Correlation between EST library size and the tissue specificity of EST expression data

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. Knowledge of tissue-specific gene expression is required for the data provided using these algorithms to be of most use. Previously we reported an EST expression matrix to elucidate the tissue type of an uncharacterised library from its expression data. 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 1 total EST count and up to 461total EST counts (the highest tested).**

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

I. INTRODUCTION

Tools such as CGAP, GEPIS and DDD may be used to compare expression levels between EST libraries from normal and cancerous tissues. However, these tools 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.

Gene expression is highly tissue-specific and the subset of genes expressed in each tissue determines the tissue function. Differential gene expression in cancer results in a primary tumour ceasing to resemble expression in the parent tissue; likewise normal tissue function is similarly distorted. Metastasis tumours result in even more complex patterns of gene expression. Furthermore, oncogenesis can occur differently in each tissue thus further complicating the interpretation if gene expression pattern. The knowledge of tissue specificity and the overall quality control of expression libraries generation and analysis are required because the methods used to generate EST libraries such as RT PCR and random selection of cDNAs for sequencing can introduce biases into EST data [1]. Disproportionate amplification during PCR [2] will lead to abnormally high expression levels of those sequences appearing in the final results [3]. Errors can also be introduced during reverse transcription because of 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 [4]. Furthermore, the analysis algorithms themselves can contain errors [5]

Inter-library correlations for tissue-specificity of expression have been attempted previously with SAGE data [6]. 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 (brain, kidney, ovary, pancreas, prostate and vascular endothelium) in three databases. Inter-database correlations were measured and were found to be high for brain, prostate and vascular endothelium, but not for kidney, ovary and pancreas [6].

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 [7]. The authors reported that widely expressed genes have higher expression levels than genes which are expressed in one or a few tissues [7].

Previous investigations only focussed on the whole genome [8] and covered aspects of the data such as GC content [9], with few investigations focusing on the tissue-specificity issues [10]. A common shortcoming of many previous reports is that tissue specificity of the genes was reported [11 – 15] but no attempts were made to actually use such data for quality control or evaluation of the expression data. 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 secondary databases, including the CGAP, are simply sophisticated information retrieval tools, lacking numerical methods for verification of the EST counts and sample origins. Existing algorithms used to analyse EST 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 pools without fully evaluating the quality of the expression data or the origins of the experimental material used, those are simply

