The use of EST expression matrices for the quality control of gene expression data and the development of improved algorithms for gene expression profiling in cancer

Andrew Timothy Milnthorpe

Royal Holloway, University of London

PhD

Declaration of Authorship

I, Andrew Milnthorpe, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed: A. Milnthorpe

Date: 23/08/2012

Acknowledgements

Andrew Milnthorpe carried out most of the experimental work and drafted the thesis. Mikhail Soloviev conceived and supervised the study, performed some of the experiments and contributed to the drafting of the thesis. Both Andrew Milnthorpe and Mikhail Soloviev contributed materials and read and approved the final thesis.

We are grateful to Bracknell Forest Council and the University of London Central Research Fund who provided travel funds for Andrew Milnthorpe to visit the National Cancer Institute, Center for Biomedical Informatics and Information Technology, Rockville MD, USA and Johns Hopkins University, Baltimore, MD 21231, USA.

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 data for gene expression levels. The Cancer Genome Anatomy Project (CGAP, run by NCBI) cDNA xProfiler and cDNA DGED tools can be used to examine EST to compare gene expression levels between cancer and normal tissue. However, neither CGAP nor other similar tools provide an easy way to compare expression in normal and cancerous tissue with e.g. expression levels in related or proximal tissues at the same time while also presenting that data for study separately. Furthermore, the expression data are often assumed to be correct and no quality control tools are made available at CGAP, dbEST and GEPIS. In this study the CGAP tools were recreated with the aim of enabling a wider range of tissues to be searched and compared in a single search. The CGAP tools were found to contain many errors in their library and gene parsing algorithms, for which solutions were implemented in the recreated algorithms. A method was also devised for the tissue origin of EST libraries to be verified and for the uncharacterised libraries to be annotated with a likely tissue of origin using EST data alone. An initial list of tissue-specific genes was optimised to create gene expression matrices which could be used to determine the tissue origin of a library. The matrices were demonstrated to show potential for cancer staging and for the indication of the degree of normalisation of a library in addition to tissue typing, making tissue-specific expression a suitable quality control method for expression data. Together the improved expression profiling algorithm and the expression matrices provide new tools to assess the quality of EST data and their suitability for expression profiling.

Table of Contents

Index of Figures

Index of Tables

1. Background

1.1. Introduction

The UniGene database (National Center for Biotechnology Information, *n.d.a*) contains 6,877,952 ESTs (expressed sequence tags) for humans mapping onto 129,525 UniGene Clusters, each of which represents an mRNA transcript (as of 20 July *2012*). Similarly, as of 10 February 2012, the Probe database (National Center for Biotechnology Information, *n.d.b*) contains 11,835,370 nucleotide sequences of many different kinds which have been collected in a variety of investigations including gene expression studies and genome mapping experiments.

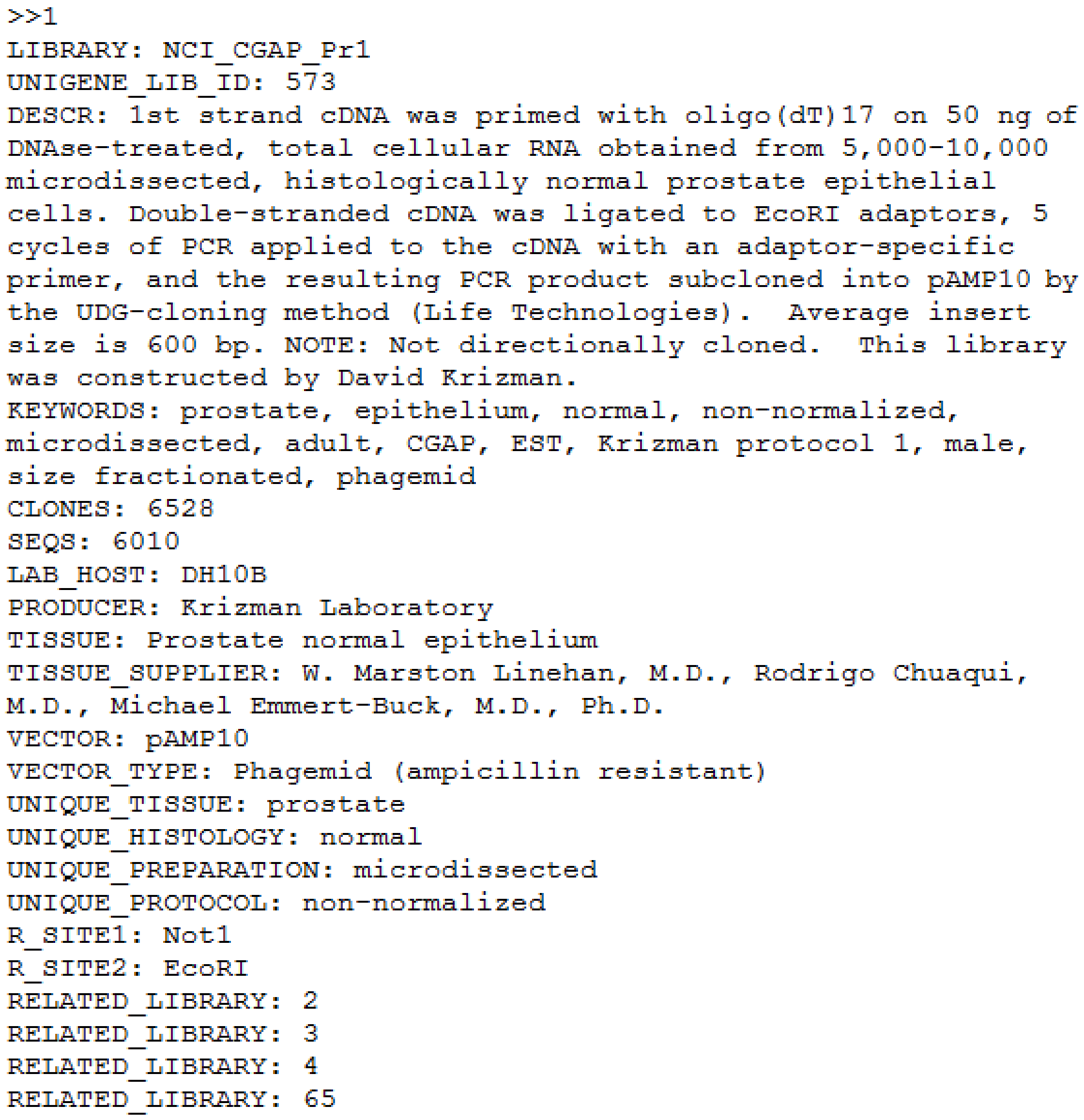
These large amounts of data require new approaches to annotation and access the data. Both UniGene and Probe host search algorithms that enable the database to be queried for data the user wishes to access. Other relevant databases and search tools include ArrayExpress Archive (European Bioinformatics Institute, *2012*), Gene Expression Omnibus (National Center for Biotechnology Information, *n.d.c*) and the Cancer Genome Anatomy Project (CGAP) (National Cancer Institute, *n.d.a*).

ESTs have previously been used for novel gene discovery and gene mapping (Adams et al, *1991*; Gudas et al, *1999*; Lee et al, *2001*) and the ever increasing amount of data can be examined to report gene expression levels and compare them between different tissues or conditions. CGAP exists to provide bioinformatics for the analysis of gene expression in cancer and the comparison of this with expression in normal tissue (Strausberg, *2001*). These tools allow analysis of a variety of different types of cancer such as breast cancer, and are very good and useful in principle.

1.2. The Cancer Genome Anatomy Project

This database, which contains EST and SAGE (Serial Analysis of Gene Expression) data, will be used in this project. Managed by the US National Cancer Institute (NCI) and operational since 1996, CGAP's (National Cancer Institute, *n.d.a*) aim is to provide the information and technological tools needed to decipher the molecular anatomy of the cancer cell (Riggins and Strausberg, *2001*). Since then CGAP tools have been used worldwide for the analysis or for validation of the differential gene expression in e.g. brain cancers and retinoblastomas (Beaty et al, *2007*; Loging et al, *2000*; Shostak et al, *2003*; Yang et al, *2008*), breast cancer (Petersson et al; *2010*; Shen et al, *2005*; Yousef et al, *2004a*), colon cancer (Ahmed et al, *2007a*, *2007b*; De Young et al, *2002*; Deyoung et al, *2002*; Haung et al, *2006*; Nam et al, *2005*; Yousef et al, *2004b*), gastric cancer (Meng et al, *2007*), lung cancer (Bidon et al, *2001*), pancreatic cancer (Alaiya et al, *2000*; Elek et al, *2000*; Yousef et al, *2004b*, prostate cancer (Mitas et al, *2002*; Wu et al, *2006*) and haematological malignancies (Sher et al, *2005*) to name just a few. The improved methods of analysing and mining this data include an NCBI classification system based on hierarchically related keywords, assigned to each new library by NCI staff. Furthermore, CGAP (National Cancer Institute, *n.d.a*) hosts two bioinformatics tools, the cDNA xProfiler (National Cancer Institute, *n.d.b*) and the cDNA Digital Gene Expression Displayer (DGED) (National Cancer Institute, *n.d.c*), which are designed to enable a user to identify differentially expressed genes, e.g. between a cancer and a normal tissue, or compare gene expression between two user-selectable pools of EST libraries.

The data used by both tools is derived from the UniGene repository (National Center for Biotechnology Information, *n.d.a*), a publicly available relational EST library database maintained by the US National Center for Biotechnology Information (NCBI) in which the tag counts from submitted human or house mouse EST libraries are mapped onto UniGene IDs, the unique transcripts they most closely match. NCBI uses tissue type/sample type annotation to create an ontology hierarchy in which libraries are grouped into tissue types according to tissue dependency (bone marrow, for example is a constituent of bone tissue, so bone marrow libraries are listed under bone tissue). The user-submitted tissue type annotations are listed for each library under the "keywords" and "unique tissue" fields, see Figure 1, which shows the first entry of the CGAP EST library database.



```
>>1
LIBRARY: NCI_CGAP_Pr1
UNIGENE_LIB_ID: 573
DESCR: 1st strand cDNA was primed with oligo(dT)17 on 50 ng of
DNAse-treated, total cellular RNA obtained from 5,000-10,000
microdissected, histologically normal prostate epithelial
cells. Double-stranded cDNA was ligated to EcoRI adaptors, 5
cycles of PCR applied to the cDNA with an adaptor-specific
primer, and the resulting PCR product subcloned into pAMP10 by
the UDG-cloning method (Life Technologies).  Average insert
size is 600 bp. NOTE: Not directionally cloned.  This library
was constructed by David Krizman.
KEYWORDS: prostate, epithelium, normal, non-normalized,
microdissected, adult, CGAP, EST, Krizman protocol 1, male,
size fractionated, phagemid
CLONES: 6528
SEQS: 6010
LAB_HOST: DH10B
PRODUCER: Krizman Laboratory
TISSUE: Prostate normal epithelium
TISSUE_SUPPLIER: W. Marston Linehan, M.D., Rodrigo Chuaqui,
M.D., Michael Emmert-Buck, M.D., Ph.D.
VECTOR: pAMP10
VECTOR_TYPE: Phagemid (ampicillin resistant)
UNIQUE_TISSUE: prostate
UNIQUE_HISTOLOGY: normal
UNIQUE_PREPARATION: microdissected
UNIQUE_PROTOCOL: non-normalized
R_SITE1: Not1
R_SITE2: EcoRI
RELATED_LIBRARY: 2
RELATED_LIBRARY: 3
RELATED_LIBRARY: 4
RELATED_LIBRARY: 65
```

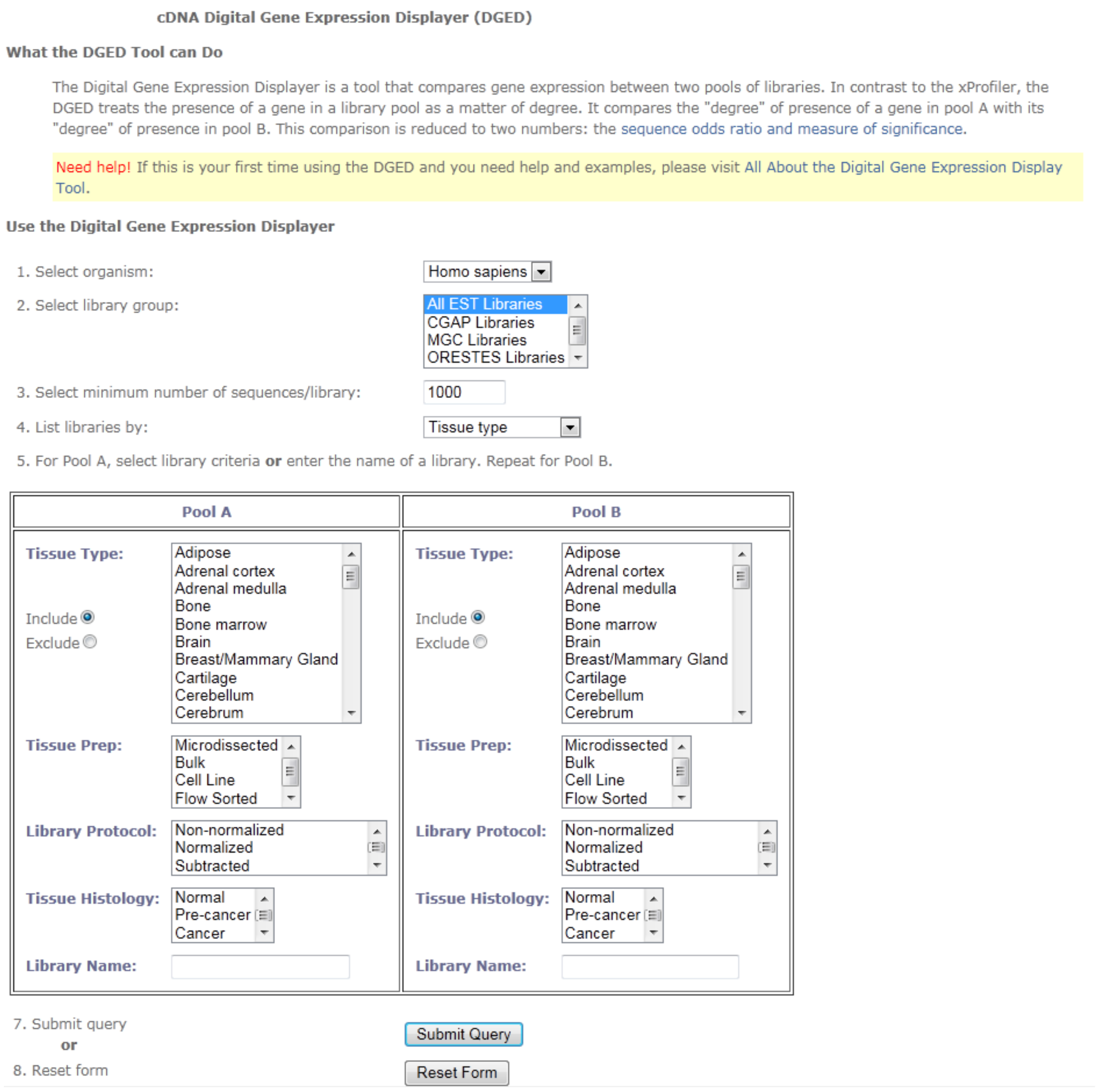
Figure 1. CGAP library database entry. Each field entry begins with its heading, shown in capital letters, followed by the value associated with that field, after the colon.

The CGAP tools use the values associated with the "keywords" field, to include or exclude libraries from a search based on the chosen tissue type and the inclusion of any dependent tissues under the selected one in CGAP's ontology hierarchy. Using this field, which also includes information on a library's histology, libraries from a secondary tumour (for example, neuroblastoma which has metastasised to bone marrow) can also be listed under the tissue in which the primary tumour is located (brain in the case of neuroblastoma) (Murray et al, *2007*; Ootsaka et al, *2008*).

cDNA DGED relies on a publically accessible relational database (National Cancer Institute, *n.d.d*), whilst the cDNA xProfiler accesses a single-file database, not available for online access, and uses a Boolean search to identify the presence or absence of a transcript in either or both of two groups (pools) of libraries which the user has chosen to compare to find differentially expressed genes. It lists the results as a table detailing how many matching transcripts have a known or unknown name and/or function (listed as known or unknown) and how many are found only in the libraries in the two pools or in at least one library outside the two pools (listed as unique or non-unique (National Cancer Institute, *n.d.e*), also reviewed in (Murray et al, *2007*).

Although the presence of a transcript in a particular library can be revealing, the outcome would depend on many parameters, including the size of the libraries used, and is therefore of limited biological significance. To overcome this, the cDNA DGED calculates sequence odds ratios for individual transcripts expressed in the two pools (Lal et al, *1999*) and calculates the statistical significance for the difference in the expression level of each gene between the two pools. Thus the cDNA DGED yields the most biologically relevant prediction – the normalised odds ratio, which at least in principle should be comparable to the results obtained through other methods based on Northern hybridisations or DNA microarrays.

The interface, see Figure 2, is straightforward and the calculations appear to be reliable. The relational database used by the cDNA DGED is also available in raw data format, enabling the use of alternative tools for interrogation. The same interface is used by both tools because they serve the identical purpose of finding genes which are differentially expressed in cancer compared to normal tissue. Once a search is initiated a list of matching libraries is presented, which the user can modify if they wish. At this point cDNA DGED allows the user to supply values for the statistical filters used by this tool to omit transcripts whose upregulation or downregulation does not meet the required significance. Once the library selection and statistical parameters are satisfactory, the search for upregulated and downregulated transcripts can proceed.



cDNA Digital Gene Expression Displayer (DGED)

What the DGED Tool can Do

The Digital Gene Expression Displayer is a tool that compares gene expression between two pools of libraries. In contrast to the xProfiler, the DGED treats the presence of a gene in a library pool as a matter of degree. It compares the "degree" of presence of a gene in pool A with its "degree" of presence in pool B. This comparison is reduced to two numbers: the sequence odds ratio and measure of significance.

