ACTB* and *ACTG1* in normalized libraries could be due to the fact that the two genes are equally abundant, as the method is designed to create libraries containing equal representation of all sequences (*7*). We compared the expression level of *ACTB* and *ACTG*1 in libraries of different treatment. It seems that normalization and subtraction did significantly reduce the detection level of these abundant genes, with little effect on their relative ratio. However, the libraries with uncharacterized treatment seem to have significantly changed the relative ratio of the genes (Table 5).

It has been realized that much larger portion of genomes are transcribed than anticipated from whole genome annotations, and non-protein-encoding transcripts comprise a substantial fraction of the human genome (*26–28*). More and more studies indicate the important biological roles of non-coding RNAs (*29*). Due to the short length and lack of polyA tail for such small RNA molecules, they are missed by most cDNA library preparation processes as a result of fractionation in selecting larger molecules and cDNA synthesis using poly-dT primers. Modification in the cDNA library preparation process, or sequencing of cDNA pools without library preparation could result in inclusion of these species and allow evaluation of their expression level in various biological states, as well as correlation with the expression level of other protein coding genes (*30*, *31*). cDNA library sequencing also provides a unique advantage in detecting alternative splicing forms (*32*, *33*). In addition, other transcripts such as fusion genes from splicing (*34*) or chromosomal translocation, variants reflecting somatic mutations or RNA editing (*15*), could also be detected through in-depth sequencing.

It was estimated that the number of genes expressed in a cell lies between 10,000 and 15,000 (*35–37*), and the total transcripts in a single cell was estimated to be between 300,000 (*35*, *38*) to a few million (*39*, *40*). Patanjali *et al* estimated that the copy numbers of expressed genes could vary from a single copy to 200,000 copies per cell (*7*), while Galau *et al* estimated that one-third of the mRNA in a single cell type is made up of species present at only 1–10 copies per cell (*41*). Accordingly, it is safe to say that the majority of genes express at a level below one copy per 10,000 transcripts. Future gene expression detection methods should be made capable of detecting species at this expression level in a high-throughput format. To reach this goal, sequencing of hundreds of thousands to millions of clones in a pool/library would be necessary. The need of much increased sequencing depth is also pointed by Zhu *et al* (*40*) and Stern *et al* (*42*).

SAGE data also point to the need for more sequencing depth for detecting low abundant genes (*42*). At the range of 100,000–150,000 total tags sequenced in a library, the number of unique tags is in the range of 30,000–40,000. The two SAGE libraries containing about 400,000 sequenced tags have about 80,000 unique tags. It is shown that the fraction of new tags identified approaches zero only when library size approaches 650,000 total tags (*43*). These analyses on SAGE data indicate that there are probably large numbers of low abundant genes that express at a level that can only be detected with much increased sequencing depth.

Table 5 Median expression levels of *ACTB* and *ACTG1* as reflected by libraries of different preparation process

Library type	Expression level (copy number/10,000 sequenced clones)	
	ACTB	*ACTG1*
Non-normalized library	24.65	19.43
Normalized library	1.97	1.06
Subtracted library	0	0
Uncharacterized library	19.85	0.16

Sequence-specific subtraction is an attracting method to increase the sequencing depth without changing the relative expression level of genes. The UniGene data analysis result (Figure 4) is consistent with analysis on SAGE libraries, which indicates that the most highly expressed 623 genes accounted for nearly one-half of the mRNA content (*43*). Sequence-specific subtraction can be more beneficial for particular tissues, such as subtraction of albumin transcript in liver libraries (Table 4).

Microarrays are limited by the prior knowledge of RNA species and are usually unable to distinguish alternative spliced forms or alternative promoter usage (*28*). It remains an open question how well array data can be quantified to reflect detailed expression level information, while fold changes of any range can be extracted in detail for cDNA sequencing data if sequencing depth is significantly increased. In summary, in-depth cDNA library sequencing stands for a very promising method in revealing valuable expression information to help our understanding of many biological and pathological processes.