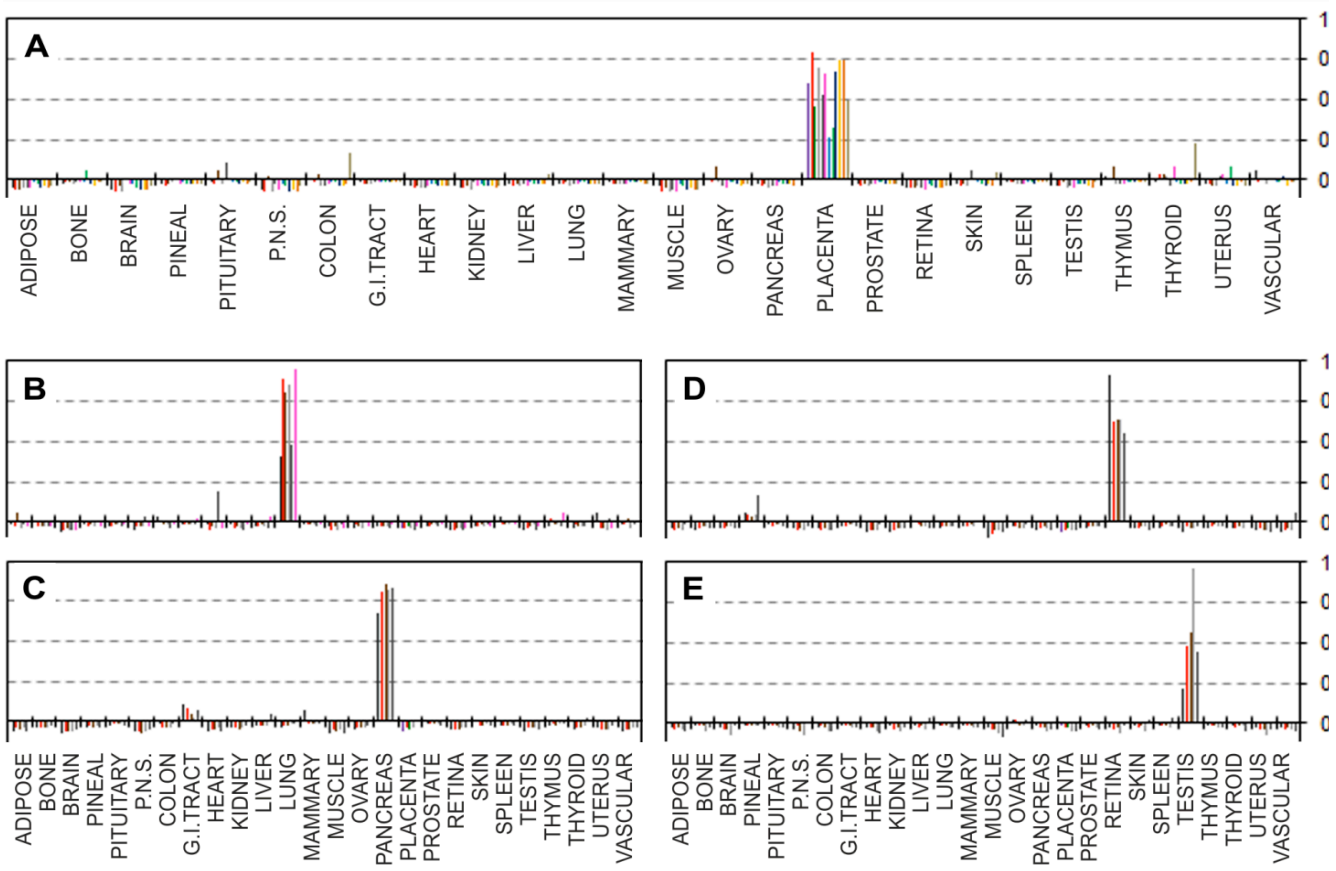

Figure 1. Correlation of the EST matrix with individual libraries from matching tissues showing no inter-tissue correlation. Pearson product-moment correlation coefficients (vertical axes) calculated for each of the individual EST libraries and the EST expression matrix: (a) placental libraries. (b) lung libraries. (c) pancreatic libraries. (d) retinal libraries. (e) testis libraries. Reproduced with permission from [23].

assumed to be correct and no numerical methods for their verification are made available [16 – 18]. It is not surprising that many such tissue distribution resources are quickly superseded by more recent developments or are being taken offline [19 – 22].

We have earlier calculated 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 [23]. 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, is summarized in Figure 1. Correlation levels clearly confirm the identity of each of the individual EST libraries.

This approach to the tissue-specificity problem is different from the previously reported attempts in that the origins of the expression data were looked into and the tissue specificity of the original preparations and the data quality were both assessed. It was possible to generate a small optimised subset of 244 different transcripts which showed high levels of intra-tissue correlation between different EST libraries while presenting low levels of inter-tissue correlation, suggesting high tissue specificity. 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, to identify normalised or subtracted libraries or various other experimental artefacts. In a few cases it was possible to identify the location of the tumour from which a cancer sample was taken, an extension not previously considered and not previously reported [23]. For the EST expression matrix see http://dx.doi.org/10.1371/journal.pone.0032966.

II. Results

We recently showed that tissue specificity can be used for the purposes of expression data quality control using small EST expression matrices [23]. Because Pearson correlation coefficients used in our previous analysis allow any linear transformation of the data, our approach should in principle work for any EST library size, however large or small. However, the number of tissue specific transcripts would gradually decrease and eventually none of the 244 genes