Need help! If this is your first time using the DGED and you need help and examples, please visit All About the Digital Gene Expression Display Tool.

Use the Digital Gene Expression Displayer

1. Select organism: Homo sapiens
2. Select library group: All EST Libraries
CGAP Libraries
MGC Libraries
ORESTES Libraries
3. Select minimum number of sequences/library: 1000
4. List libraries by: Tissue type
5. For Pool A, select library criteria **or** enter the name of a library. Repeat for Pool B.

	Pool A		Pool B
Tissue Type: Include ● Exclude ○	Adipose Adrenal cortex Adrenal medulla Bone Bone marrow Brain Breast/Mammary Gland Cartilage Cerebellum Cerebrum	Tissue Type: Include ● Exclude ○	Adipose Adrenal cortex Adrenal medulla Bone Bone marrow Brain Breast/Mammary Gland Cartilage Cerebellum Cerebrum
Tissue Prep:	Microdissected Bulk Cell Line Flow Sorted	Tissue Prep:	Microdissected Bulk Cell Line Flow Sorted
Library Protocol:	Non-normalized Normalized Subtracted	Library Protocol:	Non-normalized Normalized Subtracted
Tissue Histology:	Normal Pre-cancer Cancer	Tissue Histology:	Normal Pre-cancer Cancer
Library Name:		Library Name:	

7. Submit query

or

8. Reset form

Figure 2. CGAP's cDNA DGED input page. This screenshot of part of the input page shows the options available for comparing two pools of libraries. Some of the settings, such as tissue type and tissue preparation, apply to just one pool, while other choices, such as library group and organism, are applied to both pools.

1.3. Other tools which use EST data

CGAP's EST databases are not the only EST gene expression databases available. The dbEST (National Center for Biotechnology Information, *n.d.d*), which is the EST division of GenBank, is also available for use and includes the CGAP EST data. To query the dbEST data, tools have been developed such as Digital Differential Display (DDD) (National Center for Biotechnology Information, *n.d.e*) and DigiNorthern (no longer available). DDD provides the user with a list of EST libraries (each of these contains EST sequencing data from a single experiment on a single tissue sample) from a range of tumour and tissue types for inclusion in each of two pools for comparison, while DigiNorthern allowed the user to enter one or two query sequences to reveal all the cell lines, tissues or organs the genes those sequences map to are expressed in and show the relative expression levels between those sources. If two sequences were entered their expression profiles were compared and displayed together. DigiNorthern also provides comparison with any types of cancer the gene is found in (Chen et al, *2006*; Riggins and Strausberg, *2001*; Wang and Liang, *2003*).

The CGAP tools will be used because they enable two groups of samples, each of which may contain many individual sequences, to be compared. While both tools allow species selection at the beginning of a search, the CGAP tools also provide other options for displaying libraries matching only the desired tissue type, tissue preparation method and so on, and allow the settings for both pools to be selected at once. dbEST is less user-friendly because all the libraries in its database are presented at once for each pool without showing the other pool at the same time, and the user has to manually select each library they want to analyse for Pool A, and then repeat the process for Pool B.

Though widely used, the CGAP tools are not the first attempt to effort to acquire gene expression data. Begun in 1991, BodyMap (now unavailable) was the first systematic effort to acquire gene expression data and, like CGAP, contains the transcript compositions of various human tissues. Unlike the CGAP tools, in which the user chooses two tissues and is presented with differential gene expression levels in those tissues, BodyMap requires the user to enter a single sequence or UniGene Cluster ID of their choice to discover expression levels for that transcript in different body tissues. Thus, unlike CGAP, it was not suitable for high-throughput analyses, and therefore is not so suitable for comparing all the transcript levels in one sample with those in another (Kawamoto et al, *2000*).

The mechanisms underpinning changes in gene expression can often be better understood through the analysis of the expression levels of genes neighbouring those initially found to be differentially expressed. GEPIS (Gene Expression Profiling *In Silico*) (Genetech Inc, *n.d.*) not only compares gene expression in normal and cancerous tissues based on EST data, but presents a graphical view of a region of interest in the genome to show the expression patterns of neighbouring genes (Zhang et al, *2004*). However, the results do not take into account alternative splicing, for each gene is considered as a single entity without mention of multiple transcripts which might be produced.

1.4. EST and SAGE Libraries

ESTs are created by sequencing cDNA libraries. To create a cDNA library, the sample’s RNA content is extracted (Peterson et al, *1998*), before being purified (Israeli et al, *1993*) and copied by reverse transcriptase to form cDNA (Coutelle et al, *1985*). Once the cDNA library has been created, ESTs are produced by sequencing randomly selected transcripts, usually from the 3’ end to generate single read fragments which are often longer than several hundred base pairs in length. These are then assembled into longer, overlapping sequences mapped onto the original transcript. Though originally used for gene discovery, ESTs can also be used for expression profiling. mRNA expression levels are inferred by counting the absolute number of tags representing each transcript, and it is for this reason that this technique is sometimes referred to as a “digital” gene expression profiling method. EST libraries contain a snapshot of mRNAs expressed in the sample from which the library was created (Shmulevich and Zhang, *2002*). EST expression profiling is a well-established high-throughput method for acquiring quantitative information on a sample’s transcriptome and for studying differential gene expression, inferred from the differences in the relative numbers of EST tags between two libraries.

ESTs are sequence reads which enable the absolute expression levels of all the genes in a sample to be determined, as is also the case with a related technique called Serial Analysis of Gene Expression (SAGE). However, instead of concatenating the tags together as is done for SAGE, ESTs are individually sequenced, and differential expression is inferred from comparing the number of tags representing each gene in each of two samples to the total size of each sample. This makes EST advantageous over SAGE in the event of a sequencing error, which will only affect the EST concerned, whereas a whole SAGE concatamer (Ichikawa et al, *2004*), containing many individual tags, will be impacted. Furthermore, with lengths of more than a few hundred base pairs, ESTs are significantly longer than SAGE tags, which are far more likely than ESTs to map onto two or more transcript simultaneously due to their short length. Thus the use of ESTs vastly reduces the risk of ambiguity in the results (Adams et al, *1991*; Audic and Claverie, *1997*; Pariset et al, *2009*; Simon et al, *2009*).

If the purpose of an EST library is for gene discovery rather than expression profiling, the EST content of a library can be altered to reduce the abundance of transcripts representing genes with high expression. To achieve this, a library can be normalised by removing the most abundant transcripts in order to reduce or eliminate the differences in the relative transcript abundances to a narrow range (Arhondakis et al, *2006*; Bonaldo et al, *1996*; Sasaki et al, *1994*; Soares et al, *1994*).

To increase the likelihood of discovering novel genes, cDNA libraries can be normalised, a process in which the most abundant sequences are removed to bring the transcript abundances to within an order of magnitude of one another. Ideally this should create a library containing the same tag counts for the low abundance sequences as before, but with vastly reduced counts for abundant cDNAs. cDNA libraries can also be subtracted, a process in which, in addition to normalisation of the library, all previously known sequences are removed, producing a subtracted library which ideally contains only novel transcripts whose abundances are within an order of magnitude of one another (Bonaldo et al, *1996*).

At the beginning of the normalisation process single-stranded cDNA is produced from the double-stranded plasmids. This can be done either *in vitro* (Bonaldo et al, *1996*) or by transformation into bacteria (Panja et al, *2006*) and infection with the helper phage MK13K07 (Soares et al, *1994*). Because the rate of an enzymic reaction is directly proportional to the substrate concentration, using both techniques, proportionately less of the low abundance species will be present in single-stranded form at the end of the procedure, compared to the cDNA species which are more abundant and therefore present in higher concentration. Thus, a transcript for which there are 10,000 copies present will be digested at a rate ten times greater than the rate of digestion of a transcript for which there are only 1,000 copies present.

To remove any contaminant single-stranded DNA and therefore produce a mixture containing only the rarer transcripts, the whole mixture is passed down a column containing hydroxyapatite, to which the double-stranded molecules preferentially bind due to their greater negative charge (Andrews-Pfannkoch et al, *2010*).

The above procedures may be repeated at least once on the bound fraction to increase the enrichment of the library for low abundance transcripts and therefore increase the degree of normalisation. In order to do this a second strand is synthesised on each plasmids (Bonaldo et al, *1996*). At the end of the procedure an ideal normalised library will have the same number of plasmids containing each cDNA species.

The above procedures should result in reduced (ideally identical) transcript abundances and therefore a lower mean EST count per transcript than the equivalent non-normalised library. As a consequence of this, in a normalised library there should also be more transcripts detected after the same amount of sequencing.

1.5. Cancer

1.5.1. The cell cycle under normal conditions

The CGAP tools exist to allow the molecular changes which occur during oncogenesis to be better understood. These changes cause the cell concerned to divide indefinitely rather than in response to external signals. This makes the cell cancerous, and the cell will continue to divide to form a clump of cells called a tumour. The molecular changes which cause this indefinite proliferation involve deregulation of the cell cycle (the cycle of duplication and division of a cell), which is divided into the G_1, S, G_2 and M phases (Morgan, *2002*; Weinberg, *2007*).

In the S (synthesis) phase DNA replication occurs, taking 10-12 hours in mammalian cells, which will be the focus of this report. The resulting two identical sets of chromosomes are segregated and the cell divides in M (mitosis) phase. During this phase the cell takes under an hour to complete a series of events which begin with mitosis (nuclear division, in which the two sets of chromosomes are separated to opposite ends of the cell) and end with cytokinesis (division of the cytoplasm into two daughter cells).

The G_1 and G_2 phases are gap phases that allow the cell to grow by doubling its mass of proteins and organelles, a process which takes far longer than the 11-13 hours required for the S and M phases (Morgan, *2002*).

Each stage is under the control of various protein complexes and signalling pathways, some of which induce progress through the cycle, and therefore promote cell proliferation, and some of which repress proliferation. The signalling pathways act as quality control mechanisms because they ensure that the cell only divides when it is required to do so, and external signals triggering the pathways bring about division.

The protein complexes are called checkpoint complexes and arrest the cell cycle until any detected errors in the cell's mechanisms for DNA replication and chromosome duplication and segregation are repaired. If these errors cannot be repaired the cell may enter a resting state called G_0 or it could undergo senescence (ageing) or apoptosis (programmed cell death) (Figure 3A) (Delaval and Birnbaum, *2007*; Morgan, *2002*).

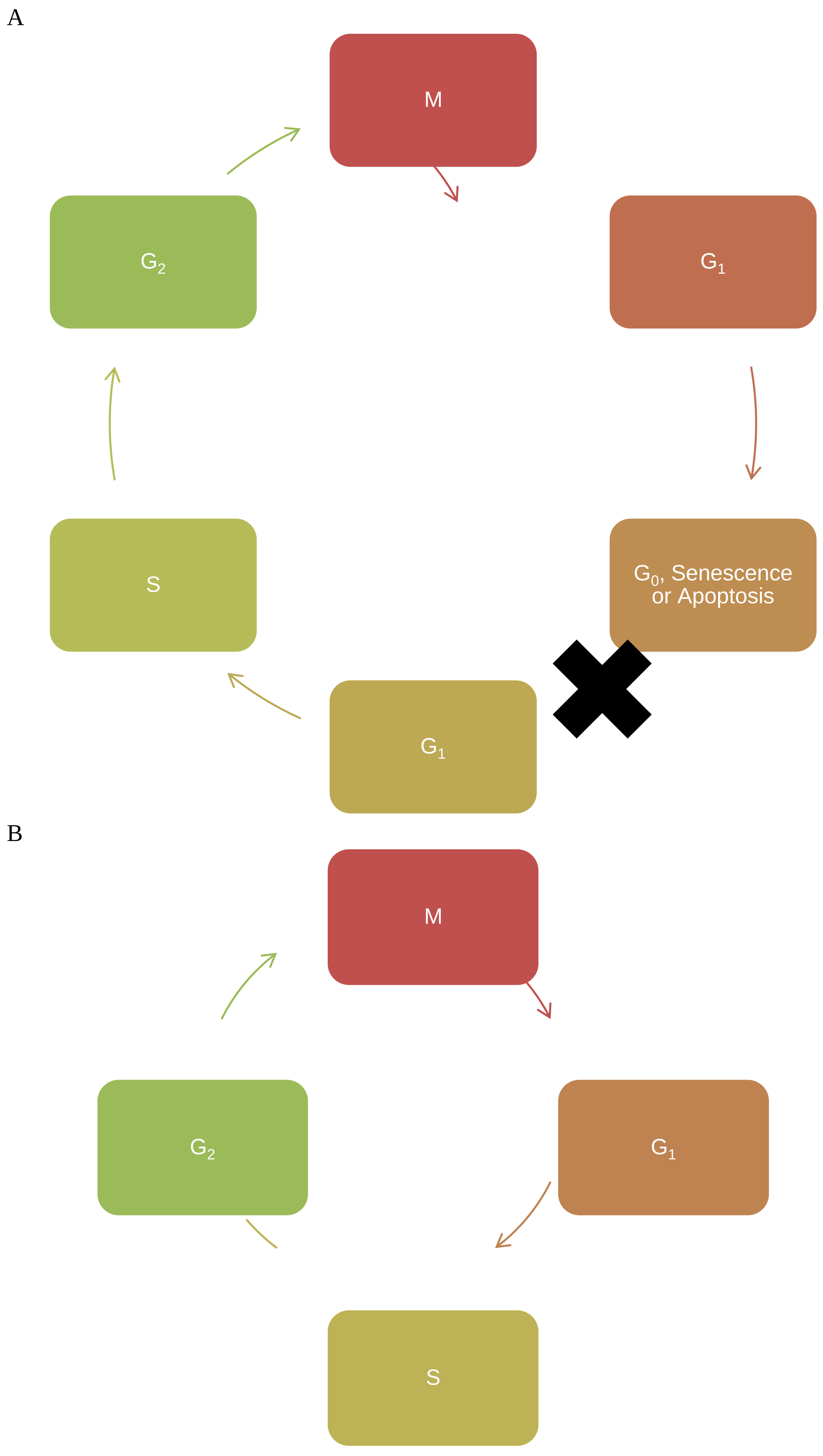

A
M
G_2
G_1
G_0, Senescence or Apoptosis
S
G_1
B
M
G_2
G_1
S

Figure 3. Changes to the cell cycle which result from the disruption to checkpoint complexes in cancer. A: Cell cycle arrest due to errors detected by undamaged checkpoint complexes in normal cell. **B:** Continuous cell cycle resulting from disruption to the checkpoint complexes due to mutations and gene expression changes in cancer.

Some of the molecular changes which disrupt the above quality control mechanisms and bring about the onset of cancer (Figure 3B) include mutations (changes in gene structure), whole others involve changes in gene expression even though those genes are not mutated. Such genes are said to be differentially expressed in cancer compared to normal tissue) and can be used as biomarkers in diagnosis or targets for anticancer therapy (Larsson et al, *2010*; Morgan, *2009*; Salama and Platel, *2009*; Troncone et al, *2010*; van Eijk et al, *2010*).

1.5.2. The importance of gene expression in diagnosis and treatment

Cancer biomarkers have been shown to be increasingly important in cancer diagnosis, a key part of which is determining what stage of cancer a patient has because tumours are stratified into two types according to disease stage. A benign tumour cannot invade adjacent tissues or metastasise to other organs, and will therefore only cause death if it presses against nearby organs or causes physiological changes through hormone imbalances. Therefore a patient with a benign tumour has a much greater chance of survival than a patient with a malignant tumour, which does invade adjacent tumours and metastasise to other organs, causing the majority of cancer-related deaths in the process (Weinberg, *2007*).

Cancers have traditionally been diagnosed according to their tissue of origin, and the stage of the disease, and the method used has usually been their appearance under the microscope or location in an MRI scan, a method called histopathology. Cancer disease biomarkers have been used on some occasions to refine traditional methods of classification but histopathology has been the usual method (Livingston and Shivdasani, *2001*; Weinberg, *2007*).

In the last few years, however, investigations have shown that biomarkers could be a potential diagnostic tool. Investigations over that time have begun to alter the way different cancers are stratified into groups using the above histopathological methods. Cancer disease biomarkers can provide clinicians with an accurate diagnosis, prognosis, assessment of treatment options, or likelihood of recurrence. This is because patients with similar histopathological diagnoses can have very different outcomes, and histopathological classification gives very little information about the above parameters (Arsanious et al, *2009*).

For example, an investigation into diffuse large B-cell lymphoma found that this disease, previously thought to be one condition, was in fact two molecularly distinct diseases whose expression patterns showed similarities to two different B-cell differentiation stages. One type was germinal centre B-cell like DLBCL, which was found in 40% of patients who had a significantly better survival rate than the other 60% of patients who had the other form, which was activated B-like DLBCL. Another investigation revealed a third type called primary mediastinal lymphoma, which, like activated B-cell like DLBCL, has a worse prognosis than germinal centre B-cell like DLBCL. This showed that classifying tumours according to their gene expression profiles could identify previously undetected types of cancer, enabling an accurate prognosis to be provided and the correct treatment to be administered. This is particularly important because incidence of the three conditions has subsequently been found to vary between countries (Alizadeh et al, *2000*; Ke et al, *2010*; Weinberg, *2007*).

Even if tumours are stratified correctly according to more established methods, they may still be classified differently according to gene expression profile. For example, gene expression profiling of prostate tumours revealed three distinct subtypes, each of which combined primary tumours and metastases in varying proportions (Lapointe et al, *2004*), suggesting that gene expression profiling may aid in locating the primary tumour by studying the expression profile of one or more metastases.

