Optimisation of a minimum data set for the identification and quality control of EST expression libraries

Keywords: mRNA expression; Transcriptomics; Gene expression; EST expression; Quality control; Tissue typing; Tissue identification; Differential expression; Tissue specific markers; Differential gene expression in cancer.

Abstract: There are currently a few bioinformatics tools, such as dbEST, DDD and GEPIS to name a few, which have been widely used to retrieve and analyse EST expression data. The Cancer Genome Anatomy Project offers cDNA xProfiler and cDNA DGED tools which can also be used to use EST counts for comparing gene expression levels between cancer and normal tissues. The outcome of any such comparison depends on EST libraries' annotations and assumes that the actual expression data (EST counts) are correct. Neither CGAP nor other similar tools provide a quality control method for the selection and evaluation of the original EST expression libraries. Here we report the selection and optimisation of a minimal gene expression data set and describe a few examples of its applications. The described methods rely solely on the expression data itself and are independent on the libraries annotations. The reported approach allows tissue typing of expression libraries of different sizes containing between as little as 249 total EST counts and up to 13,929 total EST counts (the highest tested).

1 INTRODUCTION

CGAP and other similar tools and databases such as dbEST, EST (Digital Differential Display) and GEPIS (Gene Expression Profiling In Silico) compare expression levels between EST libraries from normal and cancerous tissues. However, they assume the reported EST counts to be correct without employing a quality control method for the underlying data which would enable the identity of each library to be verified independently of any external information. A quality control method is required because the methods used to generate EST libraries may also introduce biases into EST data (Liu and Graber, *2004*). For example, during any one cycle of a PCR reaction one DNA molecule can be amplified more than once (Song, *2003*). This disproportionate amplification will lead to abnormally high expression levels of those sequences appearing in the final results (Ray et al, *2004*) Errors can also be introduced from the fact that multiple polyadenylate repeats are found in a significant percentage of mRNA species contain multiple polyadenylation sites, potentially leading to multiple ESTs being produced from one transcript (Beaudoing et al, *2000*).

Analysis of expression data for quality control purpose has been attempted previously (Huminiecki et al, *2003*). Three databases were compared – Gene Expression Atlas (oligonucleotide microarray data), SAGEmap (SAGE libraries) and TissueInfo (EST libraries). Because these databases use different formats for sample annotation and use different statistical methods for data analysis, a method called Preferential Expression Measure (PEM) was devised to score differential expression of genes in libraries grouped into six different tissue categories in three databases. Inter-database correlations were measured and were found to vary between tissues. However, inter-library correlations have yet to be applied as a quality control method within one database (Huminiecki et al, *2003*).

In a more recent study, data for 8,570 genes across 46 human tissues from the Gene Expression Omnibus (an Affymetrix microarray data repository) were categorised according to tissue specificity and subcellular localisation of their protein product (Li et al, *2011*). The analysis revealed that widely expressed genes have higher expression levels than genes which are expressed in one or a few tissues (Li et al, *2011*).

While many quality control methods were previously suggested, they only focussed on the whole genome (Liang et al, *2006*) or covered aspects of the data such as GC content (Arhondakis et al, *2006*), with few investigations focusing on tissue-specificity (Russ and Futschik, *2010*). A common shortcoming of previous

reports is that tissue specificity of the genes was reported (Hu et al, *2000*) (Krief et al, *1999*) (Miner and Rajkovic, *2003*) (Pao et al, *2006*) (Vaes et al, *2002*) but no attempts were made to use such data for quality control or evaluation. Moreover, even unique "tissue specific genes" might be of little use if they are expressed at low levels and would therefore be absent in many smaller libraries. Furthermore, many existing tools and databases, including CGAP, are simply information retrieval tools, lacking methods for verification of the EST counts and sample origins. The EST counts are assumed to be correct and the libraries to be correctly annotated (Elfilali et al, *2006*) (Strausberg et al, *2002*) (Zhang et al, *2004*). The existing algorithms used to analyse expression data place the emphasis on identification of the degree of over/under-expression for tissue/disease-specific genes by comparing EST counts between two library groups without evaluating the quality of the expression data or the origins of the experimental material used, these are simply assumed to be correct and no numerical methods for their verification are made available (Elfilali et al, *2006*) (Strausberg et al, *2002*) (Zhang et al, *2004*). It is not surprising that many such tissue distribution resources are quickly superseded by more recent developments or are being taken off-line (Brown et al, *2004*) (Kawamoto et al, *1996*) (Okubo et al, *1992*) (Skrabanek and Campagne, *2001*).

2 A NEW APPROACH TO THE QUALITY CONTROL OF EXPRESSION DATA

Tissue phenotype depends on the pattern of gene expression in the tissue as well as the influence of environmental factors. Therefore the pattern of gene expression in the identical tissues if probed under similar conditions is likely to be similar or nearly identical. Thus comparing global gene expression data in the form of EST expression levels between similarly prepared EST libraries from the identical tissues is close to "+1" (data not shown).

We hypothesised that a smaller subset of genes can be generated and used for the same purpose of identifying or checking EST libraries prior to further analysis. The main challenge was therefore finding a small enough subset of genes sufficient for the task. One other major challenge was to ensure that even small EST expression libraries often having less than ~ 1,000, or even less than ~ 100 EST counts are still identifiable. For these and other reasons we could not rely on the so called "tissue specific" markers, expression of which, whilst being specific in most cases, is often rather low, resulting in their absence from many medium size and smaller libraries.

We have hypothesised that the identity of non-normalised and non-subtracted expression libraries from normal non-cancerous (healthy) tissues may be inferred from a pattern of expression of such a subset of genes. Such an expression patterns is a more biologically relevant phenomenon rather than hypothetical "yes/no" pattern expected of the so called "tissue specific" genes. We have further hypothesised that removal of potentially differentially expressed genes and the ones which express constitutively should further improve the accuracy of tissue typing. To this end we have used a set of EST expression libraries available from CGAP database (cgap.nci.nih.gov/Info/CGAPDownload.) to find such a reduced and optimised expression dataset.