Materials and Methods

cDNA library information was downloaded from the Cancer Genome Anatomy Project (http://cgap.nci.nih.gov/Info/CGAPDownload) and processed by an in-house Perl program. This includes information on tissue origin, library preparation, pathology and histology of the tissues or cell lines, total number of sequenced clones for the library, etc. Out of a total of more than 8,000 libraries, we have chosen 142 libraries generated by non-normalized preparation and with more than 5,000 sequenced clones for most of the analysis presented here (Table 1). Among these libraries, 43 were derived from normal tissues (normal_bulk), 23 were generated from tumor tissues (neoplasia_bulk), and 64 were originated from tumor cell lines (neoplasia_cell). There are also 12 libraries generated from cell lines of normal origin, and they were used for the expression correlation analysis but not the analysis on differential gene expression. In addition to the 142 libraries with non-normalized treatment, there are 61 normalized libraries, 10 subtracted libraries, and 152 libraries with uncharacterized treatment containing more than 5,000 sequenced clones. Differential expression (*BIRC5*) and expression correlation (between *ACTB* and *ACTG1*) from the non-normalized libraries were compared with those from normalized/subtracted libraries or uncharacterized libraries.

The EST sequences for the genes in this study were extracted either by searching dbEST_human by a standalone program using blastn program or by direct download of corresponding UniGene cluster sequences from NCBI for that particular gene. For the former, the most complete cDNA sequences for the given gene in question were extracted from NCBI and were used as templates in the subsequent similarity searches. Human EST sequences in dbEST were searched using the template sequences and standalone blastn program for sequences that generate high sequence similarities to the template (*44*). In the meantime, paralogous genes for the gene in question were also used to search dbEST and the entire blast search results were compared by an in-house Perl program to eliminate EST sequences that showed better alignment to the paralogous genes than to the gene in question. EST sequences that align with the corresponding templates in a stretch of 100 bp or below, or sequences that generate longer stretch of alignment but with low (less than 95%) sequence similarity were further analyzed by aligning the EST sequences with human genome sequences to find the chromosomal regions that best align with the ESTs. Information from relative sequence similarity to the corresponding template and to its paralogous genes or the chromosomal location that best align with the EST entry was used to distinguish a real positive sequence tag from a mismatch. The method was used to get the EST entries for *RAC1*, *Albumin*, *BIRC5/Survivin*, and *MALAT1/Pro1073*; the semi-manual method allows extraction of more sequences than those from UniGene for the corresponding genes. For other abundant genes usually with many paralogs, direct extraction of sequence entries from UniGene was used. Corresponding library information for the verified ESTs was retrieved using an in-house Perl program. Differential expression of the selected genes among normal tissues, neoplasia tissues, and neoplasia cell lines was analyzed using nonparametric t-test with Welch's correction for unequal variance. Expression level correlation between *ACTB* and *ACTG1* and between *KRT8/18* was done using linear regression.

The 57-nucleotide sequence in the extra exon of *RAC1b* cDNA was used as template in searching human EST sequences and matched sequences were further evaluated for their authenticity as *RAC1b* splicing variant. Then information on their tissue and library origin was extracted using an in-house Perl

program.

Acknowledgements

We thank John Hildebrandt, Yian Ann Chen, and David Kurtz of the Medical University of South Carolina for helpful discussions for this work.