This shows that knowledge of a cancer's gene expression profile and of changes in expression as a consequence of disease progression are essential for diagnosis. Furthermore, cancers in different patients need to be correctly stratified because the term "cancer" covers a wide range of diseases, each of which has its own characteristic gene expression profile and each of which will arise from a different tissue in the body. For this reason it is vital that the gene expression profile of each normal tissue is known so that cancer samples can be compared with it, both for finding new biomarkers and testing patients' samples for diagnosis, prognosis or monitoring. Once the gene expression levels have been obtained, if they are deposited in the appropriate database, bioinformatics tools such as those hosted by CGAP can then be used to analyse the results and compare them against any other sample the user chooses to correctly diagnose the cancer.

Differentially gene expression has also been shown to be useful in providing an accurate prognosis. This was shown when the expression levels of 70 genes in primary breast tumours were found to be an indicator of the likelihood of distant metastasis within 5 years in 83% of patients studied. Furthermore, these genes were found to be involved in processes essential for tumour development such as cell cycle regulation and angiogenesis (blood vessel formation). Moreover, 70-80% of the patients deemed eligible for chemotherapy based on histology and clinical characteristics, would not benefit from treatment because the tumour would not have formed distant metastases if left untreated. This shows that gene expression profiling, together with *in silico* analysis of the results with tools such as those hosted by CGAP, is an essential tool for the provision of an accurate prognosis (van't Veer et al, *2002*).

Differential gene expression has also found to indicate the likelihood of cancer recurrence. In a study of grade two breast tumours, a stage at which it is impossible to determine the likelihood of recurrence from the tissue histology, analysis of expression levels of 97 genes found to be associated with disease stage divided the patients into two groups of either high or low risk of recurrence (Sotiriou et al, *2006*). As with the earlier mentioned study of diffuse large B-cell lymphoma, this shows that gene expression profiling and use of appropriate bioinformatics tools is essential for accurately classifying tumours and producing an accurate prognosis.

Differentially expressed genes can be used as targets for anticancer therapy if the change in expression is required for the development of the disease. An example of such a gene is the one which encodes VEGF (vascular endothelial growth factor), whose expression induces angiogenesis in a wide variety of conditions in which this occurs (Neufeld et al, *1999*). It has been shown that targeting VEGF using small interfering RNA (siRNA) resulted in almost total inhibition of secretion of this protein in a prostate cancer cell line (Takei et al, *2004*). However, the major challenge is identifying delivery strategies suitable for clinical use (Bumcrot et al, *2006*). A clinical trial has been undertaken using a targeted nanoparticle to deliver siRNA against the M2 subunit of ribonucleotide reductase to melanoma tumours. The result was a reduction in expression which correlated with the siRNA dose (Davis et al, *2010*). This shows that gene expression profiling together with *in silico* analysis of the results can be used to provide novel treatments for cancer.

1.6. Current Problems

1.6.1. Existing CGAP tools do not currently allow all searches which might be required for effective gene expression profiling in cancer

Currently, existing tools (CGAP (National Cancer Institute, *n.d.a*), DDD (National Center for Biotechnology Information, *n.d.e*) and GEPIS (Genetech Inc, *n.d.*)) are only able to compare two groups of tissues at once, for example cancer from a chosen tissue with normal samples from the same tissue. Thus it is not possible to compare gene expression levels in three or more groups of tissues side-by-side in a single search. For example if the aim of the investigation was to study just one type of cancer in a specific tissue and ignore all other cancers in that tissue and all related or proximal tissues. To do this using CGAP's algorithms, multiple searches would have to be carried out, each set to present two of the desired groups of tissues, and the results would have to be merged using other software. Such a comparison would enable gene expression within the organ concerned to be compared with expression in the rest of the system and with that in nearby unrelated tissues, with it still being possible to analyse each set of results individually. For example, genes could be reported which are expressed only in thyroid and not in the related parathyroid, the proximal oesophagus or muscle or the connected peripheral nervous system and vascular tissue. The identification of such transcripts would improve the reliability and accuracy of any diagnostic or prognostic tests developed from suggested biomarkers and would also eliminate possible side effects from any new RNAi-based treatment developed against suggested targets.

1.6.2. Data is currently assumed to be correct with no means of quality control

Furthermore, the underlying data still requires a quality control method which would enable the identity of each library to be verified independently of any external information. This is 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 (Liu and Graber, *2004*). 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 because a significant percentage of mRNA species contain multiple polyadenylation sites, potentially leading to multiple ESTs being produced from one transcript (Beaudoing et al, *2000*). Such errors may lead to false positive results or the omission of potential diagnostic biomarkers or therapeutic targets from further investigations, which in turn may lead to erroneous diagnoses or incorrect treatments.

Analysis of gene expression data for quality control purpose has been attempted previously with SAGE data (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 (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. 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 authors reported 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 the tissue-specificity issues (Russ and Futschik, *2010*). A common shortcoming of many 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 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. This is because the quality of a library could depend on the depth of sequencing (the number of ESTs sequenced for inclusion in the library). A greater depth of sequencing would provide a better quantitative estimate of gene expression (Simon et al, *2009*) because low-abundance transcripts are more likely to be included (Bashir et al, *2010*), making the library more representative of gene expression in the original sample. It is for this reason that the effect of library size on gene expression results has been previously studied and/or taken into account in statistical tests, which have been applied to a range of different types of cancer (Abba et al, *2004*; Baggerly et al, *2003, 2004*; Robinson and Smyth, 2007; Ruijter et al, *2002*; Silveira et al, *2008*; Thygesen, *2006*). However, the effect of library size on inter-library correlations has not been previously studied, despite it being known that this parameter impacts the reliability of the results (Schaaf et al, *2008*).

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. 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 EST expression data place the emphasis on identification of the degree of over/under-expressed 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, 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 offline (Brown et al, *2004*; Kawamoto et al, *1996*; Okubo et al, *1992*; Skrabanek and Campagne, *2001*).

2. Aims and Objectives

2.1. Aims

The main aim of this research was to create a new quality control method for gene expression data and improved bioinformatics methods used to analyse differential gene expression, particularly in cancer.

We aimed to address the so far unresolved problem of controlling the quality of EST expression data. The usefulness of any such expression data depends overwhelmingly on the provided annotations and there is virtually no way of experimentally testing any of the datasets. Since annotations are often incomplete and inconsistent we decided to investigate if the quality of expression data could be tested on the data alone, rather than the annotations. Such a capability would enable the characterisation of libraries from unknown or un-annotated tissue samples, as well as the identification of libraries whose annotation is erroneous or whose identity is obscured by experimental error.

We also aimed to investigate whether partially normalised libraries, which are normally considered to be unsuitable for quantitative analysis of differential gene expression because of the changes in relative transcript abundance, could be identified from the expression data (not the annotations) and still used for quantitative expression profiling.

We also sought to investigate whether cancer staging could be undertaken by correlating their expression data with that of normal libraries of known tissue identity, instead of merely relying on the cancer libraries' annotation, which may not always be correct. We also studied whether the degree of normalisation of a normalised library could be obtained in a similar manner. Similarly, we intended to look into breast cancer stratification and whether this can be aided using the developed methods.

Finally, using simple model data, we aimed to test our data evaluation method on small libraries to assess whether it could be used to correctly obtain the tissue identity of such libraries, which are equivalent in size to some of the smallest and least representative libraries in CGAP's database.

2.2. Objectives

1. The first objective was to learn the CGAP algorithms by recreating them using Microsoft Excel. Having done that we also created an easy to use development tool so we could improve and modify any of the analysis methods.
2. As part of the creation of the new algorithm the second objective was to correct errors in the existing algorithms, which we suspected have existed at the time of embarking on this project and which became apparent whilst addressing first objective.
3. The third objective was to include the facility for selecting more than two pools of tissues for side-by-side comparison in a single search (as opposed to the two allowed at present by CGAP's algorithms and other basic comparison tools. This would enable, for example, genes to be investigated which are expressed only in one specific cancer of interest from a particular tissue, and not in any other type of cancer from that tissue or the normal tissue, as well as related or proximal tissues, with each tissue presented separately so its gene expression profile can be individually studied.
4. The fourth objective was to create a gene expression matrix with a small number of genes with known expression levels across the tissues, and to use these patterns of expression levels to elucidate (i) the tissue identity of libraries, (ii) normalisation status, (iii) to discriminate normal versus cancer-derived libraries, (iv) to attempt cancer staging.

5. The fifth objective was to study the potential of the EST expression matrix in identifying normalised libraries and indicating the degree of normalisation, using annotated normalised libraries as well as simple modelled data. Having done this we investigated the potential of the matrix in cancer staging, using annotated cancer libraries.

3. Materials and Methods

3.1. Materials

All experiments were carried out using 64-bit Microsoft Excel 2010 on a 64-bit Windows 7 workstation with 16GB of RAM. The CGAP EST expression data (National Cancer Institute, *n.d.d*) were used for the development of a Microsoft Excel-based tool. Data downloaded on 15 November 2008 was used to compare the number of transcript-mapping ESTs reported for a group of libraries with the library size annotations of those libraries. Data available on 2 July 2009 was used to produce the preliminary list of 1,437 transcripts in the creation of a quality control method based on tissue-specific expression. Data downloaded on 18 August 2010 was used to optimise that list to 244 transcripts and investigate the potential use of that list as a quality control method. Data downloaded on 9 March 2011 was used to study CGAP's statistics and investigate empty and missing database entries.

3.2. Methods used to analyse the existing CGAP tools for errors

3.2.1. Analysis of the library parsing algorithm

For each of the 56 "specified" tissues in CGAP's database a search was carried out using CGAP's cDNA xProfiler in which all libraries associated with that tissue were reported. All library protocols, tissue preparation methods and tissue histology annotations were included, and the "sequences" cut-off was set to zero, with the tissue type set to the tissue of interest. The settings were the same for both pools. This was done to present as many libraries as possible, regardless of whether they contained a representative profile of *in vivo* gene expression. The data used was that accessed by the cDNA xProfiler on 14 May 2010.

Each library reported for each tissue was studied to assign that library as correctly or incorrectly reported for that tissue. A library was considered correctly reported if its "unique tissue" annotation precisely matched the selected tissue type. All other libraries were considered incorrectly reported. The phrases "germ cell," "head and neck" and "stem cell" were not contained in any libraries' unique tissue annotations, so libraries were considered to be correctly reported for these tissues if their "keywords" annotations contained one of these phrases.

The bone search to test whether libraries from dependent tissues are consistently included with their parent tissue was performed on 25 January 2012. All of these experiments were performed by Andrew Milnthorpe.

3.2.2. Analysis of the cDNA xProfiler transcript lists

To check the lists of transcripts produced by the CGAP tools normal adipose tissue (Pool A) and cancerous adipose tissue (Pool B) were compared using both the cDNA xProfiler (National Cancer Institute, *n.d.b*) and cDNA DGED (National Cancer Institute, *n.d.c*), using the version of these tools available from on 16 March 2010. Bulk and non-normalised libraries were used. The "number of sequences" display cut-off was set to zero to include all libraries. When running searches using cDNA DGED the Bayesian probability "P" value and the calculated odds "F" ratio display cut-offs were both set to one to ensure that all transcripts were displayed to enable comparison of the results with those produced by the cDNA xProfiler, which does not have statistical filters.

The problem of the cDNA xProfiler's results table misreporting some non-unique transcripts as unique was discovered when bulk non-normalised non-cancerous tissue (Pool A) was compared with bulk non-normalised cancerous tissue (Pool B). The “sequences” display cut-off value was set to zero to include all libraries, and a search was undertaken for each of the 52 available tissues, with the tissue of interest chosen for each search. This was carried out using the version of the cDNA xProfiler available between 14 November 2011 and 17 November 2011. For tissues with which the problem occurred, further searches using the same settings were carried out for those tissues only. Unlike previous searches, during the library selection stage, the search was repeated multiple times, each with a different selection of libraries, until the cause of the problem was found. All of these experiments were carried out by Andrew Milnthorpe.

3.2.3. Estimating the reliability of the cDNA DGED’s statistical prediction of the reliability of gene expression

To test two different “F” value display cut-off settings, two CGAP cDNA DGED searches were run to compare normal bone with cancerous bone (accessed on 22 August 2011). Bulk non-normalised bone libraries were used, with normal tissue selected for Pool A and cancerous tissue selected in Pool B. The “number of sequences” display cut-off was set to zero to include all libraries. The bone libraries chosen were the ones whose unique tissue field contained the exact phrase “bone” and all other libraries that the CGAP tools map onto bone tissue were excluded. The “Q” value display cut-off was set to one for both searches to present all results regardless of their reliability.

The “F” value display cut-off was set to two for the first search to display every gene whose expression differed between the two pools by a factor of two or more. The “F” display cut-off was set to three for the second search to display every gene whose expression differed between the two pools by a factor of three or more. The output of the online search results was analysed for three genes whose “Q” values were close to zero when the “F” value cut-off was set to two and these “Q” values were compared with those obtained when the “F” value display cut-off was set to three. All of these procedures were undertaken by Andrew Milnthorpe.

3.2.4. Experiments to study CGAP’s database files

The comparison of normal adipose tissue with cancerous adipose tissue described above was repeated using CGAP's cDNA DGED (the database was last accessed on 6 January 2010). The number of ESTs contained within the libraries of each pool was counted for each pool (by summing together the values reported for the individual libraries from CGAP's library database annotations). This value was compared with the value reported by CGAP DGED gene list for the number of ESTs mapping onto all transcripts in the chosen libraries.

The CGAP raw data was examined to discover the erroneous annotation of library “SARS-Cov infected lung tissue” as containing no ESTs. To investigate further, a search was run using CGAP’s cDNA DGED in which all libraries were present in all pools. All tissues, tissue preparations, library protocols and tissue histologies were included in both pools, and the “number of sequences” display cut-off was set to zero. This was undertaken on 26 January 2012.

The version of the CGAP database available for download on 9 March 2011 was searched using Excel to count the number of expression database records listed for libraries not presented in the library database. In the same way the number of library database records included for libraries containing no transcript-mapping ESTs was also counted. The number of “gene database” records listed for transcripts which did not map onto any ESTs in any libraries was similarly recorded. All experiments were undertaken by Andrew Milnthorpe.

3.3. Methods used to solve problems identified during investigation of CGAP's databases and algorithms

3.3.1. Configuration of new library search algorithm to solve problem of incorrect or irrelevant libraries being reported by CGAP

Microsoft Excel was used to test the new library parsing algorithm. Initially this was designed to mimic CGAP tools and present the same libraries for each tissue as do the CGAP tools, so the transcript parsing algorithm and the reported number of ESTs per library could both be compared with CGAP's equivalents. Without this it would not be possible to report any differences in the reported EST counts and the "gene results" compared to the CGAP tools as being solely due to differences in those two features between the tools.

The new algorithm was then modified to assign libraries to tissues using their "unique tissue" field to present only the libraries which are associated with the selected tissue. Once this was done, a search for each available tissue was undertaken in which all libraries for that tissue were reported, regardless of whether they contained a representative profile of *in vivo* gene expression. Therefore all tissue preparation, library protocol and tissue histology choices were included and the "number of transcript-mapping ESTs" threshold was set to zero. This was carried out using the version of CGAP's data available on 2 January 2010. All studies were undertaken by Andrew Milnthorpe.

3.3.2. Creation of novel transcript search algorithms to resolve issue of different transcript lists being reported by CGAP

Also using Microsoft Excel, two new transcript search routines were designed (unlike the CGAP tools, these were called transcript searches rather than gene searches). One reports the presence or absence of that transcript in each pool of libraries, as reported by the cDNA xProfiler. The other reports the number of ESTs mapping onto that transcript in all of the libraries included in each pool and calculates the odds ratio for each transcript between the two pools, as does CGAP's cDNA DGED. Both algorithms rely on the UniGene Library ID (the unique identifier) of each of the chosen libraries in the expression datasheet of the UniGene relational database used by cDNA DGED. This table lists the transcripts in each library along with the number of ESTs in that library which represent each of those transcripts. The UniGene Cluster ID (the unique identifier) of each transcript, which is used to identify it in the expression datasheet, is used to search the transcript datasheet for the details of that transcript. These are reported in a transcript list. The presence or absence of each of the presented transcripts in each pool of libraries and the number of ESTs mapping onto that transcript in all pooled libraries are reported.

The new transcript parsing routines were tested by comparing a set of normal bone libraries with a set of libraries from cancerous bone. The chosen libraries were made from bulk bone tissue and had not been normalised during their preparation, thus matching as closely as possible the *in vivo* gene expression levels. The libraries used were the same as those presented by the online CGAP tools for bone tissue, in order to show that any differences in the gene results were due to differences in the transcript parsing algorithms and not due to differences in the library parsing algorithms. The "sequences" cut-off was set to zero to include all qualifying libraries of any size. All experiments were carried out by Andrew Milnthorpe.

3.3.3. Methods used to correct problem of CGAP statistics reporting different output values for different input display filter settings

Statistical methods implemented in CGAP DGED involve the calculation of a Benjamini Hochberg False Discovery Rate “Q” value for each gene (Benjamini and Hochberg, *1995*) from a probability “P” value calculated using the Fisher Exact Test (Daya, *2002*). Unlike CGAP, the Fisher Exact Test reported by Daya (Daya, 2002) was implemented alone using the following equation (Daya, 2002):

$$P = \frac{(a+b)!(a+c)!(b+d)!(c+d)!}{N!a!b!c!d!} \qquad (1)$$

Where P is the probability that the observed expression is not due to sampling error, a is the number of ESTs mapping onto the transcript in Pool A, b is the total number of ESTs in Pool A minus the number of ESTs for the transcript in Pool A, c is the number of ESTs mapping onto the transcript in Pool B, d is the total number of ESTs in Pool B minus the number of ESTs for the transcript in Pool B, and N is the total number of ESTs in both pools.