2.1 Finding Tissue specific transcripts and generating EST expression matrices

Initial lists of 2,295 tissue-specific transcripts and 37,575 transcripts found in non-normalised non-cancerous libraries from many tissues were obtained from the CGAP database. We selected UniGene IDs which appeared to be the most abundant in their target tissues relative to all the other tissues and which were also abundant in absolute terms in the relevant target tissues. The high relative abundance (high odds ratio) defines the tissue specificity. The high absolute abundance (above 0.1%) was chosen to ensure that such transcripts would still be found even in smaller libraries with small number of total EST counts. Up to thirty individual transcripts were eventually selected using these criteria, from each of the individual tissue types. The analysis of 25 human tissues yielded just over 1,000 transcripts of which about half were expressed in more than one tissue type, making our approach different from the ones reliant onto the "tissue specific" genes which are not supposed to be expressed in more than one specific tissue. We then attempted to optimise our selection.

For the majority of the tissues, the original selection was made based on the very small number of libraries available in CGAP for those tissues (typically 2-4 libraries, with brain and placenta being exceptions where more than 10 libraries were available). Because of that and also because of the stringent selection requirements, it was reasonable to assume that some suitable transcripts could have been omitted because of the very limited choice of libraries available for the analysis and not because of them being unsuitable tissue markers. Therefore, a

search was undertaken for additional candidate transcripts by looking solely into individual EST counts across 155 non-normalised libraries from all non-cancerous tissue types. Genes having expression patterns similar to the original list of ~ 1,000 ESTs across all of the libraries were selected, which expanded the list of potential marker genes to 1,437 transcripts.

Because of the relaxed criteria used for selecting the potential tissue markers, and in order to find the best makers and also to reduce the list to a more manageable size we attempted to optimise the selection using new selection criteria independent of the ones used in the original rounds of selection. For this first round the EST counts for the 1,437 transcripts were summed together from all the libraries in each tissue to make a super-library for that tissue. All possible Pearson correlations were calculated between all of such super-libraries (equation 1).

$$Correl(X,Y) = \frac{\Sigma(x-m)(y-n)}{\Sigma(x-m)^2\,\Sigma(y-n)^2} \qquad (1)$$

Where x and y are the EST count for the transcript concerned in super-libraries X and Y respectively, where m and n are the mean EST counts across all 1,437 transcripts in super-libraries X and Y, respectively, and where Correl(X,Y) is the calculated Pearson Correlation Coefficient between the two super-libraries.

Higher correlation value here means higher inter-tissue correlation and is undesirable; ideally all such inter-tissue correlations should be equal to "0". Hence we calculated sum of squares of deviations of the calculated correlation value from "1" (equation 2).

$$S = \sum([Correl - 1)]^2 \qquad (2)$$

Where Correl is the calculated Pearson Correlation coefficient between two super-libraries and where S is the calculated sum of squares value for the correlations between all possible pairs of super-libraries.

Individual genes were then removed and the correlation values and the equation (2) total were recalculated. Gene, removal of which resulted in the lowest overall inter-tissue correlations' (as calculated per equation (2)) was permanently removed and the iteration steps were repeated again. The decrease in inter-tissue correlations slowed shortly before the 1,000th gene was removed (Figure 1). The remaining ~ 500 genes included the set of high-quality tissue-specific markers and these were retained. A similar optimisation was then repeated for the remaining ~ 500 genes but this time we aimed to improve intra-tissue correlations between the individual libraries from within the same tissues and hence used the original individual EST libraries, rather than the super libraries (data not shown).Transcripts were removed one by one and the correlations recalculated. The transcript whose removal resulted in the improvement of intra-tissue correlation was permanently removed. The finally optimised set of tissue-specific markers contained 244 transcripts for which EST expression matrix (244 transcripts x 26 tissues) was created.

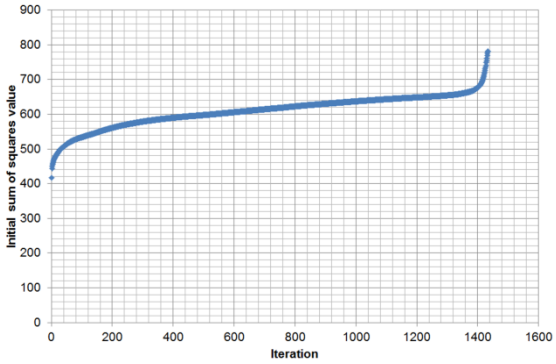

Figure 1: Inter-tissue correlation during optimization of marker list for genes with improved tissue specificity. The increase in the sum of squares value (which corresponds to a decrease in the inter-tissue correlation) (y-axis) is plotted against the gene removal iteration (x-axis), after each of which the gene was permanently removed whose temporary removal had produced the greatest improvement in the tissue-specificity of the gene list.

2.2 Confirming identity of known libraries of varying sizes using inter-tissue correlations using EST expression matrix

Correlation values between tissue expression profiles of the 244 transcripts from the EST expression matrix and the relevant EST counts from 113 largest libraries representing 26 main human tissues were calculated. The correlation data for a group of tissues which contained libraries for which virtually no inter-tissue correlation was found, and where all the libraries shown good positive correlation (values ranging approximately within +0.2 to +1) with the relevant source tissues but not with any of the other tissues. Figure 2 summarises the results for five such representative tissues where

correlation levels clearly confirm the identity of each of the individual EST libraries.

Further, in order to systematically investigate the robustness of this approach, we used modelled EST data to simulate small EST expression datasets. These were generated from the reported EST expression data taken from CGAP database, by proportionally reducing the reported EST counts and rounding any fractional values to the nearest whole EST count each time until each library ceased to present any ESTs mapping onto the 244 marker transcripts or ceased to be identified as a positive tissue match for the tissue from which it was created in the first place. Using this approach we gradually scaled down the real EST expression data and compared all of the generated model libraries with the original libraries including from all the other tissues by calculating the correlation values for the 244 UniGene IDs from our optimised matrix set (Figure 3 – Figure 7). Virtually every library continues to correlate well with the tissue of origin until the very last UniGene ID and the last EST mapping onto one of the transcripts in the matrix is removed. Furthermore, we realised that the majority of the scaled down libraries remain identifiable until the total library EST counts falls below the range of 10 to 50 which is equal to some of the smallest libraries currently in the CGAP database.

Our results are summarised in Tables 1-5, which report tissue matching results for each of the original EST libraries used and the relevant scaled down model data sets. The initial and the final (reduced) number of total ESTs are shown and the relevant correlation values are indicated for each pair. Remarkably, the final mapped EST counts across all transcripts in each library which still yield positive intra-tissue correlation for the transcripts in the matrix are below 100 ESTs for all but 3 libraries tested and are below 10 total ESTs for 15 out of 33 libraries tested. The quality of tissue typing does not change dramatically and for lung the correlations it actually improved as the total EST counts were reduced. These findings show that the matrix can be used to confirm the tissue identity of very small libraries, making it a very robust method for the quality control of expression libraries and tissue typing.