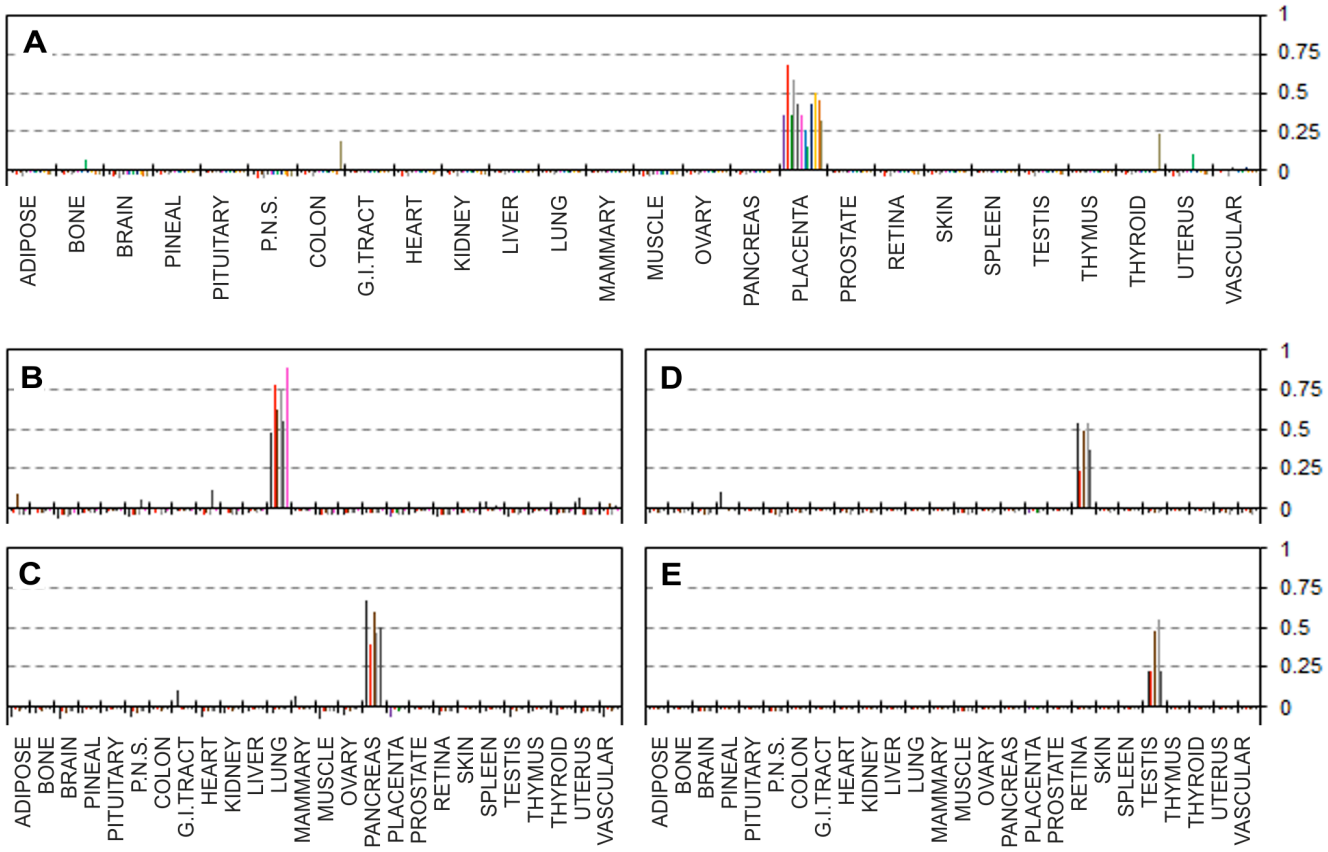

Figure 2. Correlation of the EST matrix with individual modelled (small) libraries from matching tissues showing no inter-tissue correlation. Pearson product-moment correlation coefficients (vertical axes) calculated for each of the individual scaled down libraries and the EST expression matrix: (a) placenta, total number of EST counts in the libraries were 7, 31, 34, 7, 13, 24, 2, 11, 10, 2, 21 and 69 (left to right). (b) lung, total number of EST counts in the libraries were 461, 10, 255, 6, 83 and 11 respectively. (c) pancreas, total number of EST counts in the libraries were 231, 4, 4, 2 and 4. (d) retina, total number of EST counts in the libraries were 13, 7, 4, 18 and 18. (e) testis libraries, total number of EST counts in the libraries were 7, 4, 1, 4 and 17.

included in the matrix could match a small EST library. We have therefore attempted to find the minimum library size for which this approach would still work. In our previous studies the expression matrix was used to confirm the tissue identity of uncharacterised libraries, for cancer staging and to indicate the degree of normalisation of a normalised library [23]. Here 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