That equation was used to calculate the probability (reported in the manner of a Chi-squared “P” value (Yousef et al, *2004a*)) of the observed expression difference for each transcript being due to chance in a comparison of a group of bulk non-normalised normal bone libraries with a group of bulk non-normalised cancerous bone libraries (the relevant CGAP database was available for download on 9 March 2011). The “number of ESTs mapping onto transcripts in library” display cut-off was set to zero to ensure that the libraries studied were the same as those included in the equivalent CGAP search mentioned above. The bone libraries whose “unique tissue” field contained “bone” were the only ones reported, and the additional libraries the CGAP tools map onto bone tissue were not presented.

The factorials shown in Equation (1) were calculated using Stirling's Approximation, a method for efficient and highly accurate approximation of large factorials (Bracken, *2003*; Mortici, *2011*). The "P" values calculated using Equation (1) range between zero and one. The statistically significant values are those closest to zero, in the manner of a Chi-squared "P" value (Yousef et al, *2004a*). To check whether the "P" values calculated using Equation (1) change when the display cut-off value "F" is changed, the transcript parsing algorithm was run twice, once with the "F" display cut-off set to two and once with the "F" display cut-off set to three.

The difference between these statistics and those used by CGAP is that the CGAP cDNA DGED calculates a "Q" Benjamini-Hochberg False Discovery Rate value from the Fisher Exact "P" values as shown below (Benjamini and Hochberg, *1995*).

$$Q_1 = \frac{P_1}{n}, Q_2 = \frac{P_2}{n-1}, Q_3 = \frac{P_3}{n-2} \dots Q_n = \frac{P_n}{1} \qquad (2)$$

Where Q is the Benjamin-Hochberg False Discovery Rate for each gene, P is the Fisher Exact Probability value for each transcript, and n is the number of transcript expressed in either or both pools.

All experiments were undertaken by Andrew Milnthorpe.

3.3.4. Rectification of incorrectly reported library sizes and empty or missing database entries

For each library the number of ESTs representing all the genes in that library were routinely counted and reported by the transcript-parsing algorithm created using Excel to resemble that of CGAP's cDNA DGED. These values were compared to the ones reported by both CGAP tools' library lists and the number of ESTs mapping onto the transcripts in the same libraries as reported by cDNA DGED's transcript list for a comparison of bulk non-normalised libraries from non-cancerous bone (Pool A) with bulk non-normalised cancerous libraries from bone (Pool B). The respective library size display cut-off settings were set to zero so that all qualifying libraries were present for comparison. The data available for all searches was that available from CGAP on 15 November 2008.

The presentation and use of the number of ESTs mapping onto transcripts in each library instead of CGAP's "sequences" annotations was also used to solve the problem of the erroneous annotation of the library entitled "SARS-Cov infected lung tissue" as containing no ESTs.

The problem of the inclusion of empty libraries, unmapped transcripts and unlisted libraries in CGAP's database was corrected using Excel so that the empty libraries would not be reported in the library list of the Excel-based tool and so that Excel-based searches would take less time due to the unlisted libraries and unmapped genes not being present. To achieve this all such entries were deleted from the database files used by Excel. All procedures were undertaken by Andrew Milnthorpe.

3.4. Procedures undertaken to further improve the new tools

The new tools were programmed using Excel 2010's Visual Basic for Applications. The implementation of multiple pools in the library parsing algorithm and the inclusion of the formula in the algorithm based on CGA's cDNA xProfiler for filtering the transcript list were originally carried out previously for an MSc in Biological Sciences Research, also at Royal Holloway, which was completed in August 2008. Everything else presented in section 4.3, and all other work presented in this thesis, was undertaken since October 2008 for this PhD.

3.5. Methods used to create a procedure for the quality control of EST data

3.5.1. Selection of tissue specific transcripts

Candidate tissue-specific transcripts were selected based on a number of criteria. Firstly, CGAP database was manually searched for the highly abundant and tissue-specific transcripts (each of these unique cDNA species has an entry in CGAP's database, where it is annotated with a UniGene cluster ID as its unique identifier) defined by their EST counts, for all individual tissue types available using the cDNA DGED (National Cancer Institute, *n.d.c*) on 15 October 2009. Separate searches were conducted for "Normal" and "Cancer" histology for all tissue types. The minimum number of sequences per library was set at 10, the tissue preparation was set to "bulk" and the library protocol to "non-normalised" in all searches. The EST library group was set to "All", which included all CGAP, MGC, ORESTES and un-annotated libraries, the latter constituted the vast majority (~72%) of all the libraries used. The transcript lists were downloaded from CGAP and then searched for the transcripts with the high odds ratio (i.e. the normalised EST abundance in the selected tissue type divided by normalised abundance in all other libraries, typically above 10), which was also statistically significant (typically P < 0.05). Additional selection criteria were high relative EST expression levels in the targeted tissue (typically above 0.1% of all the ESTs counts) and low expression levels in the rest of tissue types (typically below 0.1% of all the ESTs counts). Where possible only ESTs identified in at least two libraries and counted at least three times in the tissue studied were selected. Up to thirty individual transcripts having the highest odds ratios and meeting all of the above criteria were selected from each of the individual tissue types. Where less than thirty or none were available, the selection criteria were relaxed and the transcripts which satisfied most of the search criteria were selected. These steps were carried out by Andrew Milnthorpe. All the transcripts were combined (totalling 2,295 from all tissues types) and the duplicates were removed, yielding 1,089 individual UniGene cluster IDs. This final step was performed by Mikhail Soloviev.

The second round of search for additional tissue markers was on the basis of their absolute abundance level only. For this EST counts for each of the 37,575 different transcripts from 155 non-normalised libraries from all non-cancerous tissue types were determined (the version of the database used for this was the one available on 2 July 2009). This first step was carried out by Andrew Milnthorpe. Expression thresholds were set at 1, 2, 4, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 63, 128 and 256, and subsets of genes based on their maximum expression level recorded across all these libraries (across all the tissues) were identified. Statistical relationships between these subsets and the previously constructed list of 1,089 genes were identified. The maximum positive correlation value of +0.48 was recorded for the subset of transcripts with the maximum EST counts of at least 18 in at least one of the 155 libraries tested. That subset contained 909 transcripts, of which 483 were already among the earlier found genes (the 1,089 set). The newly identified 426 transcripts were added to the original selection yielding 1,515 UniGene IDs. These procedures were undertaken by Mikhail Soloviev. Following a more recent update to the version of the database available on 18 August 2010, this list was reduced to 1,437 transcripts by excluding 78 transcripts (due to removal of these entries from the subsequent CGAP database release). Expression levels (EST counts) were then calculated for each of these 1,437 UniGene clusters for each of the main 26 human tissues matching tissue definitions of CGAP database, except for bone marrow, which was combined with bone, its parent tissue, and cerebellum and cerebrum, which were combined with brain, of which they are dependent tissues. However, some tissues, e.g. brain and nearby pituitary gland were not combined because despite being close together, therefore relevant EST libraries were assigned to different tissues. Also, having a few tissues with only limited (often single) suitable EST libraries would not allow the consistent analysis of all dependent tissues at many levels of resolution. To avoid such inconsistency, dependent tissues were not analysed. The produced expression matrix (1,437 transcripts x 26 tissues) was used for further optimisation. These procedures were performed by Andrew Milnthorpe.

3.5.2. Optimised selection of UniGene clusters to achieve improved tissue-specificity

The first round of optimisation aimed to reduce inter-tissue correlations. Tissue-specific expression "super-libraries" were created for 26 tissues from 126 bulk, non-normalised libraries made from normal tissues with at least 200 total EST counts, by combining EST counts for the selected set of 1,437 transcripts from the same tissue, where more than one EST library per tissue was available. Pearson product-moment correlation coefficients were calculated for all pair-wise combinations of such tissue specific expression data sets. The Pearson correlation is invariant to the changes in location and scale in the variables, the calculated correlation coefficients yield comparable values within the same scale interval (–1 to +1) for all tissues and libraries irrespective of their size, coverage, the number of ESTs or any preceding linear data transformations. This was done using Excel formulae (Zou et al, *2003*):

$$Correl(X,Y) = \frac{\sum(x-m)(y-n)}{\sum(x-m)^2 \sum(y-n)^2} \tag{3}$$

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.

Sum of squared errors was calculated from the deviations of the calculated correlation values from 1 as follows:

$$S = \sum(Correl - 1)^2 \tag{4}$$

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.

These values were used as a measure of discrepancy between the calculated correlation data and the model (no inter-tissue correlation of expression data for the selected markers). The change in the inter-tissue correlation values following removal of individual cluster expression data from the subset of 1,437 transcripts was then tested. The individual transcripts, removal of which had favourable effect on the reduction of the overall inter-tissue correlations were permanently removed and the iterative rounds of transcript removal were repeated. This experiment was undertaken by Andrew Milnthorpe. The best remaining transcripts (the last 505) were used for the second optimisation round, which was aimed to improve intra-tissue correlations. EST counts for each of the remaining 505 transcripts for each individual non-normalised library from normal (non-cancer) tissues (the same libraries as used before) were compared to each other. This time individual library expression data (not the super-libraries) were used to calculate sum of squared differences between the calculated correlation data and the model (high intra-tissue correlation of expression data for the tissues where two or more individual libraries were available). The change in intra-tissue correlation values was tested following removal of individual cluster expression data from the subset of 505 clusters. This step was performed by Andrew Milnthorpe. After repeating this procedure for all of the 505 remaining clusters, all the transcripts were scored and the ones, removal of which improved the correlations most were permanently removed. 244 transcripts were eventually selected as the generic EST expression tissue-specific dataset. This step was performed by Mikhail Soloviev. The reduced expression matrix (244 transcripts x 26 tissues, referred to as EST expression matrix) was used by Andrew Milnthorpe for all subsequent analyses.

4. Results

4.1 Problems found to be present in CGAP's tools during creation of the new tools

4.1.1. Errors in library search algorithm used by CGAP tools

In the attempt to replicate the cDNA xProfiler and cDNA DGED algorithms it was found that the core hierarchical classification system on which both the cDNA xProfiler and the cDNA DGED rely is not flawless. For example a search of cDNA database for the "ear" tissue resulted in over 100 libraries of which only six were actually generated from ear or related tissues (see Figure 4, database access date 13 May 2010).

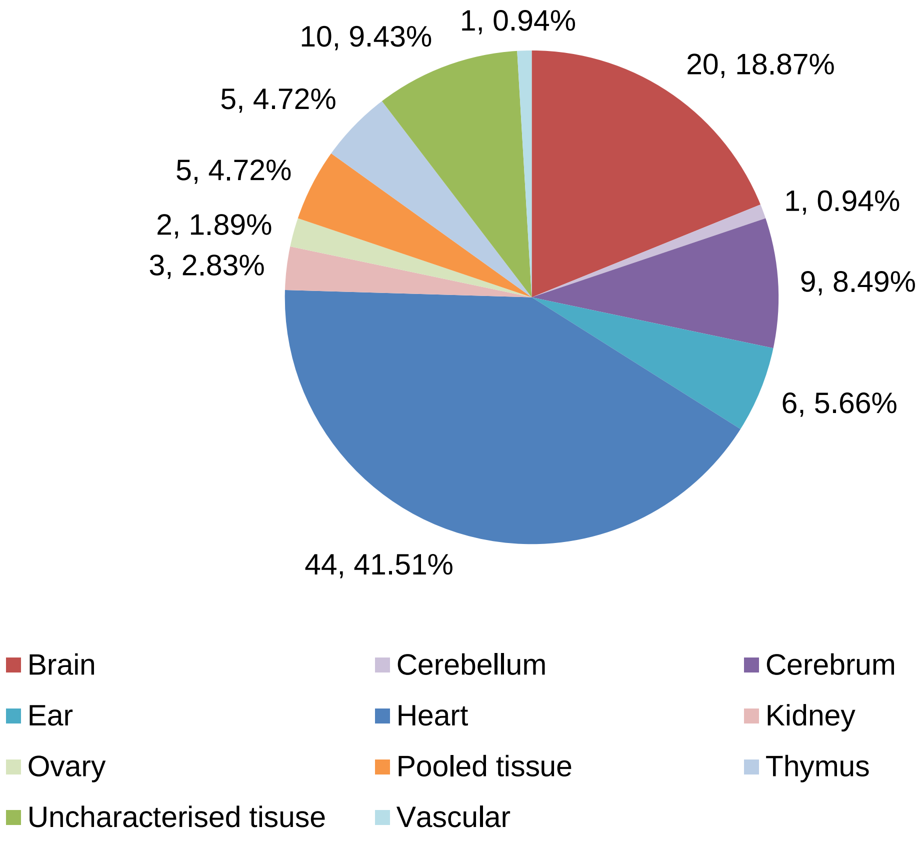


Figure 4. Tissue type origin of libraries reported for by CGAP tools after searching for "ear" tissue. All the libraries reported in this search (database access date 13 May 2010) were then manually checked for their "unique tissue" annotations and the percentage of the reported libraries which originate from all tissues were calculated.

The remaining ~94% of the libraries would be from irrelevant tissues such as the heart and brain. Other tissues also contained irrelevant libraries, e.g. brain library pool contained nine other unrelated tissues, or if eye libraries were selected, out of the 73 libraries, five were mixed tissues. Table 1 reports the correct/incorrect library inclusion rates for all other tissue types available and listed on the CGAP server. We have reason to believe that the CGAP library selection algorithms had serious flaws; these are detailed below and in Figure 4, using the "ear" library search as an example.

All libraries containing a text string "heart" in their "keywords" field (see Figure 1) seem to be included indicating that CGAP search for the correct string "ear" using any text matches, regardless of whether that string is part of a longer string such as "heart" or is a standalone word (as in ear tissue). This deficiency also brings into the results some libraries whose "unique tissue" field contains "brain", "cerebellum", "cerebrum", "thymus" or "vascular" because their "keywords" contain the phrase "heart disease" in their "keywords" field. This results in the inclusion of dependent or irrelevant tissues.

Other heart libraries which do not contain "heart" in their "keywords" field but contain "pericardium" instead are still included despite the fact that the letters "ear" do not appear in their "keywords" field. We found these to contain "heart" in their "unique tissue" field. Therefore CGAP must be searching "unique tissue" field similarly to the "keywords" fields and erroneously include partial text string matches.

A number of other libraries that contain "kidney" or "ovary" in their "unique tissue" field were included. We identified the reason for these - their "keywords" field contains text "clear cell renal carcinoma" or "clear cell ovarian tumor", where a search string "ear" can be found in "clear".

Table 1. Error rates for the CGAP library selection tools.

Tissue types available	Number of correctly reported libraries	Number of incorrectly reported libraries	Percentage of correctly reported libraries	Percentage of incorrectly reported libraries
Adipose	18	0	100.00	0.00
Adrenal cortex	3	0	100.00	0.00
Adrenal medulla	1	0	100.00	0.00
Bone	38	77	33.04	66.96
Bone marrow	54	2	96.43	3.57
Brain	543	476	53.29	46.71
Breast/Mammary Gland	1,137	7	99.39	0.61
Cartilage	24	0	100.00	0.00
Cerebellum	13	1	92.86	7.14
Cerebrum	428	2	99.53	0.47
Cervix	36	0	100.00	0.00
Colon	974	11	98.88	1.12
Ear	6	100	5.66	94.34
Embryonic tissue	43	466	8.45	91.55
Endocrine	26	437	5.62	94.38
Eye	44	28	61.11	38.89
Gastrointestinal tract	48	1,335	3.47	96.53
Genitourinary system[1]	0	0	0.00	0.00
Germ cell	6	46	11.54	88.46
Head and neck	4	951	0.42	99.58
Heart	44	41	51.76	48.24
Kidney	199	12	94.31	5.69
Limb	0	1	0.00	100.00
Liver	128	25	83.66	16.34
Lung	392	11	97.27	2.73
Lymph node	28	0	100.00	0.00
Lymphoreticular	16	97	14.16	85.84
Mammary gland/Breast	1,137	7	99.39	0.61
Muscle	36	104	25.71	74.29
Nervous	10	1,072	0.92	99.08
Oesophagus	21	1	95.45	4.55
Ovary	188	8	95.92	4.08
Pancreas	33	16	67.35	32.65
Pancreatic islet	14	0	100.00	0.00
Parathyroid	4	3	57.14	42.86
Peripheral nervous system	6	42	12.50	87.50
Pineal gland	7	1	87.50	12.50
Pituitary gland	14	1	93.33	6.67
Placenta	382	2	99.48	0.52
Pooled tissue[2]	Not available	Not available	Not available	Not available
Prostate	346	9	97.46	2.54
Retina	23	0	100.00	0.00
Salivary gland	10	6	62.50	37.50
Skin	89	11	89.00	11.00
Soft tissue	2	113	1.74	98.26
Spleen	22	6	78.57	21.43
Stem cell	29	57	33.72	66.28
Stomach	333	21	94.07	5.93
Synovium	20	0	100.00	0.00
Testis	222	3	98.67	1.33
Thymus	39	1	97.50	2.50
Thyroid	401	10	97.57	2.43
Uncharacterised tissue	1,959	5	99.75	0.25
Uterus	255	2	99.22	0.78
Vascular	34	3	91.89	8.11
White Blood Cells[1]	0	0	0.00	0.00

[1] No libraries were present in the database for these tissues

[2] Pooled tissue was not available in the CGAP tools, which listed these libraries under each of the tissues they were produced from.

Another keyword field error for "ear" search was found for libraries whose "unique tissue" field contains "uncharacterised tissue". Under their "keywords" field we found text "peripheral blood mononuclear cell" which is an incorrectly match for the search string "ear".