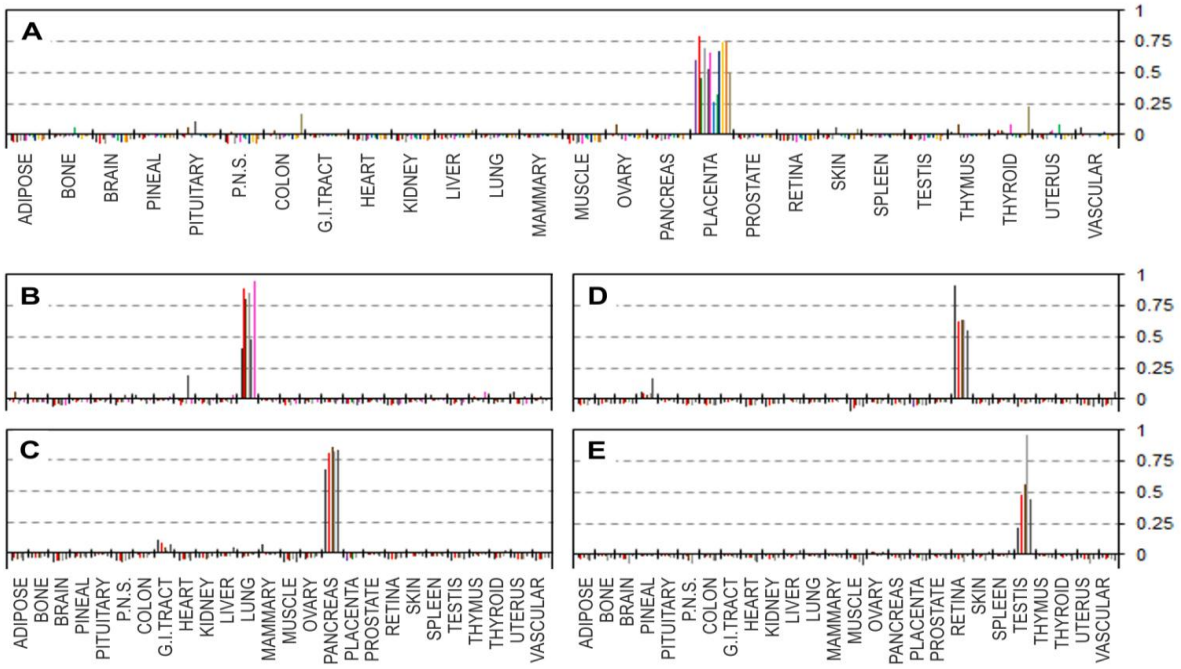

Figure 2: Correlation of the EST matrix with individual libraries from matching tissues showing no inter-tissue correlation. Pearson product-moment correlation coefficients (vertical axes) calculated for each of the individual EST libraries and the EST expression matrix. A: Placental libraries. B: Lung libraries. C: Pancreatic libraries. D: Retinal libraries. E: Testis libraries. Used with permission (Milnthorpe and Soloviev, *2012*).