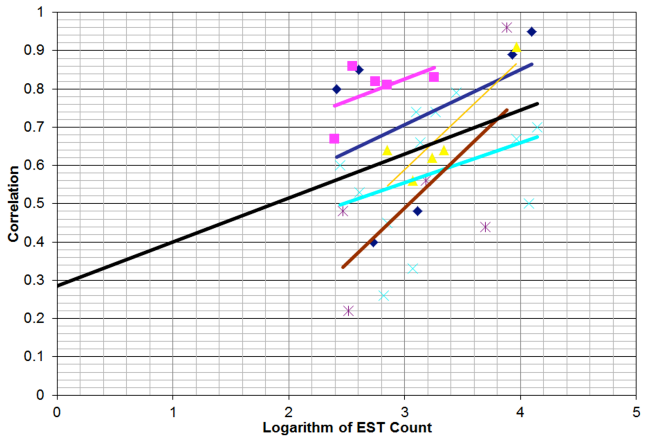

Figure 3. Pearson correlation values of original EST libraries (y-axis) vs. EST count (x-axis). The black trendline is fitted to all of the data points shown (all tissues), while the other trendlines are fitted to the individual tissues: (dark blue) lung. (pink) pancreas. (light blue) placenta. (yellow) retina. (brown) testis.

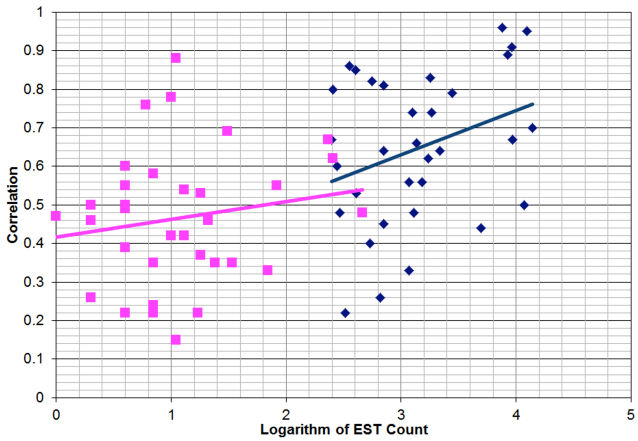

Figure 4. Pearson correlation values of original and scaled down libraries (y-axis vs. EST count (x-axis): (blue) dta points corresponding to the original libraries, as shown in Figure 3. (pink) data points representing the modelled scaled down libraries, although the modelling involved non-linear transformation of the data, the graph shows similar degree of positive correlation between 0.15 and 0.88.