Finally, and unexpectedly, libraries created from mixed tissue samples (and therefore contain "pooled tissue" in their "unique tissue" field) were still included even if they did not contain the ear tissues. The reason is the same as described above - these libraries contained "heart" in their "keywords" field. In the same manner, as of 25 February 2012, libraries labelled with multiple tissue preparations, library protocols or tissue histologies will appear in the results if one their keywords matches the chosen library protocol or tissue preparation.

These errors in CGAP's library parsing algorithm were corrected once the findings were reported to NCBI. However, when last checked on 25 January 2012, “pooled tissue” libraries whose “keywords” field contains the required phrase were still being included with the tissue concerned. For example, a search of the database for “brain” tissue resulted in the inclusion of 13 libraries from mixed tissue samples along with the 984 libraries from brain and its dependent tissues which were correctly reported. Furthermore, one library from uncharacterised tissue was included. The inclusion of these mixed tissue libraries is erroneous because their gene expression levels are likely to be different from the gene expression levels in brain tissue.

We have also discovered that after correction by NCBI, while the only irrelevant libraries included are those from tissue samples labelled as mixed or uncharacterised, the inclusion of libraries from any dependent tissues is inconsistent. We have discovered that bone marrow is not included with bone even though bone marrow is a constituent of bone tissue, resulting in the exclusion of 58.7% of the libraries which quality for inclusion in the results for bone (Figure 5). This is despite the fact that other dependent tissues are correctly grouped with their parental tissue, for example cerebellum and cerebrum are correctly reported alongside brain, of which they are constituents.

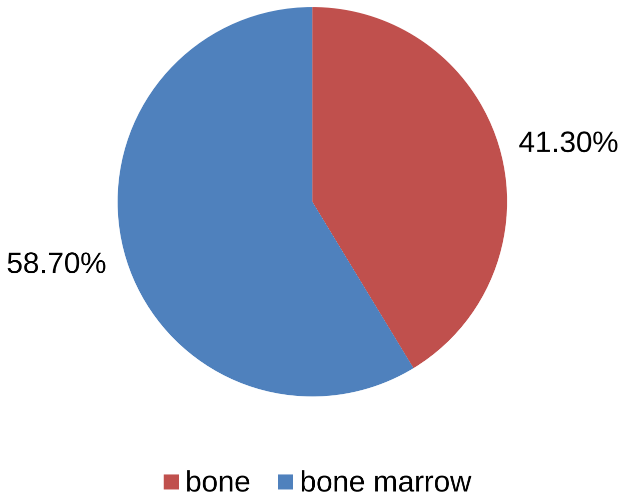


Figure 5. Percentage of libraries qualifying for inclusion under bone tissue by CGAP's tools which originate from bone or bone marrow. All the libraries reported in this search (database access date 26 January 2012) could be reported for inclusion when CGAP's tools are used to search for bone tissue, but only the bone libraries are presented.

4.1.2. Errors in CGAP’s gene search algorithm

We have also found that the cDNA xProfiler yields different lists of transcripts compared to cDNA DGED when all the same parameters are used. For example when normal adipose tissue was compared with cancerous adipose tissue using the cDNA xProfiler and the cDNA DGED, 1,359 transcripts were reported by both tools, 150 additional transcripts were reported by the cDNA xProfiler only and 273 by the cDNA DGED only, see Figure 6. This problem was not limited to this tissue alone. It was also discovered that the xProfiler reports additional transcripts to be present in its summary table of transcript results, compared to the transcript lists, see Table 2. As this table also shows, the total number of transcripts reported by cDNA DGED is greater than the number reported by the cDNA xProfiler gene lists and less than the number reported by the cDNA xProfiler results table. An attempt was made to analyse these discrepancies by looking into transcript annotations for the transcripts which were listed incorrectly, i.e. not listed by both tools.

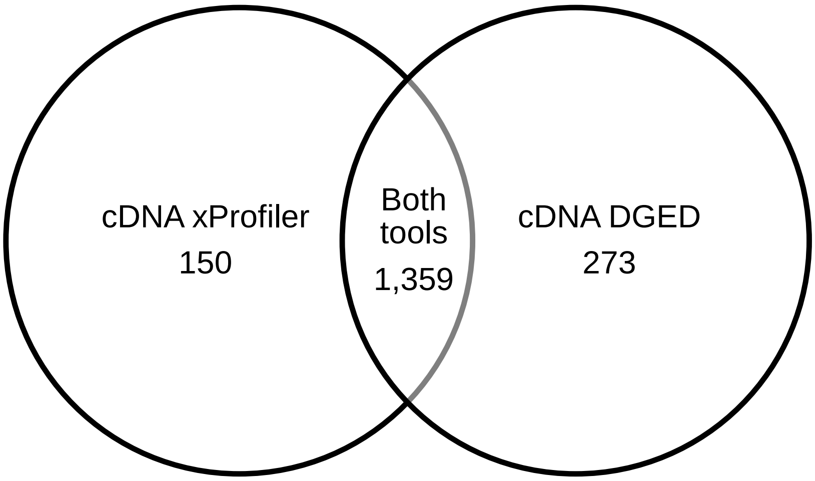


Figure 6. Differences in the number of transcripts reported by CGAP tools for an identical query. The total number of transcripts reported to be present when normal adipose libraries (in one pool) are compared with cancerous bone libraries (in the other pool) by the cDNA xProfiler's transcript lists (left circle) and the cDNA DGED (right circle). The overlap between the two circles represents the transcripts reported by both tools.

Table 2. Number of transcripts reported to be present in both pools when normal adipose libraries (in one pool) were compared with cancerous adipose tissues (in the other pool), by the cDNA xProfiler's transcripts lists and summary table of results, and by the cDNA DGED.

Tool and output method used	Number of transcripts reported
cDNA xProfiler results table	1,688
cDNA xProfiler gene lists	1,509
cDNA DGED	1,632

The correct list of transcripts to report for this particular comparison of normal adipose tissue with cancerous adipose tissue should include both the 1,359 transcripts reported by both tools and the 273 transcripts reported only by cDNA DGED. Both our transcript search routine and CGAP DGED appear to produce correct gene lists whilst the cDNA xProfiler missed 273 transcripts and also incorrectly selected 150 transcripts. This is because the cDNA DGED accesses the publically available CGAP relational database using the same method as our algorithm (see section 3.3.2 in the Methods section, entitled "Creation of novel transcript search algorithms to resolve issue of different transcript lists being reported by CGAP") to find transcripts which are represented by ESTs in one or more of the libraries presented in each pool, whilst the cDNA xProfiler accesses a single-file database which appears to miss 273 transcripts from the presented libraries and also incorrectly lists as being present in the chosen libraries the 150 additional transcripts. This discovery was confirmed by closer inspection of the results for the comparison of normal and cancerous bone libraries shown in Table 3, which revealed that the cDNA xProfiler reported 237 additional transcripts not reported by the cDNA DGED, while 707 of the genes reported by the cDNA DGED were omitted from the cDNA xProfiler’s results.

Table 3. Number of transcripts reported to be present in either or both pools when normal bone libraries (in one pool) are compared with cancerous bone libraries (in the other pool) by the cDNA xProfiler's transcripts lists and summary table of results, the cDNA DGED and using the new algorithm.

Tool and output method used	**Number of transcripts reported**
Reporting the presence or absence of each gene in a Boolean manner	
cDNA xProfiler results table	10,108
Our algorithm	9,996
Reporting the sequence odds ratio for each gene	
cDNA DGED	9,996
Our algorithm	9,996

These problems were corrected once the findings were reported to NCBI. However, subsequently we have discovered the inclusion of some libraries such as the brain library “NIH_MGC_181” in some searches where known, unknown, unique and non-unique transcripts are all reported (for example, normal brain vs. cancerous brain) will cause discrepancies in the distinguishing of transcripts as unique or non-unique. While some transcripts that are found in both pools (one in this case) are correctly reported as non-unique (found in at least one library outside the two pools) in the lists for each of the two pools (which list the transcripts in the relevant pool regardless of whether they are also found in the other pool), the transcripts concerned are incorrectly reported as unique (found only within the libraries included in the pools) in the lists of transcripts found in both pools. It was also discovered that the pooled tissue library “NIH_MGC_184” causes this problem when included with endocrine tissue, but does not cause this problem with brain.

As Table 4 shows for brain, this problem causes the number of unknown, non-unique transcripts reported as being present in both pools to not match what would be calculated from the figure reported for each pool on its own and the figures for each pool including transcripts common to both (this can be seen with the two columns representing unknown transcripts). Table 5 presents the figures that would be displayed if the results were reported correctly. In the internally available flat file database accessed by the cDNA xProfiler, the transcripts listed for each library are categorised according to whether they are unique to that library, so it appears that the cDNA xProfiler is incorrectly processing this information and misreporting some transcripts as unique when they are in fact non-unique.

Table 4. Numbers of known, unknown, unique and non-unique transcripts reported to be present in either or both pools when normal brain libraries (in pool A) are compared with cancerous brain libraries (in pool B) by CGAP's cDNA xProfiler.

Subset	Known unique transcripts	Unknown unique transcripts	Known non-unique transcripts	Unknown non-unique transcripts	Total
A	6	425	9,514	2,260	12,205
B	2	201	7,912	1,327	9,442
A or B	8	627	11,674	3,239	15,548
A and B	0	1	5,752	346	6,099
A minus B	6	425	3,762	1,913	6,106
B minus A	2	201	2,160	980	3,343

Table 5. Numbers of known, unknown, unique and non-unique transcripts which should be reported for either or both pools when CGAP's cDNA xProfiler is used to compare normal brain libraries (in pool A) with cancerous brain libraries (in pool B).

Subset	**Known unique transcripts**	**Unknown unique transcripts**	**Known non-unique transcripts**	**Unknown non-unique transcripts**	**Total**
A	6	425	9,514	2,260	12,205
B	2	201	7,912	1,327	9,442
A or B	8	626	11,674	3,240	15,548
A and B	0	0	5,752	347	6,099
A minus B	6	425	3,762	1,913	6,106
B minus A	2	201	2,160	980	3,343

This problem occurred for brain, endocrine, gastrointestinal tract, nervous, pancreas and thyroid. However, when last checked on 27 June 2012, this problem only occurred for nervous and pancreas. This suggests that while the cDNA xProfiler now reports the correct results for most libraries, some problems still remain.

4.1.3. CGAP errors in estimating significance of the predicted value of gene overexpression

The False Discovery “Q” values are used to indicate the likelihood of each expression result being a false discovery. These values are reported by the cDNA DGED and are based on Benjamini-Hochberg statistics and the Fisher Exact Test.

The value “Q” should indicate the reliability of the calculated odds ratio “F”. It was found that the “Q” value calculated by CGAP DGED would change depending on the user-selected display cut-off for the odds ratio “F”. This is certainly incorrect, as the probability of finding upregulation should not depend on whether a whole list of transcripts or part of that list is looked at. We believe that the probability of the result being correct should not depend on the display cut-off setting (all three are upregulated more than twofold). Table 6 illustrates this for three distinct entries. All the “Q” values are reported as very close to zero which indicates statistically very significant results. However, if the display cut-off is increased from two to three, the “Q” values for all three results change, indicating apparently increased statistical significance, which is not the case.

Table 6. Change in probability values reported by the cDNA DGED and the new algorithm when the display cut-off value "F" is changed are exemplified for three transcripts that are presented in the results list when normal bone libraries are compared with cancerous bone libraries.

UniGene Cluster ID	Name	Symbol	CGAP "Q" values[1]		The new "P" values[2]	
			"F" = 2	"F" = 3	"F" = 2	"F" = 3
280130	Ribosomal protein S24	RPS24	$1.10\text{x}10^{-10}$	$1.05\text{x}10^{-10}$	$1.23\text{x}10^{-13}$	$1.23\text{x}10^{-13}$
172928	Collagen, type I, alpha 1	COL1A1	$3.15\text{x}10^{-44}$	$2.98\text{x}10^{-44}$	$2.70\text{x}10^{-48}$	$2.70\text{x}10^{-48}$
436568	CD74 molecule, major histocompatibility complex, class II invariant chain	CD74	$9.37\text{x}10^{-20}$	$8.88\text{x}10^{-20}$	$6.73\text{x}10^{-9}$	$6.73\text{x}10^{-9}$

[1] Calculated using online tools from CGAP; calculations based on equations (1 and 2) in this report). The calculated "Q" value is close to zero (on a scale of zero to one) if the probability is high that the observed expression difference is genuinely greater than the user-specified "F" value, and is not a false discovery (Benjamini and Hochberg, *1995*).

[2] Calculated using equation (1) in this report. This produces a "P" value of between zero and one. The calculated “P” value is close to zero (on a scale of zero to one) if the probability is high of the observed expression difference being genuine and not due to sampling error, in the manner of a Chi-Squared “P” value (Daya, *2002*; Yousef et al, *2004*).

4.1.4. CGAP incorrectly calculates number of ESTs per library and includes empty and missing entries in their database

We believe that the number of ESTs in each library is an indirect indicator of the library quality because a library containing only a few ESTs is less likely to provide a representative picture of gene expression in the sample from which it was created than a library in which many ESTs map onto transcripts. As Table 7 shows, when normal adipose tissue was compared with cancerous adipose tissue using cDNA DGED, it was discovered that the sum of the number of ESTs in each library based on the annotations in the library database (in the "sequences" field, see Figure 1) was always greater than the number of ESTs cDNA DGED reported to be mapped onto all the transcripts in each pool (cDNA DGED uses the latter for its statistical tests, more on these below). This problem is not limited to adipose tissue and it affects the majority of the library database.

Table 7. Total number of ESTs reported for normal adipose tissue libraries and for cancerous adipose tissue by cDNA DGED library list and transcript list.

Results list	**ESTs in normal adipose libraries**	**ESTs in cancerous adipose libraries**
Library list	2,285	1,740
Transcript list	1,799	721

Despite reporting these findings to NCBI, we have discovered that this problem still exists. It has also been discovered that the library entitled “SARS-Cov infected lung tissue” has an incorrect “sequences” annotation of zero, suggesting it is an empty library (more on these below, see section entitled “CGAP incorrectly lists libraries which do not contain any transcript-mapping ESTs”). However, when the number of ESTs mapping onto the transcripts in each library was calculated (see section 4.2.4 below entitled “Solution to the problem with number of sequences per library and erroneous inclusion of empty of missing database entries”), it was discovered that this library contained 1,083 ESTs which mapped onto 1,023 transcripts. When last checked on 26 January 2012 the CGAP tools were found to misreport this library as containing no ESTs, potentially leading to its omission from a search, removing potentially useful gene expression data from the results and leading to the omission of potentially valid biomarkers or targets from further study or the study of false positive results in further investigations.

We have also discovered that the expression data file of the relational database used by cDNA DGED contains information for 37 libraries which are not listed in the library data file, in addition to the 8,378 libraries for which there is expression data (more on this below). As a result of this error the additional libraries will never be reported in any searches and the 3,846 expression data records they refer to (in which a total of 7,805 ESTs map onto 2,919 transcripts (all of which were also found in the 8,378 libraries in the library list) will never be of use to any study, leading to the omission of potentially valid biomarkers or targets from further investigation or the inclusion of false positive results in further experiments.

In addition to the 8,378 libraries which contain transcript-mapping ESTs (these are the 8,378 also presented in the expression database as mentioned above), the CGAP library database has been found to contain 529 libraries which do not contain any tags that map onto transcripts, of which 164 are annotated as SAGE (these are the only SAGE libraries in the database). Thus 5.94% of the libraries contain no expression information, as demonstrated in Figure 7. Of these, 228 libraries, including all those annotated as SAGE, do not contain any tags at all according to their "sequences" property. This is certainly incorrect because these libraries contain no expression data and their inclusion in a search will provide no contribution towards the results of any study.

Similarly the transcript database contains 3,202 transcripts (2.59% of the total) which do not map onto any ESTs in any libraries, as presented in Figure 8. This is an error because this database is not intended to be a transcript catalogue and is instead intended for use in gene expression profiling investigations. These transcripts will not appear in any results and will therefore not be further studied.

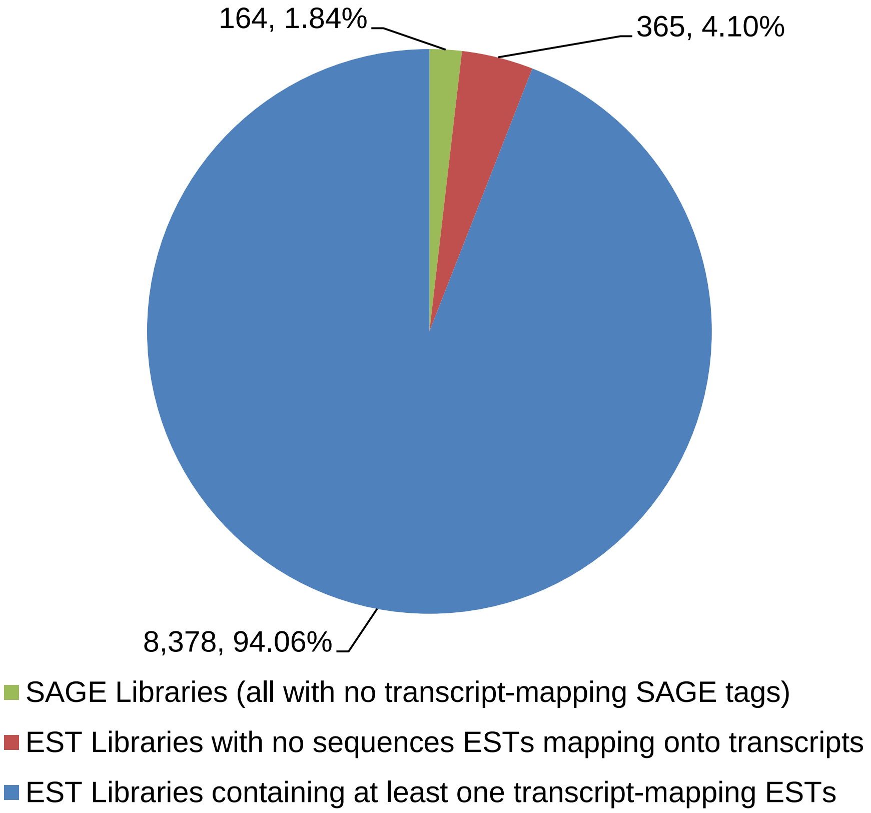


Figure 7. Percentage and number of libraries in CGAP's database which do not contain any expression information. All the libraries in CGAP's library database are presented in this pie chart, which shows the fraction of the libraries which do not contain any expression information (no ESTs or SAGE tags mapping onto any transcripts) and which therefore serve no purpose in any investigations.