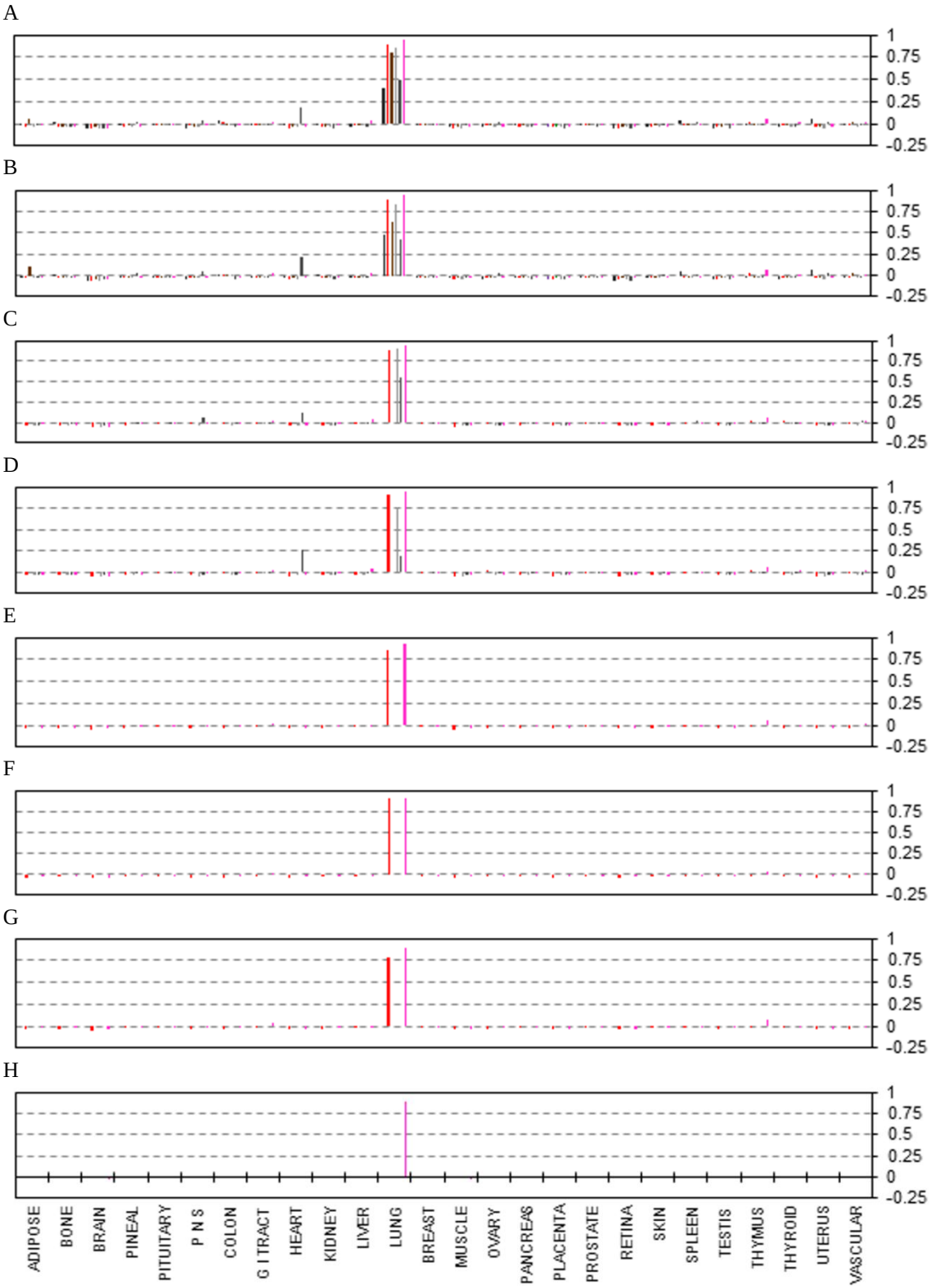

Figure 3: Correlation of the EST matrix with individual libraries of reduced size from lung tissue. Pearson product-moment coefficients (vertical axes) calculated for each individual EST library and the EST expression matrix. A: original libraries. B: Reduced to 50% of original size. C: 20% of original size. D: reduced to 10% of original counts. E: lowered to 5% of original size. F: lowered to 2% of original size. G: reduced to 1% of original size. H: lowered to 0.5% of original size. The original sizes for each of the libraries used are listed in Table 1.

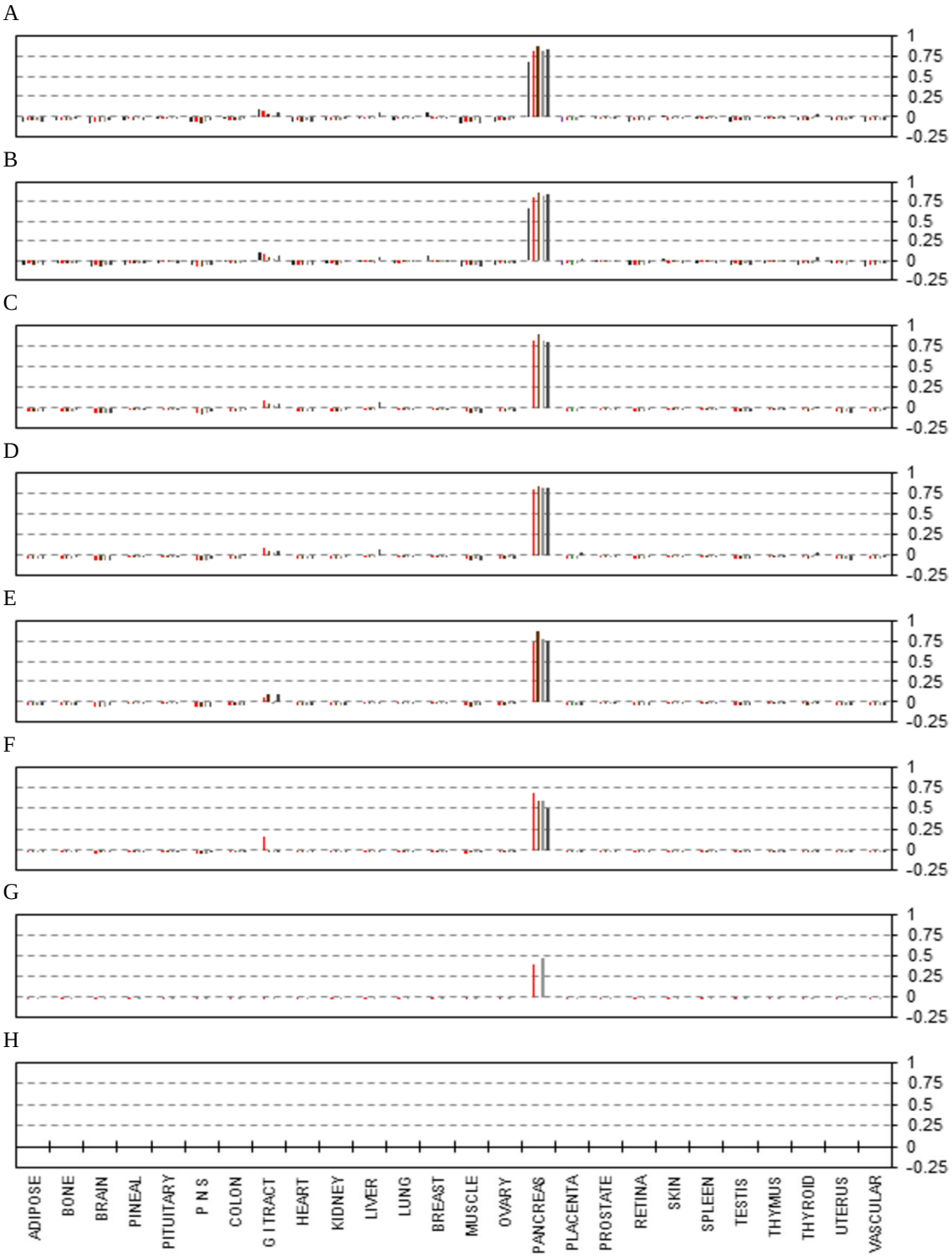

Figure 4: Correlation of the EST matrix with individual libraries of gradually reduced size from pancreas. Pearson product-moment coefficients (vertical axes) calculated for each individual EST library and the EST expression matrix. A: original libraries. B: Reduced to 50% of original size. C: 20% of original size. D: reduced to 10% of original counts. E: lowered to 5% of original size. F: lowered to 2% of original size. G: reduced to 1% of original size. H: reduced to 0.5% of original size. The original sizes for each of the libraries used are listed in Table 2.