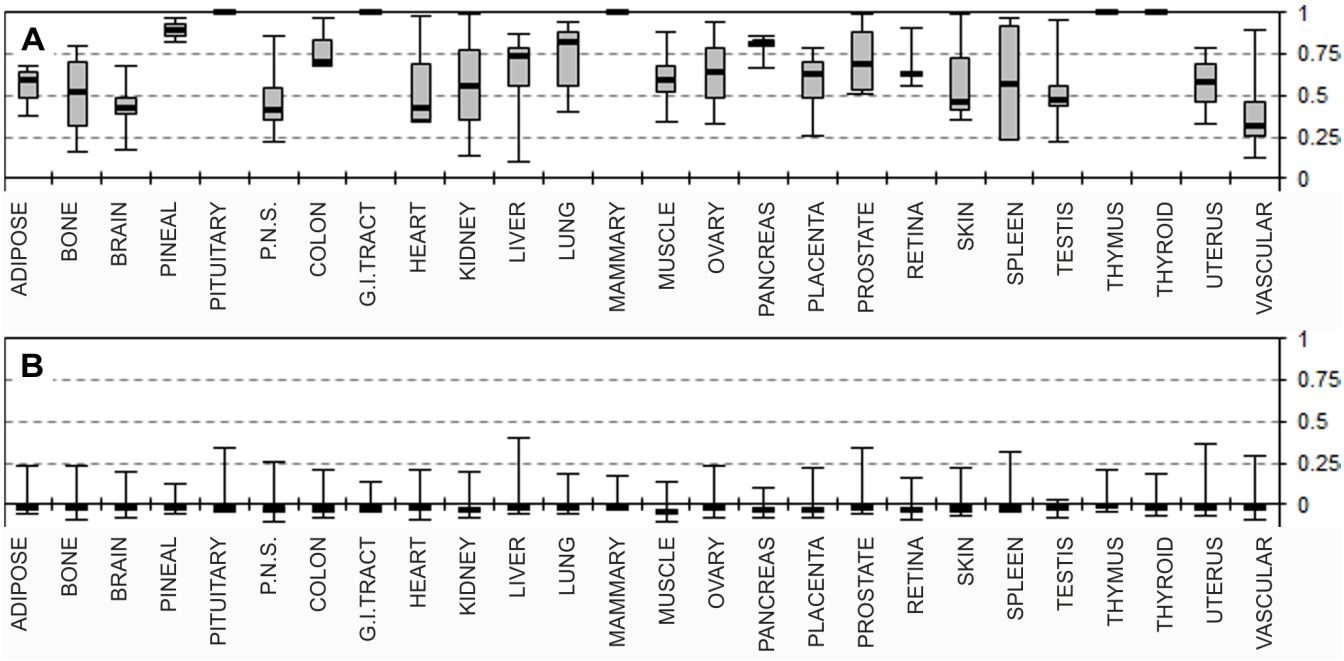

Figure 5. Intra-tissue and inter-tissue correlations. Correlation coefficients calculated for all of the 113 EST libraries against the EST expression matrix: (a) positive correlations between all expected matching libraries, e.g. all individual "Adipose" libraries vs. the "Adipose" expression matrix etc. Correlation value of "1" is for tissues where only one EST library was available. (b) Correlations for all expected non-matching libraries, e.g. all "Adipose" libraries available vs. all but the "Adipose" expression arrays from our EST matrix etc. The presumed mixed tissue brain library "NIH_MGC_181" was excluded from calculations. Reproduced with permission from [23].

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 modelled real EST expression data by scaling down the reported EST counts and rounding any fractional values to the nearest whole EST count each time. We continued this until each library ceased to present any ESTs mapping onto the 244 transcripts or ceased to be identified as a positive tissue match for the tissue from which it was created in the first place. We 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. Most of the libraries tested continue 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. Our results indicate that the majority of the scaled down libraries remain identifiable until the total library EST counts falls below the range of 10 to 50 total EST counts, which corresponds to some of the smallest libraries currently in the CGAP database, Figure 2.

We have further investigated whether the quality of tissue matching (the positive correlations calculated) depends on the library size or not. We carried such analysis for the same five tissues as reported in Figure 1; each tissue was analysed separately. The results presented here show a clear positive correlation between the size of the library and the quality of tissue matches (positive correlation and matching of the correct tissue, with values of between 0.22 and 0.96), see Figure 3. Smaller libraries tested were also identifiable, albeit with smaller correlation values of between 0.15 and 0.88, see Figure 4. We have earlier reported inter-tissue correlation coefficients for 26 individual tissues for which EST expression data were available from CGAP, see Figure 5. These indicate that the EST expression matrix is capable of correctly identifying all the non-matching tissues (median correlation values of ~ –0.02 yet some individual expression libraries may yield relatively high false positive correlation values (median ~ 0.21). Taking into account the trend data shown in Figure 3, which indicate that on average a library as small as one EST count is likely to yield correct positive correlation of ~ 0.3, our expression matrices should in principle be suitable for EST libraries containing just a few EST counts.

III. CONCLUSIONS

An EST expression matrix has been optimised and tested here on EST libraries of a range of sizes. We showed that the approach is correct and robust, applicable to EST libraries of different sizes. The matrix can be also be used to identify the disease state of a tissue and identify normalized or otherwise modified expression libraries [23]. The reported data indicate

that even a small subset of EST expression data sets containing 10 – 50 ESTs can still contain sufficient information to determine the tissue specificity and be of use for quality control. Although originally developed to test EST expression libraries, testing our approach with SAGE or siRNA expression data and other transcriptomics analyses such as microarray data is now justified.

REFERENCES

[1] S. Liu and J. H. Graber, "Detecting and profiling tissue-selective genes," Physiol. Genomics, vol. 26, Feb. 2006 pp. 158-162.

[2] J. Song, "What a wise SAGE Once Said about Gene Expression", BioTeach. J., vol. 1, Apr. 2003 pp. 99-104.

[3] A. Ray, S. Macwana, P. Ayoubi, L. T. Hall, R. Prade et al., "Detecting and profiling tissue-selective genes," Physiol. Genomics, vol. 26, Feb. 2006 pp. 158-162.

[4] E. Beaudoing, S. Freier, J. R. Wyatt, J. M. Claverie and D. Gautheret, "Patterns of variant polyadenylation signal usage in human genes," Genome Res., vol. 10, Jul. 2000 pp. 1001-1010.