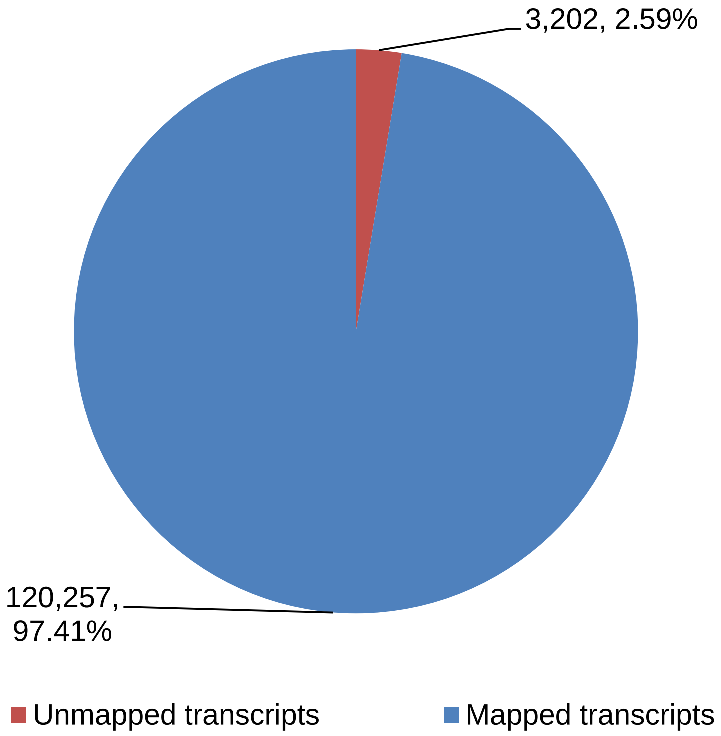


Figure 8. Percentage and number of mapped and unmapped transcripts in CGAP's database. All the genes in CGAP's gene database are shown in this pie chart, which shows the fraction which do not map onto any ESTs in any libraries (the unmapped fraction) and which therefore will never be reported in any study.

4.2. Solutions to CGAP errors implemented in new tools

In the attempt to identify the causes of errors and to further improve CGAP's cDNA xProfiler and cDNA DGED algorithms, the library and transcript parsing algorithms were studied by recreating them using Microsoft Excel. Revisions were then implemented to rectify the errors detailed above.

4.2.1. Solution to the errors in the library search algorithm

A library parsing algorithm has been designed to search only for the exact tissue type in each library's "unique tissue" field. For example, if ear is selected, we select libraries which only contain the exact string "ear" in this field and which do not have any other annotations in this field, resulting in the selection of libraries from the chosen tissue type without the inclusion of libraries from irrelevant tissues. Dependent tissues are also excluded, if the user chooses not to display them (see below).

If the required phrase is part of a longer phrase in a library's annotation (for example the phrase "bone" is part of the "bone marrow" annotation in the "unique" tissue field of a bone marrow library), the library with the longer phrase is ignored and not included in the results. In this example the selection would only contain bone libraries and not include bone marrow libraries. The recreated algorithm does this by searching for the required phrase as the only annotation in the "unique tissue" field, which does not contain any information other than the correct tissue type annotation for each library and therefore does not select dependent or irrelevant tissues. Furthermore, the recreated tool lists "pooled tissue" libraries as a separate user-selectable tissue type and does not include these libraries in the results for any other tissue, even if their "keywords" field contains the required phrase for the chosen tissue. This means that the results obtained will be based solely on libraries from the chosen tissue and will not be due to the inclusion of mixed tissue libraries. In the same way, libraries created from multiple tissue preparations or histologies or made using multiple library protocols are presented under separate user-selectable settings instead of being included when one of their keywords matches another setting. The recreated algorithm is also consistent about the inclusion of libraries from dependent tissues in the results, and allows the user control over whether such libraries are presented or not, something which is not possible with the CGAP algorithms. As a result, if the user chooses to display libraries from dependent tissues and chooses to search for bone tissue in at least one pool, both bone and bone marrow libraries will be reported in the pool(s) concerned, and therefore both sets of libraries presented in Figure 5 will be displayed.

4.2.2. Solution to the errors in the gene search algorithm

Two transcript parsing algorithms have been devised which search the CGAP relational database (as does the cDNA DGED) for transcripts contained within the presented libraries. One reports the expression information for each transcript as a Boolean type result identifying the presence or absence of a transcript in a pool (as does the cDNA xProfiler), while the other calculates an EST odds ratio for each individual transcript. Both report the results as a single list of all the transcript present in at least one pool along with the expression information. As Table 3 shows, these algorithms report the same transcript counts. The cDNA DGED reports the same transcript count for the same set of libraries whilst the cDNA xProfiler does not.

The recreated cDNA xProfiler-type algorithm avoids the problem of misreporting non-unique transcripts as unique by searching the publically available relational database, which does not present this information. However the availability of multiple pools by the recreated algorithm makes it possible to elucidate whether a transcript is unique to the libraries included using the recreated algorithm. This would be achieved by setting up an additional pool containing all libraries from the database except for those in the other pools.

4.2.3. Correct way to estimate the significance of the calculated gene expression

Though CGAP's statistics could not be completely implemented, the CGAP problem of apparent statistical significance changing after changing the display value "F" was solved by the implementation of only the Fisher Exact Test from which CGAP's "Q" values are calculated. As Table 6 shows, the resulting "P" values are the same regardless of the "F" display cut-off. That method calculates the probability of obtaining by chance the observed expression difference for a transcript between the two pools, using Equation (1) in section 3.3.3.

4.2.4. Solution to the problem with number of sequences per library and erroneous inclusion of empty or missing database entries

The problem of ESTs being reported by CGAP tools for each library which did not map onto the UniGene transcripts within that library was solved. The number of ESTs in each library which map onto the transcripts reported for that library was calculated, and this information was added to the library database for reporting in the list of libraries. The library parsing algorithm uses this calculated number instead of the "sequences" figure submitted by the library creator and included in the database by CGAP. This new approach reports the same total for the number of libraries in each pool as it does for the number of ESTs which map onto the transcripts in that pool (Table 8). This approach also means that the library entitled "SARS-Cov infected lung tissue" is correctly reported as containing 1,083 ESTs which map onto 1,023 transcripts, allowing it to be included in any study in which the sequences filter is used to only display high quality libraries with, for example, at least 1,000 transcript-mapping ESTs.

Table 8. Total number of ESTs reported for normal bone libraries and for cancerous bone libraries by the library list and transcript list produced by the CGAP tools and by the new routine.

Number of ESTs reported	**ESTs reported for normal bone libraries**	**ESTs reported for cancerous bone libraries**
Library list from CGAP tools	19,308	18,197
Transcript list from cDNA DGED	17,844	16,635
Library list from our new algorithm	17,844	16,635
Transcript list from our new algorithm	17,844	16,635

The problem of libraries being included in CGAP's expression database which are not included in the library database was solved by removing the expression data records concerned from the copy of the relational database used by the recreated tools. This has the added advantage of reducing the time required for a search. Therefore, expression data is only present in the copy used by the recreated algorithm for the libraries that are found in both files, with data for the missing libraries deleted.

The problem of the inclusion of libraries containing no transcript-mapping ESTs was solved by the removal of these libraries from the copy of the library database used by the recreated algorithms, leaving only the libraries which contain transcript-mapping ESTs. This also provided a decrease in search time. As a consequence of this, the only libraries found in the library database used by the recreated algorithm are the libraries presented in Figure 7 as containing transcript-mapping ESTs.

Similarly the transcripts which were found to not map onto any ESTs in any libraries were also removed because the purpose of the database is for gene expression profiling and not as a transcript catalogue. This also reduces the time needed for an investigation. Therefore the copy of the database used by the recreated algorithm contains only the transcripts which are shown by Figure 8 to be found in at least one library.

4.3. Further improvements made to new tools

In addition to the solutions to the errors in the online tools detailed above, the key improvement in the recreated tools is the provision of more than two pools. The tools are designed to accommodate any number of pools up to and including the maximum of 8,192 pools. Because each pool requires two columns this is the maximum capacity of an Excel 2007/2010 worksheet. This enables the finding of transcripts in a specific tissue which are not expressed in unrelated tissues, but which are expressed in tissues that are connected or located nearby, for example:

$$thyroid\ NOT \begin{pmatrix} parathyroid \\ OR \\ vascular \\ OR \\ PNS \\ OR \\ oesophagus \\ OR \\ muscle \end{pmatrix} \qquad (5)$$

Where PNS refers to the peripheral nervous system.

This six-pool search would present a list of transcripts which are expressed only in the thyroid and not in the parathyroid, oesophagus or muscles which are nearby. The reported transcripts would also not be expressed in the peripheral nervous system or the vasculature, which innervate and vascularise the thyroid, respectively.

This feature can also be used to find transcripts which are found only in a specific organ and not within any other organs in the same organ system, as well as being absent from unrelated tissues which are connected or proximal. For example:

$$\begin{pmatrix} kidney \\ OR \\ adrenal\ medulla \\ OR \\ adrenal\ cortex \end{pmatrix} NOT \begin{pmatrix} endocrine \\ OR \\ bladder \\ OR \\ PNS \\ OR \\ vascular \end{pmatrix} \qquad (6)$$

Where PNS refers to the peripheral nervous system.

This search requires 7 pools and would provide a list of transcripts which are only found in the kidney or its two outer layers (the adrenal medulla and adrenal cortex) and not in the blood vessels in this vascularised organ, or in the bladder (elsewhere in the urinary system) or other glands of the endocrine system.

The above queries could be extended to produce a list of transcripts which are only found in specific types of cancer that are associated with the organ of interested. For example:

$$papillary\ thyroid\ carcinoma\ NOT \begin{pmatrix} medullary\ thyroid\ carcinoma \\ OR \\ thyroid \\ NOT \\ \begin{pmatrix} parathyroid \\ OR \\ PNS \\ OR \\ oesphagus \\ OR \\ muscle \end{pmatrix} \end{pmatrix} \tag{7}$$

Where PNS refers to the peripheral nervous system.

This search, which would require 7 pools, would present a list of transcripts which are only expressed in papillary thyroid carcinoma and which are not reported for medullary thyroid carcinoma or in any of the normal tissues listed.

When choosing options at the beginning of a search the facility is available for libraries whose tissue preparation, library protocol and/or tissue histology is annotated as mixed or uncharacterised, to be studied separately from other libraries. There is a major need for this when choosing between library protocols. The multiple preparation, protocol and histology libraries are presented by the CGAP algorithms if one of their keywords matches another chosen setting. However, the uncharacterised libraries are only shown if no choice is made so that all libraries are included.

Similarly, libraries which are not annotated as CGAP, MGC or ORESTES or which are annotated as belonging to more than one of these groups can be selected separately using the recreated algorithm. This is unlike the CGAP tools where these libraries are only included if the list is not filtered according to library group or, in the case of the multiple group libraries, one of the groups they belong to is selected.

There is also the provision for the user to filter the libraries by developmental stage. If the aim is to study adult tissue only, this would enable, for example, libraries labelled as "embryo", "infant" or "pediatric" to be excluded. This would be useful because it is known that during these stages genes are expressed which are implicated in growth and development, some of which may also be expressed in cancer. Therefore, use of only adult libraries in this situation will not lead to these genes being erroneously excluded from further investigation as potential tumour markers.

The new tools incorporate a filter that allows the user to display only the libraries which are annotated with a specific gender. This is because the gene expression levels in cancer may be different in each gender and the user may desire to study cancer only in males or females.

There is also the facility for filtering the library list according to whether the library is annotated as being from a pregnant individual or not. This will enable the user to exclude libraries from individuals who are pregnant because the gene expression levels will be altered in this state.

If the user chooses to search for a library name in either pool, the CGAP tools report all libraries whose names contain the specified search string. For example, if "aorta endothelial cells" is entered, two libraries, one with that exact name and the other named "Aorta endothelial cells, TNF alpha-treated", would be presented. Because these libraries are expected to present different gene expression profiles, the new tools also allow the user to select whether an exact or partial match is required.

In addition to the availability of a filter based on the minimum library size, there is also the ability to exclude libraries whose maximum EST count exceeds a specified value. This would enable libraries of a similar size to be presented without the inclusion of, for example, one much larger library which may skew the results. If the two thresholds are set to values which are close to one another in magnitude, it is easier to determine the tissue coverage (and therefore the reliability) of the results for each pool from the number of libraries in that pool.

Using the algorithm based on CGAP's cDNA xProfiler tool, it is possible to filter the transcript list according to user-specified criteria. This can either be done manually using Excel's filter command to filter the columns, or, if no more than four pools are involved, can be done automatically by entering the pools into the provided formula. This formula is:

$$\left(P1\begin{pmatrix}AND\\OR\\NOT\end{pmatrix}P2\right)\begin{pmatrix}AND\\OR\\NOT\end{pmatrix}\left(P3\begin{pmatrix}AND\\OR\\NOT\end{pmatrix}P4\right) \tag{8}$$

Where P1, P2, P3 and P4 are four pools. The user can choose which pools these refer to.

When the above facility is used, a list of transcripts is produced matching the chosen settings with the pools shown in which each transcript is found. A results table is also produced similar to the one CGAP's cDNA xProfiler produces, but referring to the filtered list.

The algorithm designed to replicate CGAP's cDNA DGED also provides improvements on the functionality of that tool. Two collections of libraries are compared, each created using a formula identical to Equation (8). This enables up to eight pools to be compared at once, enabling queries such as those shown at the beginning of this section to be undertaken. The odds ratio is calculated between the two collections as follows:

$$\frac{\left(\frac{ESTs\ mapping\ onto\ transcript\ in\ Collection\ 1}{Total\ ESTs\ in\ Collection\ 1}\right)}{\left(\frac{ESTs\ mapping\ onto\ transcript\ in\ Collection\ 2}{Total\ ESTs\ in\ Collection\ 2}\right)} \tag{9}$$

In addition to the CGAP functionality, the Excel-based algorithm also presents the name of each library that a transcript is reported in and the percentage expression within that library. This allows the consistency of expression within those libraries to be verified. The tool also allows only known or unknown genes to be displayed, if that is required.

4.4. Creation of a procedure for the quality control of gene expression data

In addition to the improvements made to the search algorithms and the correction of errors therein, experiments were also carried out with a view to providing a means to quality control the expression data itself without reference to any other source of information. It has been hypothesised that if a list of tissue-specific markers could be produced the expression levels of the transcripts in that list in libraries of known identity could be used as a quality control method by enabling the elucidation of the true tissue type of a library.

4.4.1. Tissue specific transcripts and EST expression matrix

It was hypothesised that to be suitable for the role of universal tissue specific markers, the transcripts should be (i) highly abundant in their target tissues relative to all the other tissues and (ii) should be abundant in absolute terms in 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 criteria described in the Methods section, from each of the individual tissue types. Of the 1,089 transcripts identified, 1,044 were present in normal (non-cancer) tissues (although non-exclusively) and 479 originated from more than one tissue type. Whilst that was allowed, a further optimisation of the selected subset was necessary.

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 for all of the 37,575 different transcripts from 155 non-normalised libraries from all non-cancerous tissue types. Following the procedures outlined in the Methods section the list of potential tissue markers was expanded to include 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, an attempt was made 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 (3)). A 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 (4)).

Individual genes were then removed and the correlation values and the equation (2) total were recalculated. The gene, whose removal 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 (this was discovered by expanding Figure 9). The remaining 505 genes included the set of high-quality tissue-specific markers and these were retained. A similar optimisation was then repeated for the remaining 505 genes but this time the aim was 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. 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.