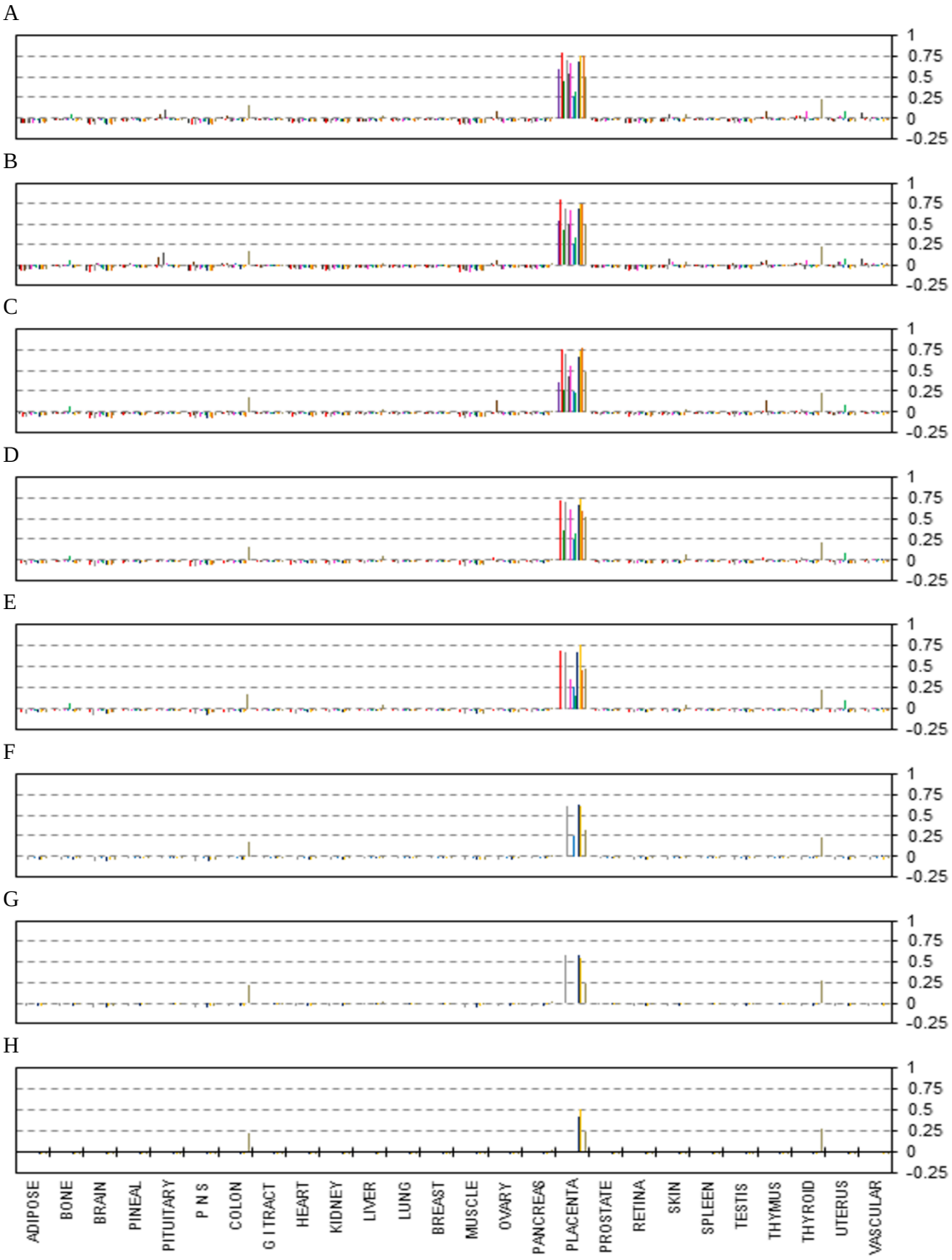

Figure 5: Correlation of the EST matrix with individual libraries of gradually reduced size from placenta. Pearson product-moment coefficients (vertical axes) calculated for each individual EST library and the EST expression matrix. A: original libraries. B: Reduced to 50% of original size. C: 20% of original size. D: reduced to 10% of original counts. E: lowered to 5% of original size. F: lowered to 2% of original size. G: reduced to 1% of original size. H: lowered to 0.5% of original size. The original sizes for each of the libraries used are listed in Table 4.