[5] A. T. Milnthorpe and M. Soloviev (2011) "Errors in CGAP xProfiler and Cdna dged: the importance of library parsing and gene selection algorithms" BMC Bioinformatics, vol. 12 p. 97.

[6] L. Huminieck, A. T. Lloyd and K. H. Wolfe, "Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases", BMC Genomics, vol. 4, Jul. 2003 p. 31

[7] Q. Li, X. Liu, Q. He, L. Hu, Y. Ling et al., "Systematic analysis of gene expression level with tissue-specificity, function and protein subcellular localization in human transcriptome", Mol. Biol. Rep., vol. 38, Apr. 2011 pp. 2597-2602.

[8] S. Liang, Y. Li, X. Be, S. Howes and W. Liu, "Detecting and profiling tissue-selective genes", Physiol. Genomics, vol. 26, Jul. 2006 pp. 158-162.

[9] S. Arhondakis, O. Clay and G. Bernardi, "Compositional properties of human cDNA libraries: practical implications", FEBS Lett., vol. 580, Oct. 2006 pp.5772-5778.

[10] J. Russ and M. E. Futschik, "Comparison and consolidation of microarray data sets of human tissue expression", BMC Genomics, vol. 11, May 2010 p. 305.

[11] R.M. Hu, Z. G. Han, H. D. Song, Y. D. Peng, Q. H. Huang et al., "Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning", Proc. Natl. Acad. Sci. U. S. A., vol. 97, Aug. 2000 pp. 9543-9548.

[12] S. Krief, J. F. Faivre, P. Robert, B. Le Douarin, N. Brument-Larignon et al., "Identification and characterization of cvHsp. A novel human small stress protein selectively expressed in cardiovascular and insulin-sensitive tissues", J. Biol. Chem., vol. 274, Dec. 1999 pp. 36592-36600.

[13] D. Miner and A. Rajkovic, "Identification of expressed sequence tags preferentially expressed in human placentas by in silico subtraction", Prenat. Diagn., vol. 23, May 2000 pp. 410-419.

[14] S. Y. Pao, W. L. Lin and M. J. Hwang, "In silico identification and comparative analysis of differentially expressed genes in human and mouse tissues', BMC Genomics, vol. 7, Apr. 2006 p. 86.

[15] B. L. Vaes, K. J. Dechering, A. Feijen, J. M. Hendriks, C. Lefèvre et al, "Comprehensive microarray analysis of bone morphogenetic protein 2-induced osteo-blast differentiation resulting in the identification of novel markers for bone development", J. Bone Miner. Res., vol. 17, Dec. 2002 pp. 2106-2118.

[16] A. Elfilali, S. Lair, C. Verbeke, P. La Rosa, F. Radvanyi et al., "ITTACA: a new database for integrated tumor transcriptome array and clinical data analysis", Nucleic Acids Res., vol. 34, Jan 2006 pp. D613-D616.

[17] R. L. Strausberg, A. A. Camargo, G. J. Riggins, C. F. Schaefer, S. J. de Souza et al., "An international database and integrated analysis tools for the study of cancer gene expression", Pharmacogenomics J., vol. 2, Feb. 2002 pp. 156-164.

[18] Y. Zhang, D. A. Eberhard, G. D. Frantz, P. Dowd, T. D. Wu et al., "GEPIS—quantitative gene expression profiling in normal and cancer tissues", Bioinformatics, vol. 20, Oct. 2004 pp. 2390-2398.

[19] A. C. Brown, K. Kai, M. E. May, D. C. Brown and D. C. Roopenian, "ExQuest, a novel method for displaying quantitative gene expression from ESTs", Genomics, vol 83, Mar. 2004 pp. 528-539.

[20] S. Kawamoto, Y. Matsumoto, K. Okubo and K. Matsubara, "Expression profiles of active genes in human and mouse livers", Gene, vol. 174, Sep. 1996 pp. 151-158.

[21] K. Okubo, N. Hori, R. Matoba, T. Niiyama, A. Fukushima et al., "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression", Nat. Genet., vol. 2, Nov. 1992 pp. 173-179.

[22] L. Skrabanek and F. Campagne, "TissueInfo: high-throughput identification of tissue expression profiles and specificity", Nucleic Acids Res., vol. 29, Nov. 2001 p. E102.

[23] A. T. Milnthorpe and M. Soloviev, "The use of EST expression matrixes for the quality control of gene expression data", PLoS One, vol. 7, Mar. 2012 e32966.