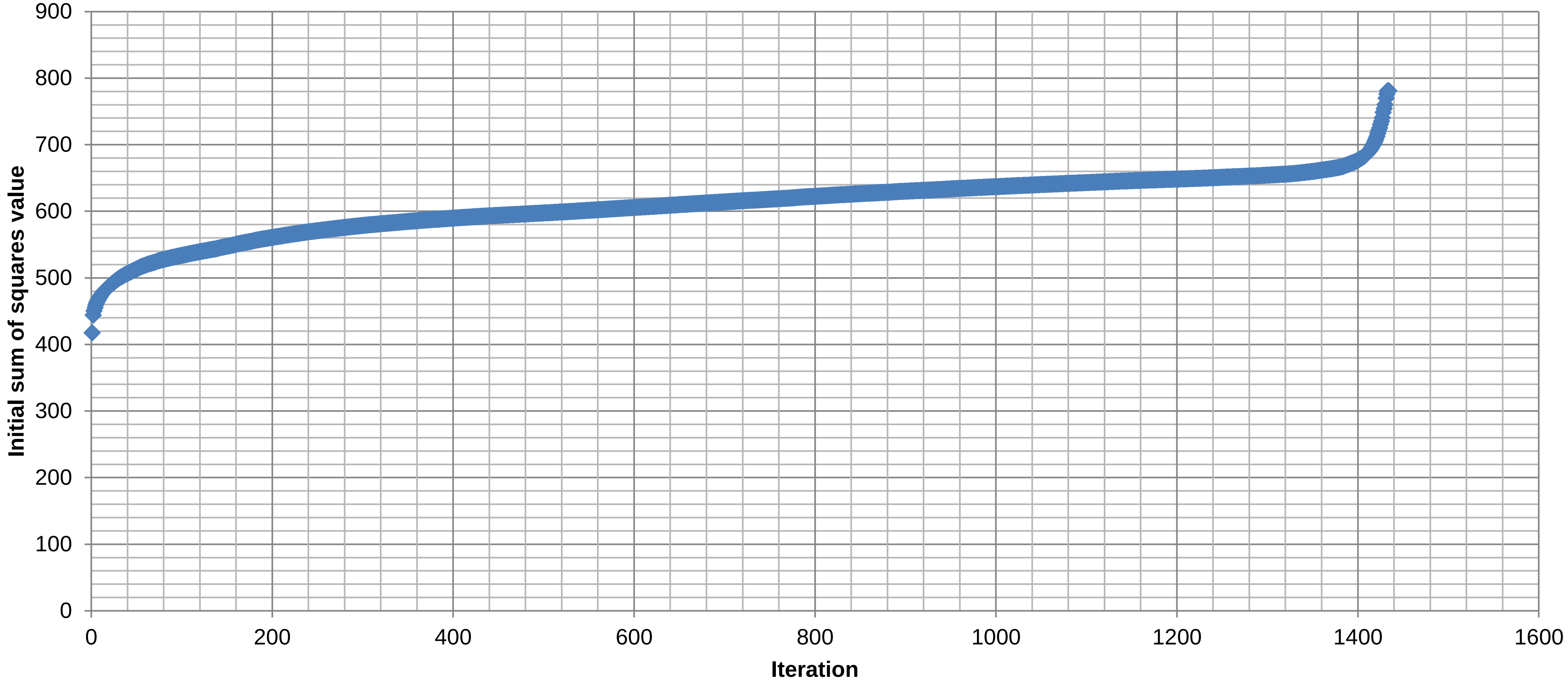


Figure 9. Inter-tissue correlation during optimization of marker list for genes with improved tissue specificity. The increase in the sum of squares value (which corresponds to a decrease in the inter-tissue correlation) (y-axis) is plotted against the gene removal iteration (x-axis), after each of which the gene was permanently removed whose temporary removal had produced the greatest improvement in the tissue-specificity of the gene list.

4.4.2. Inter-tissue correlations and intra-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 fell into three main categories. The first group contained groups of 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 10 summarises the results for five such representative tissues where correlation levels clearly confirm the identity of each of the individual EST libraries. The second group contained tissues for which only one or two non-normalised bulk EST libraries were available. In the former case (one library per tissue) positive correlations of +1 were expected, because for these tissues only the EST matrix was based on those expression data. Nevertheless, no other tissues having positive correlation above ~0.2 were identified, confirming the absence of cross-tissue correlations for the EST matrix entries (Figure 11). The third group included tissues with some degree of multiple tissue positive correlations. These were brain tissue libraries which shown partial positive correlation with peripheral nervous system EST libraries, the peripheral nervous system libraries showed a degree of positive correlation with brain derived libraries, heart libraries showed weak positive correlation with muscle libraries and muscle libraries shown some positive correlation with heart libraries (Figure 12). Some positive correlation between these groups of libraries is likely because of the very similar nature of those tissues. But this was unexpected, because one of the original optimisation rounds specifically aimed to exclude such correlation where possible. However such partial positive correlation proves that the EST matrix is also capable of identifying more distant but related tissue types. One particular brain library out of the 13 brain libraries tested (NIH_MGC_181) showed unexpectedly high correlation with pituitary gland. This was much stronger than with the brain expression pattern from the EST expression matrix – the supposed origin of this particular library (Figure 12A). A plausible explanation might be an unintentional inclusion of pituitary gland tissue with the brain tissues for the original library preparation; this is likely due to the close proximity of pituitary gland which is located at the base of the midbrain. Despite the inclusion of this library in the original selection and into the subsequent optimisation steps as "brain" derived, the EST matrix was still able to pick this inaccurately annotated library, thus confirming the robustness of the approach to cluster selection for the EST expression matrix. Using just tissue-specificity (the traditional approach which relies on comparing gene expression between tissues) would have counted such pituitary library as brain derived, which would have influenced the selection of "tissue specific" genes, for which incorrect tissue specificity would have been assigned.

Figure 13A summarises the correlation ranges for all the expected matching tissues, including the tissues detailed in Figures 10, 11 and 12. The first and third quartiles for all the positively correlated libraries from all tissues studied are 0.4 and 0.8 respectively (full range 0.2 to 1). The negative inter-tissue correlations are shown in Figure 13B. These values are based on all of the non-matching inter-tissue correlations, with first and third quartile values of –0.04 and –0.02 respectively. The expected inter-tissue correlations (such as brain with peripheral nervous system and heart with muscle) are summarised in Figure 13C. These correlations values are lower than the tissue-specific intra-tissue matches (Figure 13A), but notably higher than correlations between any non-matching tissues (Figure 13B), with the first and third quartiles at ~0 and +0.14

respectively. Figure 13D compares all three correlations ranges for all cases (positive tissue matches, related tissues, and non-matching tissues).

In order to systematically investigate the robustness of this approach to quality control, modelled EST data were used 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 the real EST expression data were gradually scaled down and all of the generated model libraries were compared with the original libraries including from all other tissues by calculating the correlation values for the 244 UniGene IDs from the optimised matrix set (Figure 14 – Figure 18). Virtually every library continued to correlate well with the tissue of origin until the very last EST mapping onto one of the transcripts in the matrix is removed. Furthermore, 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. Remarkably, some of the libraries remain identifiable when the libraries are scaled down to 0.5% of their original size. This shows that the EST matrix can be used to characterise small libraries despite the fact that they are less representative due to the reduced likelihood of rare transcripts being included in such libraries.

These results are summarised in Table 9 – Table 13, 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.

As Figure 19 shows, a clear positive correlation exists for all five tissues between the size of the library and the quality of the match, with positive correlations ranging between 0.22 and 0.96. This shows that library size has an impact on the quality of a library because a small library is less likely to be representative of expression in the original sample. Despite this, the model scaled down libraries all presented a good tissue match, albeit with reduced correlation values. This is summarised in Figure 20, where there is still positive correlation between size and the quality of the match for the scaled down libraries as well as for the original libraries, even though the values are reduced to between 0.15 and 0.88.

Therefore, the quality of tissue typing does not change dramatically and for lung the correlations 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.

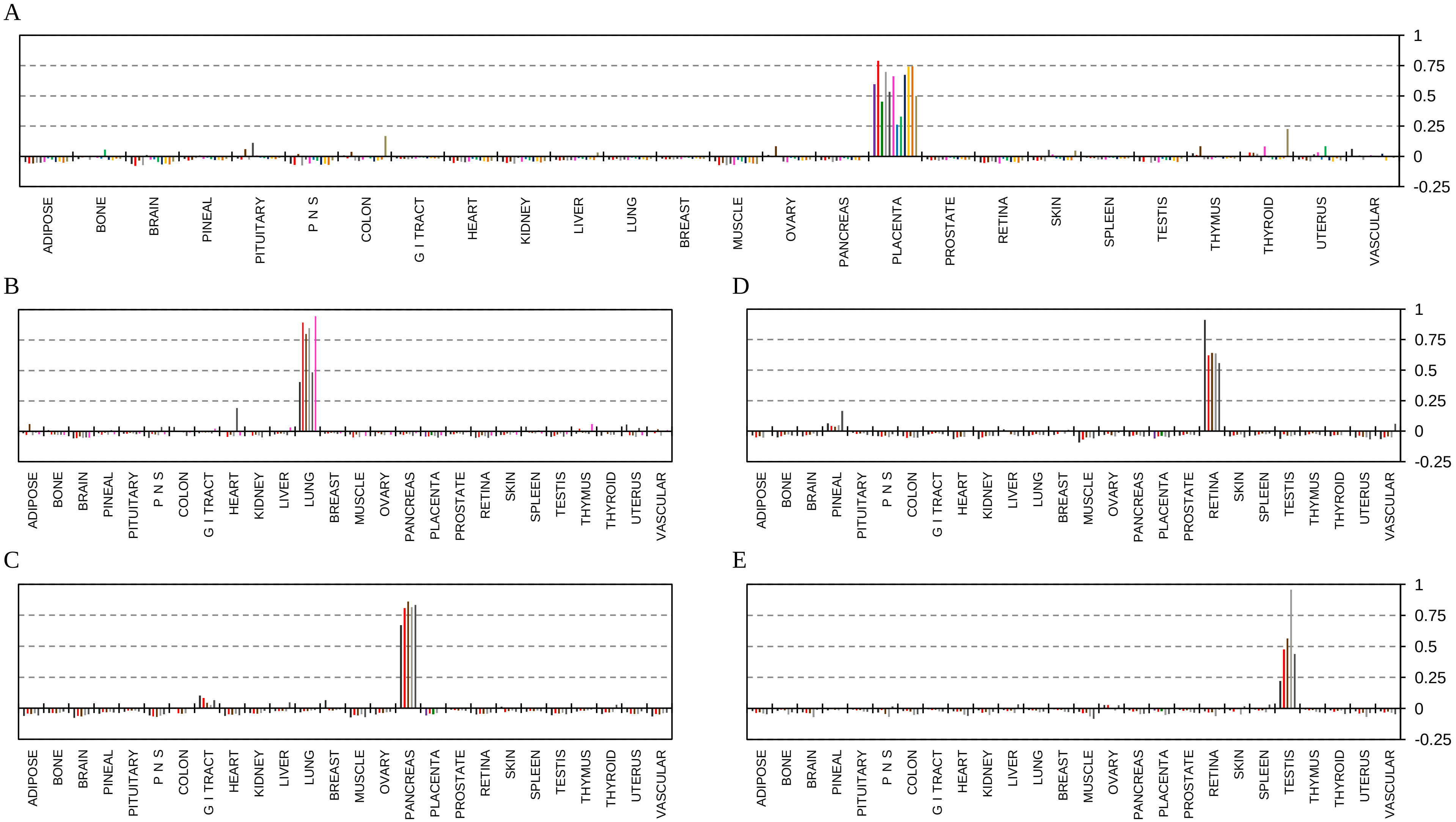

A
B
C
D
E
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 10. 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.

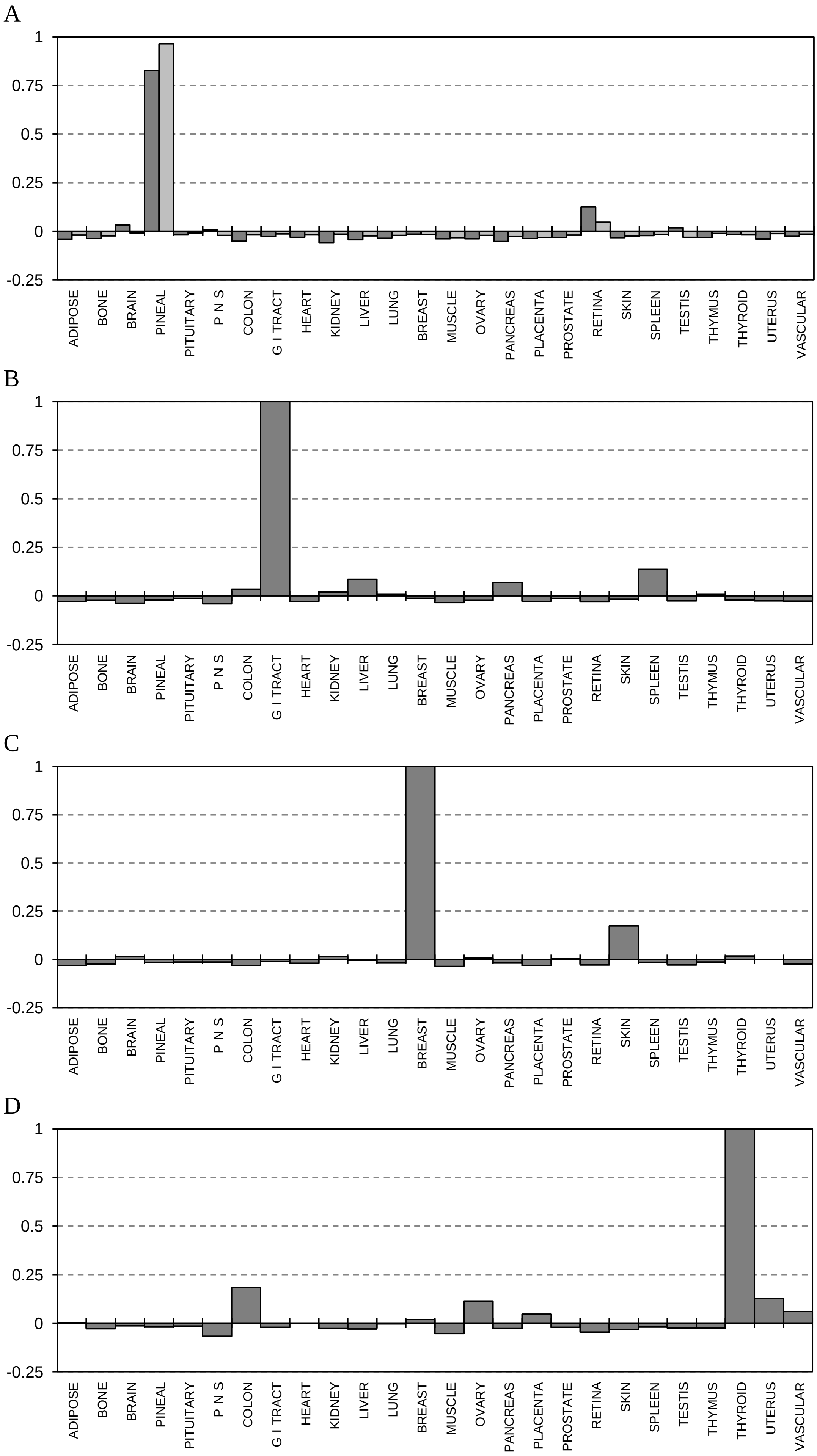

A
B
C
D
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 11. Correlation of the EST expression matrix with tissues with one or two libraries were available. Pearson product-moment correlation coefficients (vertical axes) calculated for each of the individual EST libraries and the EST expression matrix). **A:** "Soares_pineal_gland_N3HPG" library (dark bars), "Pineal gland II" (lighter bars). **B:** "Small intestine I" EST library. **C:** "NCI_CGAP_Br7" library from mammary gland. **D:** "Thyroid" EST library.

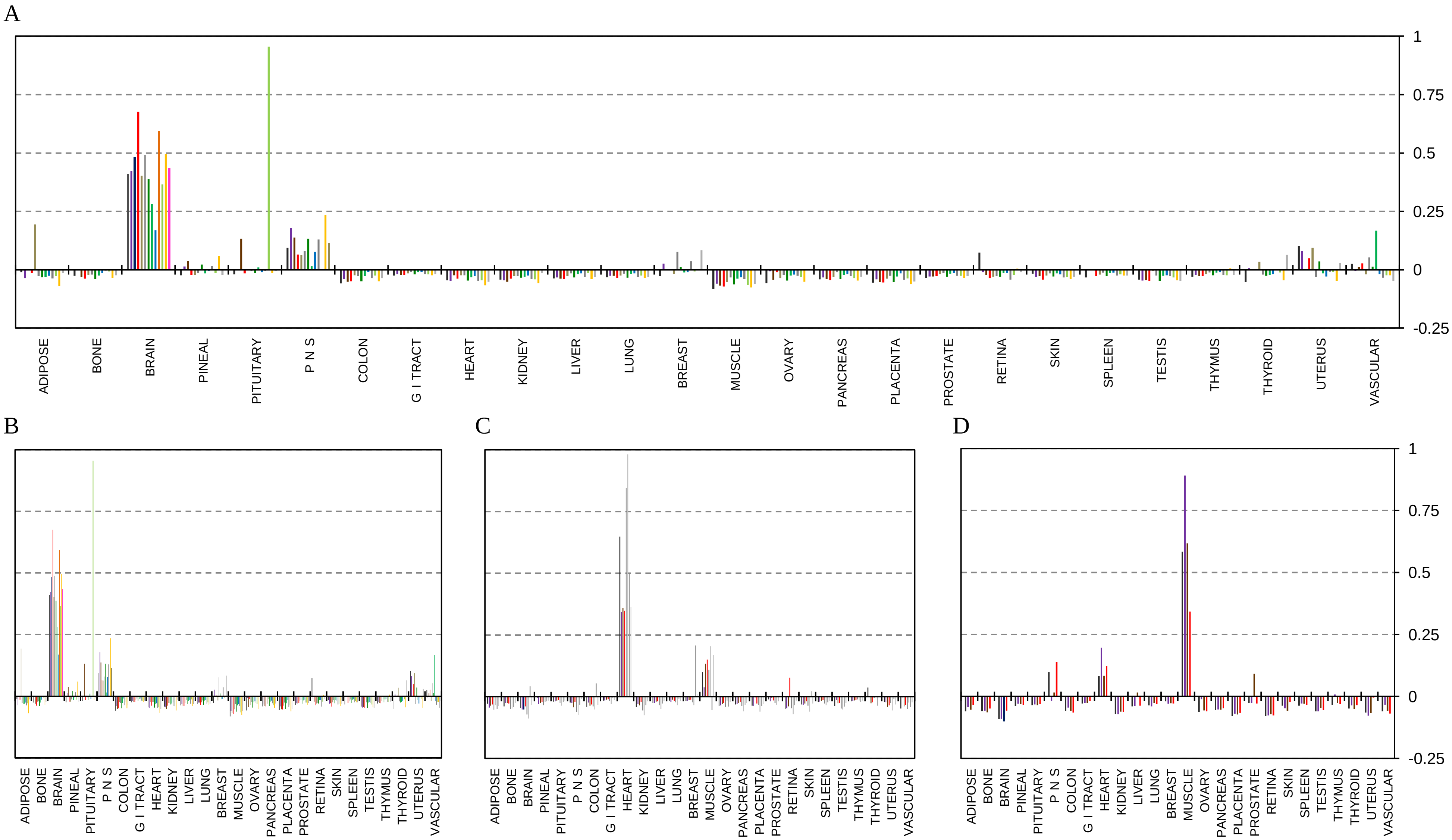

A
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
B
C
D

Figure 12. Correlation of the EST expression matrix with individual EST libraries from related tissues. Pearson product-moment correlation coefficients (vertical axes) calculated for each of the individual EST libraries and the EST expression matrix. **A:** Brain EST libraries, these include one cerebellum and one cerebrum EST libraries. Assumed mixed tissue brain library showing positive correlation with pituitary gland is "NIH_MGC_181". **B:** Peripheral nervous system libraries showing a degree of positive correlation with brain libraries. **C:** Heart libraries showing a degree of positive correlation with muscle libraries. **D:** Muscle libraries showing a degree of positive correlation with heart libraries.