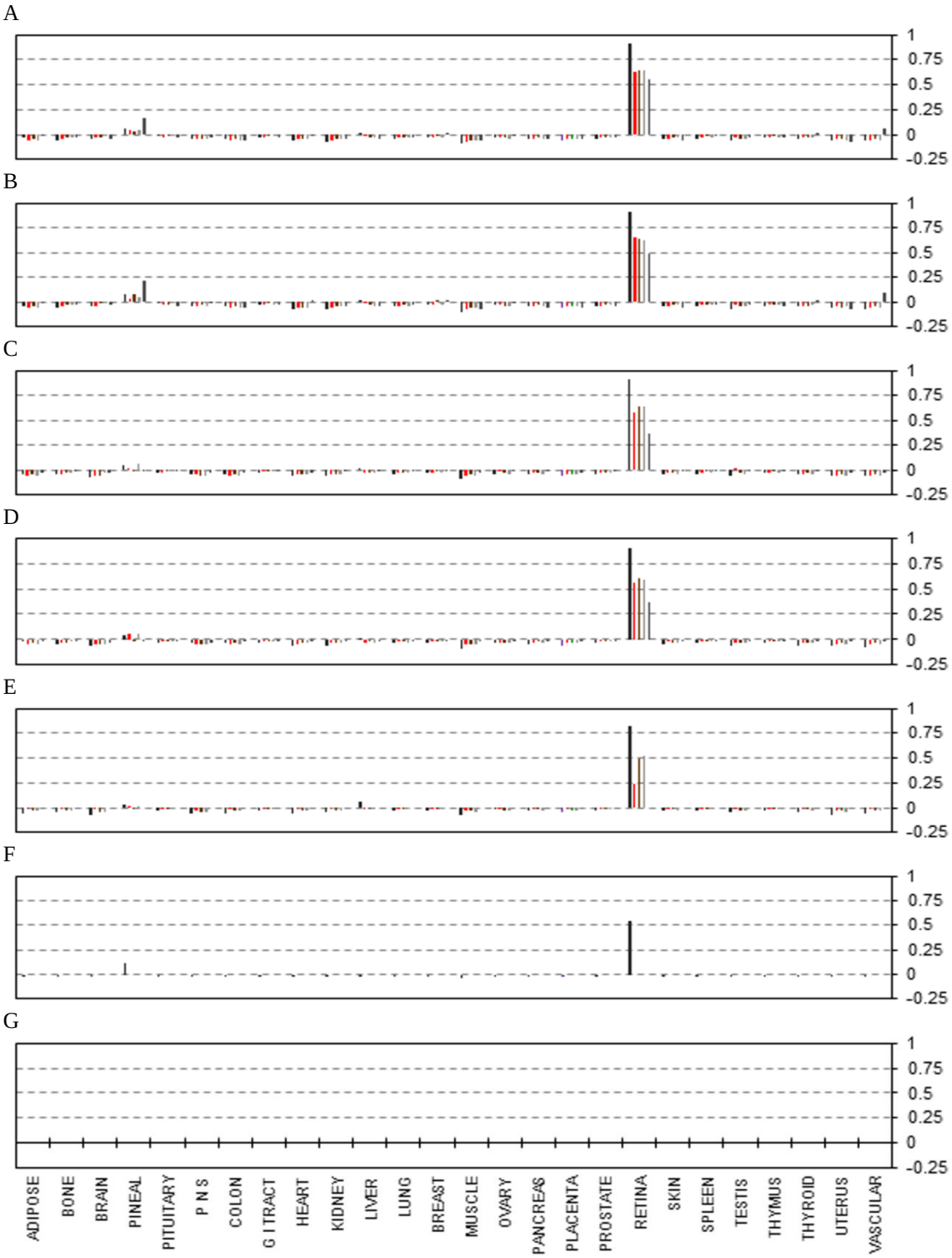

Figure 6: Correlation of the EST matrix with individual libraries of gradually reduced size from retina. Pearson product-moment coefficients (vertical axes) calculated for each individual EST library and the EST expression matrix. A: original libraries. B: Reduced to 50% of original size. C: 20% of original size. D: reduced to 10% of original counts. E: lowered to 5% of original size. F: lowered to 2% of original size. G: reduced to 1% of original size. The original sizes for each of the libraries used are listed in Table 3.

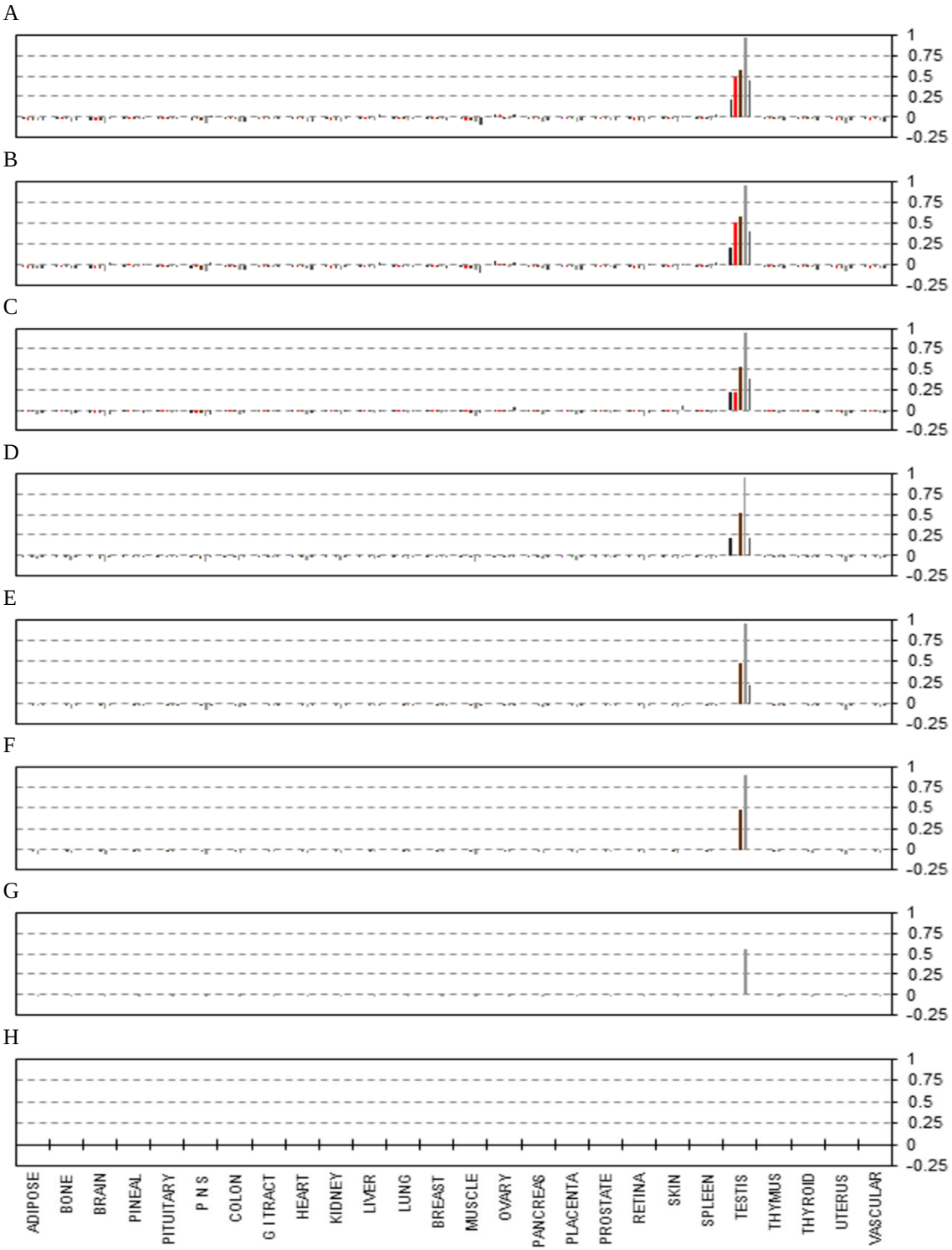

Figure 7: Correlation of the EST matrix with individual libraries of gradually reduced size from testis. Pearson product-moment coefficients (vertical axes) calculated for each individual EST library and the EST expression matrix. A: original libraries. B: Reduced to 50% of original size. C: 20% of original size. D: reduced to 10% of original counts. E: lowered to 5% of original size. F: lowered to 2% of original size. G: reduced to 1% of original size. H: lowered to 0.5% of original size. The original sizes for each of the libraries used are listed in Table 5.

Table 1:Library sizes and correlations for EST libraries from lung.

Library Name	Original library, the number of mapped[1] ESTs	Positive correlation with the tissue of origin using EST expression matrices[2]	Modelled scaled down library, the number of remaining ESTs[3]	Positive correlation with the tissue of origin for the modelled scaled down library using the same matrices[4]
Human Lung	536	0.40	461	0.48
Stratagene lung (#937210)	8,511	0.89	10	0.78
Human adult lung 3' directed MboICdna	257	0.80	255	0.62
Lung	401	0.85	6	0.76
Fetal lung II	1,289	0.48	83	0.55
NIH_MGC_77	12,494	0.95	11	0.88

Table 2:Library sizes and correlations for EST libraries from pancreas.

Library Name	Original library, the number of mapped[1] ESTs	Positive correlation with the tissue of origin using EST expression matrices[2]	Modelled scaled down library, the number of remaining ESTs[3]	Positive correlation with the tissue of origin for the modelled scaled down library using the same matrices[4]
Human Pancreas	249	0.67	231	0.67
Barstead pancreas HPLRB1	709	0.81	4	0.39
NCI_CGAP_Pan3	356	0.86	4	0.60
NIH_MGC_78	557	0.82	2	0.46
Pancreatic Islet	1,789	0.83	4	0.50

Table 3: Library sizes and correlations for EST libraries from retina.