Authors' contributions

WY conceived and conducted the analyses and prepared the manuscript. YLL participated in the analyses and revision of the manuscript. DY assisted with the data analyses. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Adams, M.D., *et al.* 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.
2. Boguski, M.S., *et al.* 1993. dbEST—database for "expressed sequence tags". *Nat. Genet.* 4: 332-333.
3. Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Res.* 7: 986-995.
4. Asmann, Y.W., *et al.* 2002. Identification of differentially expressed genes in normal and malignant prostate by electronic profiling of expressed sequence tags. *Cancer Res.* 62: 3308-3314.
5. Greller, L.D. and Tobin, F.L. 1999. Detecting selective expression of genes and proteins. *Genome Res.* 9: 282-296.
6. Claverie, J.M. 1999. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* 8: 1821-1832.
7. Patanjali, S.R., *et al.* 1991. Construction of a uniform-abundance (normalized) cDNA library. *Proc. Natl. Acad. Sci. USA* 88: 1943-1947.
8. Bonaldo, M.F., *et al.* 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* 6: 791-806.
9. Carninci, P., *et al.* 2000. Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* 10: 1617-1630.
10. Dias Neto, E., *et al.* 2000. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* 97: 3491-3496.
11. Perou, C.M., *et al.* 2000. Molecular portraits of human breast tumours. *Nature* 406: 747-52.
12. Jordan, P., *et al.* 1999. Cloning of a novel human Rac1b splice variant with increased expression in colorectal tumors. *Oncogene* 18: 6835-6839.
13. Singh, A., *et al.* 2004. Rac1b, a tumor associated, constitutively active Rac1 splice variant, promotes cellular transformation. *Oncogene* 23: 9369-9380.
14. Schuler, G.D., *et al.* 1996. A gene map of the human genome. *Science* 274: 540-546.
15. Sugarbaker, D.J., *et al.* 2008. Transcriptome sequencing of malignant pleural mesothelioma tumors. *Proc. Natl. Acad. Sci. USA* 105: 3521-3526.
16. Lossos, I.S., *et al.* 2004. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N. Engl. J. Med.* 350: 1828-1837.
17. Nutt, C.L., *et al.* 2003. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res.* 63: 1602-1607.
18. Moon, W.S. and Tarnawski, A.S. 2003. Nuclear translocation of survivin in hepatocellular carcinoma: a key to cancer cell growth? *Hum. Pathol.* 34: 1119-1126.
19. Ning, S., *et al.* 2004. siRNA-mediated down-regulation of survivin inhibits bladder cancer cell growth. *Int. J. Oncol.* 25: 1065-1071.
20. Boidot, R., *et al.* 2008. The expression of BIRC5 is correlated with loss of specific chromosomal regions in breast carcinomas. *Genes Chromosomes Cancer* 47: 299-308.
21. Hutchinson, J.N., *et al.* 2007. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 8: 39.
22. Ge, H., *et al.* 2001. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* 29: 482-486.
23. Wu, L.F., *et al.* 2002. Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.* 31: 255-265.
24. Thompson, H.G., *et al.* 2002. Identification and confirmation of a module of coexpressed genes. *Genome Res.* 12: 1517-1522.
25. Rhodes, D.R., *et al.* 2004. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl. Acad. Sci. USA* 101: 9309-9314.
26. Bertone, P., *et al.* 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* 306: 2242-2246.

27. Kapranov, P., *et al.* 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316: 1484-1488.
28. Strausberg, R.L. and Levy, S. 2007. Promoting transcriptome diversity. *Genome Res.* 17: 965-968.
29. Prasanth, K.V. and Spector, D.L. 2007. Eukaryotic regulatory RNAs: an answer to the "genome complexity" conundrum. *Genes Dev.* 21: 11-42.
30. Harbers, M. 2008. The current status of cDNA cloning. *Genomics* 91: 232-242.
31. Hafner, M., *et al.* 2008. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* 44: 3-12.
32. Eyras, E., *et al.* 2004. ESTGenes: alternative splicing from ESTs in Ensembl. *Genome Res.* 14: 976-987.
33. Florea, L., *et al.* 2005. Gene and alternative splicing annotation with AIR. *Genome Res.* 15: 54-66.
34. Yang, W. and Hildebrandt, J.D. 2006. Genomic analysis of G protein gamma subunits in human and mouse—the relationship between conserved gene structure and G protein betagamma dimer formation. *Cell. Signal.* 18: 194-201.
35. Hastie, N.D. and Bishop, J.O. 1976. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* 9: 761-774.
36. Jongeneel, C.V., *et al.* 2003. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *Proc. Natl. Acad. Sci. USA* 100: 4702-4705.
37. Brentani, H., *et al.* 2003. The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl. Acad. Sci. USA* 100: 13418-13423.
38. Bishop, J.O., *et al.* 1974. Three abundance classes in HeLa cell messenger RNA. *Nature* 250: 199-204.
39. Carter, M.G., *et al.* Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome Biol.* 6: R61.
40. Zhu, J., *et al.* 2008. Modeling transcriptome based on transcript-sampling data. *PLoS ONE* 3: e1659.
41. Galau, G.A., *et al.* 1977. Synthesis and turnover of polysomal mRNAs in sea urchin embryos. *Cell* 10: 415-432.
42. Stern, M.D., *et al.* 2003. Can transcriptome size be estimated from SAGE catalogs? *Bioinformatics* 19: 443-448.
43. Velculescu, V.E., *et al.* 1999. Analysis of human transcriptomes. *Nat. Genet.* 23: 387-388.
44. Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.