Library Name	Original library, the number of mapped[1] ESTs	Positive correlation with the tissue of origin using EST expression matrices[2]	Modelled scaled down library, the number of remaining ESTs[3]	Positive correlation with the tissue of origin for the modelled scaled down library using the same matrices[4]
Soares retina N2b4HR	9,160	0.91	13	0.54
Soares retina N2b5HR	1,722	0.62	7	0.24
Human retina cDNATsp509I-cleavedsublibrary	706	0.64	4	0.49
Human retina cDNA randomly primed sublibrary	2,169	0.64	18	0.53
Retina II	1,171	0.56	18	0.37

[1] Mapped ESTs are the ESTs in each library which map onto transcripts.

[2] Using the matrices and as described in "Confirming identity of known libraries of varying sizes using inter-tissue correlations using EST expression matrix"

[3] Each individual library was scaled down to model a smaller EST library and any fractional EST counts were rounded to the nearest whole number. The reduced modelled EST counts below "0.5" were rounded down to "0".

[4] Gradual disappearance of low abundant ESTs resulted in the progressive change lowering in of the positive correlation with the tissue of origin and in many cases the eventual loss of that correlation. Each library was scaled down until such positive correlation was lost.

Table 4: Library sizes and correlations for EST libraries from placenta.

Library Name	Original library, the number of mapped[1] ESTs	Positive correlation with the tissue of origin using EST expression matrices[2]	Modelled scaled down library, the number of remaining ESTs[3]	Positive correlation with the tissue of origin for the modelled scaled down library using the same matrices[4]
Human Placenta	276	0.60	7	0.35
Stratagene placenta (#937225)	2,784	0.79	31	0.69
Clontech human placenta polyA+ mRNA (#6518)	705	0.45	34	0.35
Soares_placenta_8to9 weeks_2NbHP8to9W	13,929	0.70	7	0.58
Human placenta polyA+ (TFujiwara)	405	0.53	13	0.42
Human placenta cDNA (TFujiwara)	1,367	0.66	24	0.35
Placenta II	662	0.26	2	0.26
Placenta I	1,168	0.33	11	0.15
NIH_MGC_79	9,271	0.67	10	0.42
NCI_CGAP_Pl1	1,856	0.74	2	0.50
NCI_CGAP_Pl4	1,261	0.74	21	0.46
Homo sapiens PLACENTA	11,864	0.50	69	0.33

Table 5:Library sizes and correlations for EST libraries from testis.

Library Name	Original library, the number of mapped[1] ESTs	Positive correlation with the tissue of origin using EST expression matrices[2]	Modelled scaled down library, the number of remaining ESTs[3]	Positive correlation with the tissue of origin for the modelled scaled down library using the same matrices[4]
TEST1, Human adult Testis tissue	326	0.22	7	0.22
Human Testis	293	0.48	4	0.22
Testis I	1,525	0.56	1	0.47
NIH_MGC_82	7,602	0.96	4	0.55
NIH_MGC_180	4,984	0.44	17	0.22

[1] Mapped ESTs are the ESTs in each library which map onto transcripts.

[2] Using the matrices and as described in "Confirming identity of known libraries of varying sizes using inter-tissue correlations using EST expression matrix"

[3] Each individual library was scaled down to model a smaller EST library. and any fractional EST counts were rounded to the nearest whole number. The reduced modelled EST counts below "0.5" were rounded down to "0".

[4] Gradual disappearance of low abundant ESTs resulted in the progressive change lowering in of the positive correlation with the tissue of origin and in many cases the eventual loss of that correlation. Each library was scaled down until such positive correlation was lost.

3 DISCUSSION

In our quest to create a quality control method based purely on the expression data itself, we initially selected 1,437 transcripts, and then performed two rounds of optimisation to reduce inter-tissue correlations and improve intra-tissue correlations to produce a final list of transcripts. As a result, the 244 chosen transcripts are highly abundant in the tissue of interest when compared to all other tissues (high odds ratio), but are not necessarily the "tissue specific" markers in the traditional understanding of this term, as many are expressed in more than one distinct tissue.

An EST expression matrix of these markers in 26 tissues was created and used as the control against which other libraries were compared. The findings presented in Figure 3 – Figure 7 and Table 1 – Table 5 show that the EST expression matrix is capable identifying the tissue of origin for expression libraries of different sizes containing between as little as ~ 1 EST counts (modelled scaled down library Testis I) and up to 13,929 EST counts (Soares_placenta_8to9weeks_2NbHP8to9W).These findings show that tissue-specific gene expression can be used as a quality control method.

Earlier investigations focussed on the whole genome (Liang et al, *2006*), or studied aspects such as GC content (Arhondakis et al, *2006*) or did not use tissue-specific gene expression data for quality control or evaluation purposes (Hu et al, *2000*) (Krief et al, *1999*) (Miner and Rajkovic, *2003*) (Pao et al, *2006*) (Vaes et al, *2002*), (Russ and Futschik, *2010*). Furthermore, tissue-specific genes have been identified in this investigation which are also highly expressed in their target tissues, unlike the tissue-specific genes reported previously in (Li et al, *2011*). This study is also an improvement on many existing search tools and secondary databases, including those hosted by CGAP, which are merely information repositories and retrieval algorithms with few numerical procedures for verifying the reported EST counts and the origins of the samples studied, both of which are assumed to be accurately reported (Elfilali et al, *2006*) (Strausberg et al, *2002*) (Zhang et al, *2004*).

The reported EST expression matrix can be used to confirm tissue identities of EST expression datasets for all main human tissue types, to provide insight into the origin of uncharacterised libraries and to identify various experimental artefacts. The next step is to further improve this method by incorporating other gene expression data, such as SAGE data (Leyritz et al, *2008*), DNA microarray data (Baron et al, *2011*) and northern blots (Schlamp et al, *2008*). It is envisaged that with the increasing amounts of expression data, the optimised expression data set could be generated and the tissue range might be further expanded. It is also envisaged that increasing amounts of available data may allow accurate analysis and tissue typing of the related and dependent tissues.

4 CONCLUSION

An EST expression matrix has been optimised and tested here on EST libraries of a range of sizes. We showed that the tissue type annotations of EST libraries could be verified by using a small expression matrix. Furthermore, the robustness of the new quality control method was confirmed by using it to correctly identity libraries which contain only a handful of ESTs.

REFERENCES

Arhondakis, S., Clay, O.,Bernardi, G. 2006. Compositional properties of human cDNA libraries: practical implications *FEBS Letters* **580** (24) pp. 5,772 – 5,778

Baron, D., Dubois, E., Bihouée, A., Teusan, R., Steenman, M., Jourdon, P., Magot, A., Péréon, Y., Veitia, R., Savagner, F., Ramstein, G., Houlgatte, R. 2011. Meta-analysis of muscle transcriptome data using the MADMuscle database reveals biologically relevant gene patterns *BMC Genomics* **12** (113)

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M.,Gautheret, D. 2000. Patterns of variant polyadenylation signal usage in human genes *Genome Research* **10** pp. 1,001 – 1,010

Brown, A.C., Kai, K., May, M.E., Brown, D.C.,Roopenian, D.C. 2004.ExQuest, a novel method for displaying quantitative gene expression from ESTs *Genomics* **83** (3) pp. 528 – 539

Elfilali, A., Lair, S., Verbeke, C., La Rosa, P., Radvanyi, F.,Barillot, E. 2006. ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis *Nucleic Acids Research* **34** (Database issue) D613 - D616

Hu, R.M., Han, Z.G., Song, H.D., Peng, Y.D., Huang, Q.H., Ren, S.X., Gu, Y.J., Huang, C.H., Li, Y.B., C. L. Jiang, Fu G., Zhang Q.H., Gu, B.W., Dai, M., Mao, Y.F., Gao, G.F., Rong, R., Ye, M., Zhou, J., Xu, S.H., Gu, J., Shi, J.X., Jin, W.R., Zhang, C.K., Wu, T.M., Huang, G.Y., Chen, Z., Chen, M.D., Chen, J.L. 2000. Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning *Proceedings of the National Academy*

of Sciences of the United States of America **97** (17) pp. 9,543 – 9,548

Huminiecki, L., Lloyd, A.T., Wolfe, K.H. 2003. Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases *BMC Genomics* **4** (31)

Kawamoto, S., Matsumoto, Y., Mizuno, K., Okubo, K., Matsubara, K. 1996. Expression profiles of active genes in human and mouse livers *Gene* **174** (1) 151 – 158

Krief, S., Faivre, J.F., Robert, P., Le Douarin, B., Brument-Larignon, N., Lefrère, I., Bouzyk, M.M., Anderson, K.M., Greller, L.D., Tobin, F.L., Souchet, M., Bril, A. 1999. Identification and characterization of cvHsp. A novel human small stress protein selectively expressed in cardiovascular and insulin-sensitive tissues *The Journal of Biological Chemistry* **274** (51) 36,592 – 36,600.

Leyritz. J., Schicklin, S., Blachon, S., Keime, C., Robardet, C., Boulicaut, J.F., Besson, J., Pensa, R.G., O. Gandrillon, O. 2008. SQUAT: A web tool to mine human, murine and avian SAGE data *BMC Bioinformatics* **9** (378)

Li, Q., Liu, X., He, Q., Hu, L., Ling, Y., Wu, Y., Yang, X., Yu, L. 2011. Systematic analysis of gene expression level with tissue-specificity, function and protein subcellular localization in human transcriptome *Molecular Biology Reports* **38** (4) pp. 2,597 – 2,602

Liang, S., Li, Y., Be, X., Howes, S., Liu, W. 2006 Detecting and profiling tissue-selective genes *Physiological Genomics* **26** (2) pp. 158 – 162

Liu, D., Graber, J.H. 2006. Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation *BMC Bioinformatics* **7** (77)

Milnthorpe, A.T., Soloviev, M. 2012. The use of EST expression matrixes for the quality control of gene expression data *PLoS One* **7** (3) e32,966

Miner, D., Rajkovic, A. 2003. Identification of expressed sequence tags preferentially expressed in human placentas by in silico subtraction *Prenatal Diagnosis* **23** (5) pp. 410 – 419.

Okubo, K., Hori, N., Matoba, R., Niiyama, Fukushima, T.A., Kojima, Y., Matsubara, K. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression *Nature Genetics* **2** (3) pp. 173 – 179

Pao, S.Y., Lin, W.L., and Hwang, M.J. 2006. In silico identification and comparative analysis of differentially expressed genes in human and mouse tissues *BMC Genomics* **7** (86)

Ray, A., Macwana, S., Ayoubi, P., Hall, L.T., Prade, R., Mort, A.J. 2004 Negative subtraction hybridization: an efficient method to isolate large numbers of condition-specific cDNAs *BMC Genomics* **5** (22)

Russ, J.,Futschik, M.E.2010. Comparison and consolidation of microarray data sets of human tissue expression *BMC Genomics* **11** (305)

Schlamp, K., Weinmann, A., Krupp, M., Maass, T., Galle, P.,Teufel, A. 2008.BlotBase: a northern blot database *Gene* **427** (1 – 2) pp. 47 – 50

Skrabanek, L. and Campagne, F. 2001. TissueInfo: high-throughput identification of tissue expression profiles and specificity *Nucleic Acids Research* **29** (21) E102

Song, J. 2003. What a Wise SAGE Once Said about Gene Expression… *BioTeach Online Journal* **1** pp. 99 – 104

Strausberg, R.L., Camargo, A.A., Riggins, G.J., Schaefer, C. F., de Souza, S. J., Grouse, L.H., Lal, A., Buetow, K.H., Boon, K., Greenhut, S.F., Simpson, A.J. 2002. An international database and integrated analysis tools for the study of cancer gene expression *The Pharmacogenomics Journal* **2** (3) pp. 156 – 164

Vaes, B.L., Dechering, K.J., Feijen, A., Hendriks, J.M., Lefèvre, C., Mummery, C.L., Olijve, W., van Zoelen, E.J.,Steegenga, W.T. 2002. Comprehensive microarray analysis of bone morphogenetic protein 2-induced osteo-blast differentiation resulting in the identification of novel markers for bone development *Journal of Bone and Mineral Research* **17** (12) 2,106 – 2118

Zhang, Y., Eberhard. D.A., Frantz, G.D., Dowd, P., Wu, T.D., Zhou, Y., Watanabe, C., Luoh, S.M., P. Polakis, P., Hillan, K.J., Wood, W.I., Zhang, Z. 2004. GEPIS—quantitative gene expression profiling in normal and cancer tissues *Bioinformatics* **20** (15) 2,390 – 2,398