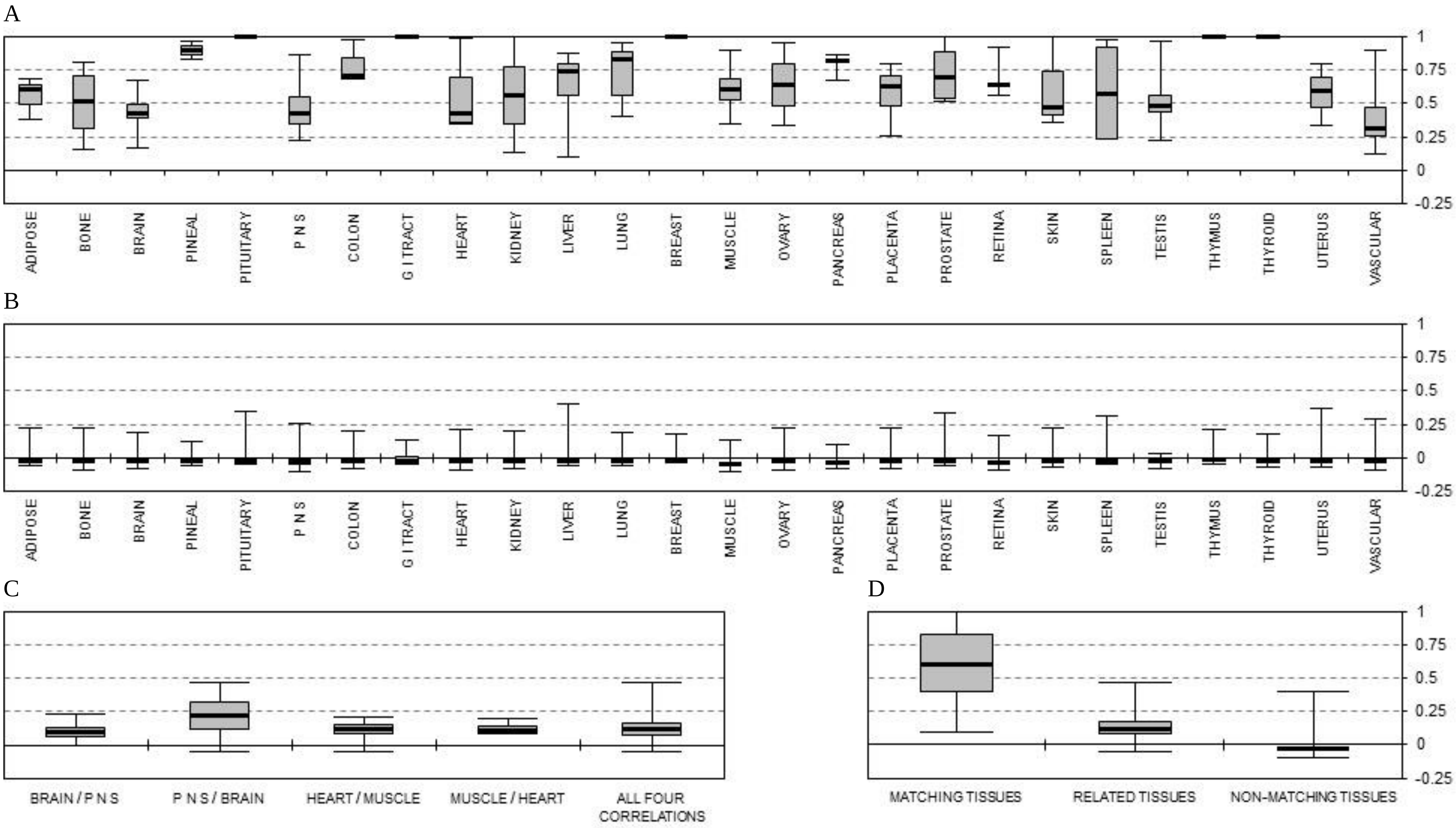

A
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
B
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
C
BRAIN / P N S
P N S / BRAIN
HEART / MUSCLE
MUSCLE / HEART
ALL FOUR CORRELATIONS
D
1
0.75
0.5
0.25
0
-0.25
MATCHING TISSUES
RELATED TISSUES
NON-MATCHING TISSUES

Figure 13. Intra-tissue and inter-tissue correlations. Correlation coefficients calculated for all of the 113 EST libraries against the EST expression matrix. The data also include the tissues detailed previously in Figures 10 – 12. **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. **C:** Correlations for all expected related tissues, e.g. all individual "Brain" libraries available vs. the "Peripheral nervous system" expression matrix, etc. **D:** All expected positive correlations from all matching libraries as in panel A (left box plot). Correlations from all related tissues as in panel B (middle box plot). All expected correlations from non-matching tissues, as in panel C (right). In all panels the boxes are drawn from the first to third quartiles. Plots also show minimum value, median (thick line) and the maximum correlation values recorded.

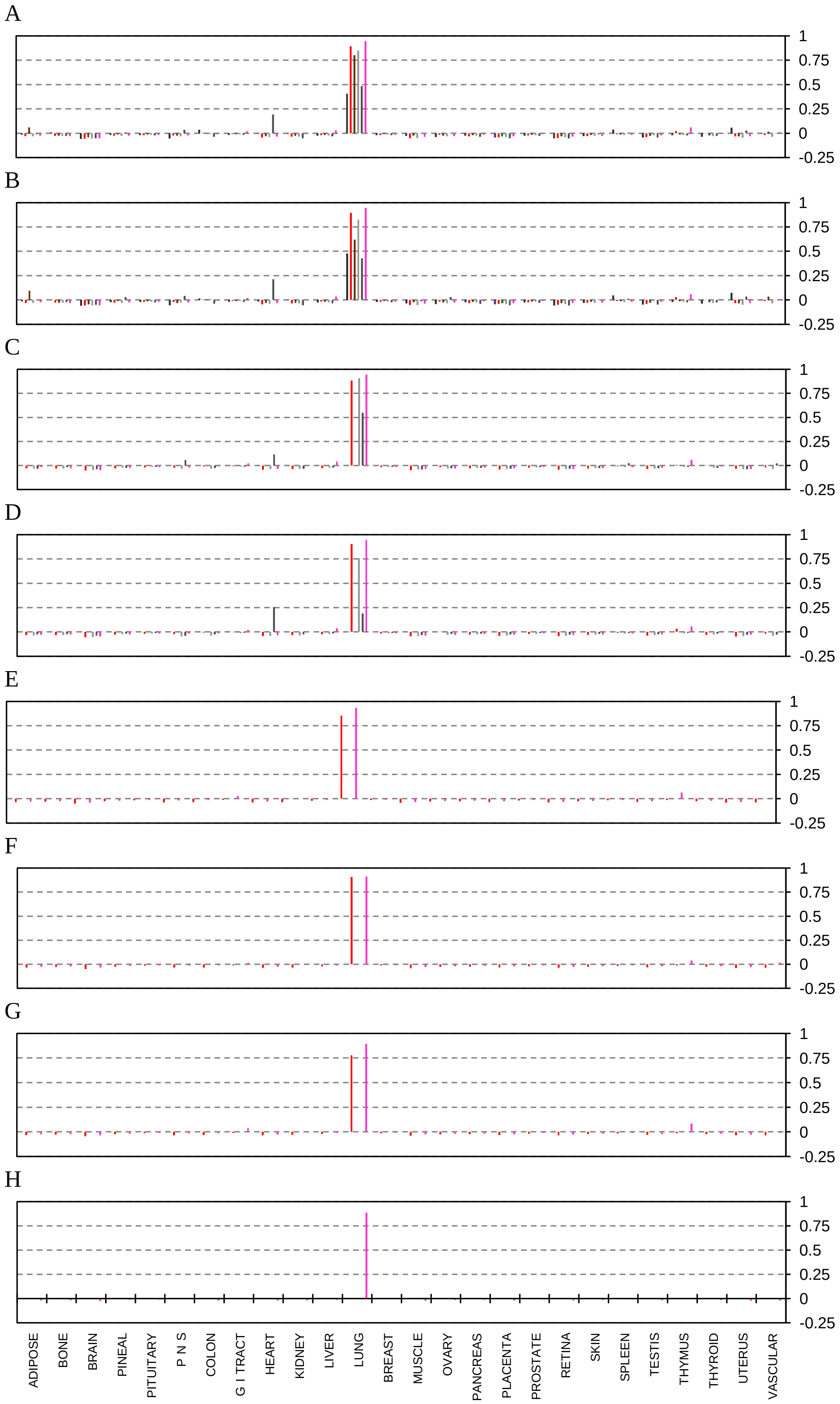

A
B
C
D
E
F
G
H
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 14. 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 9.

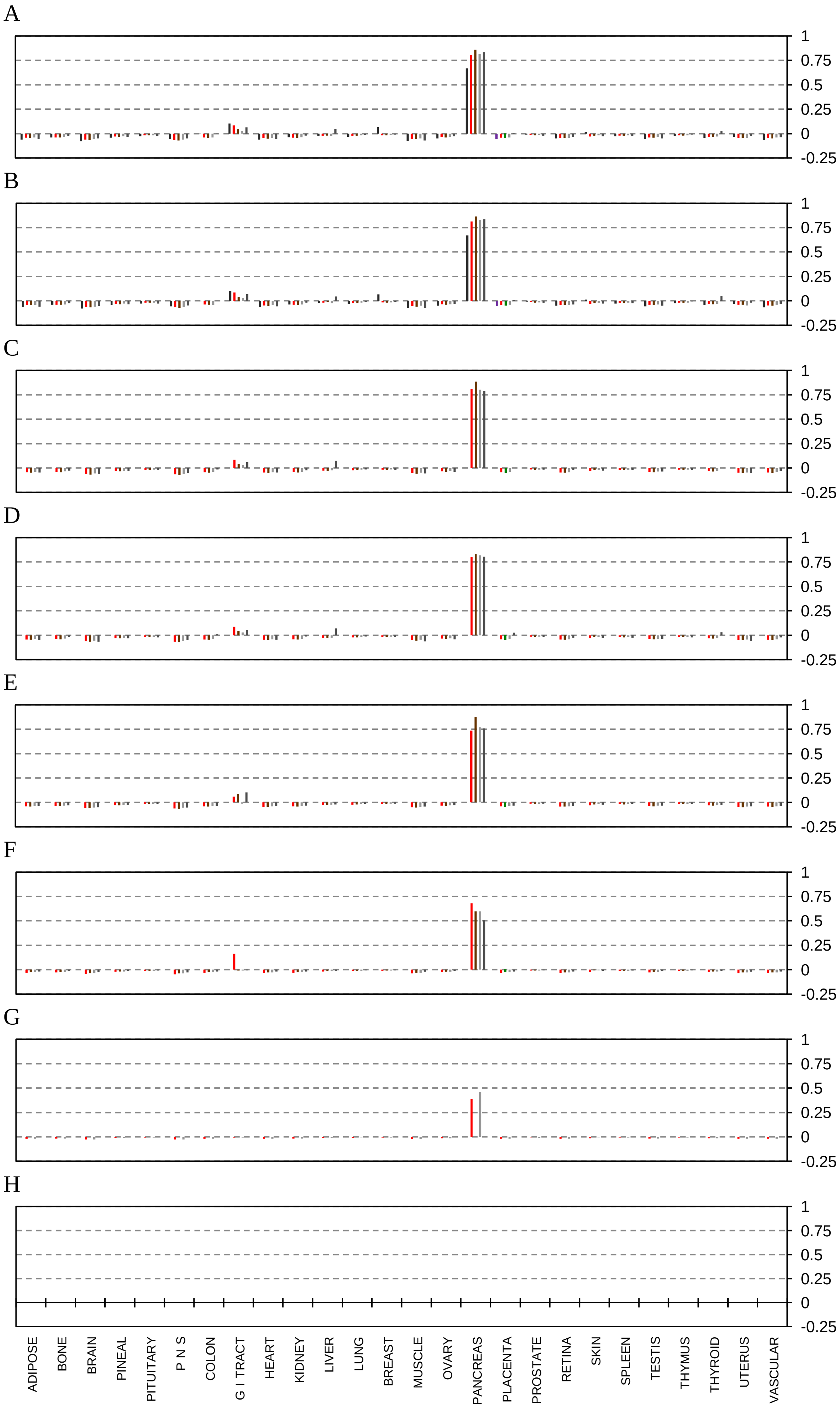

A
B
C
D
E
F
G
H
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 15. 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 10.

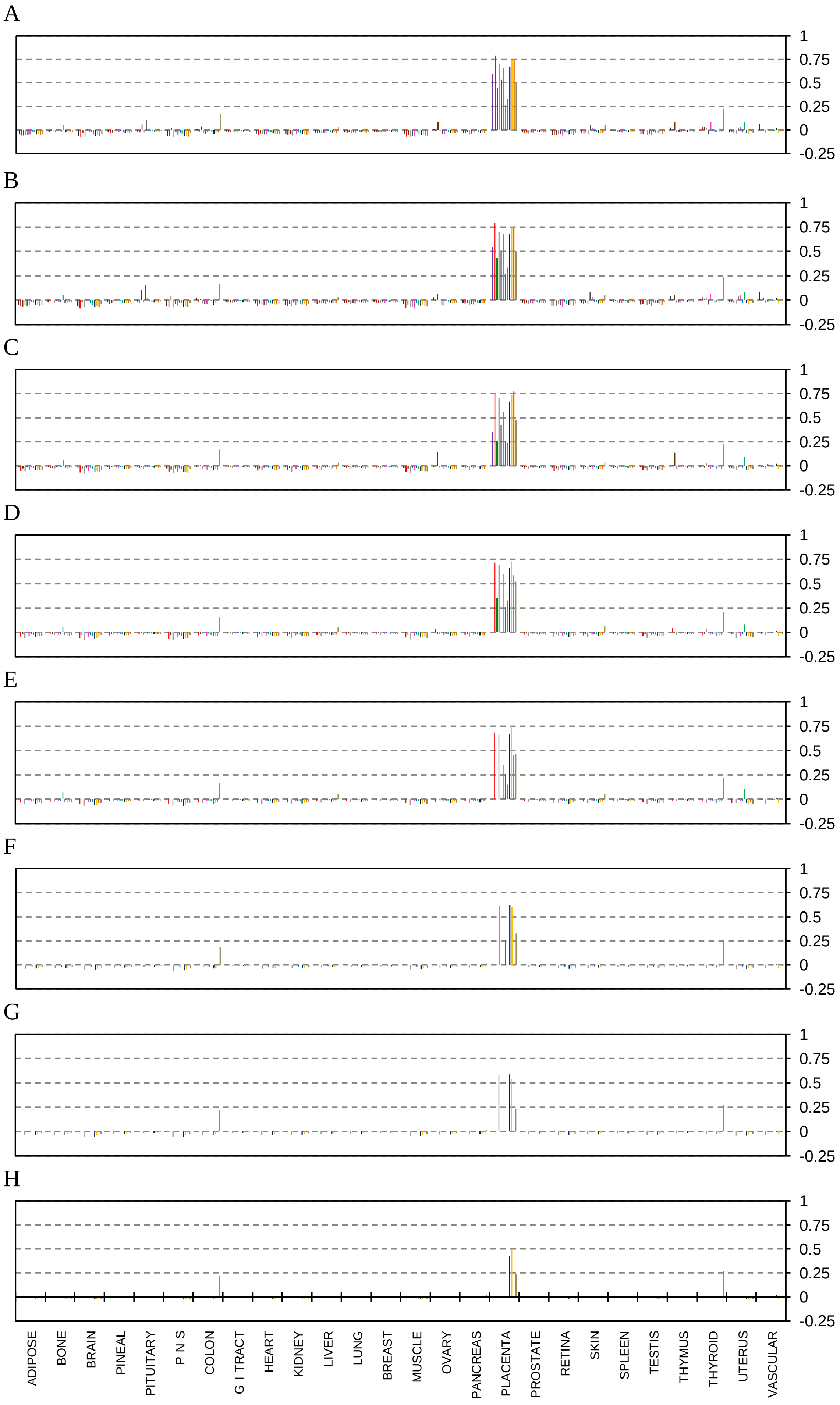

A
B
C
D
E
F
G
H
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 16. 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 11.

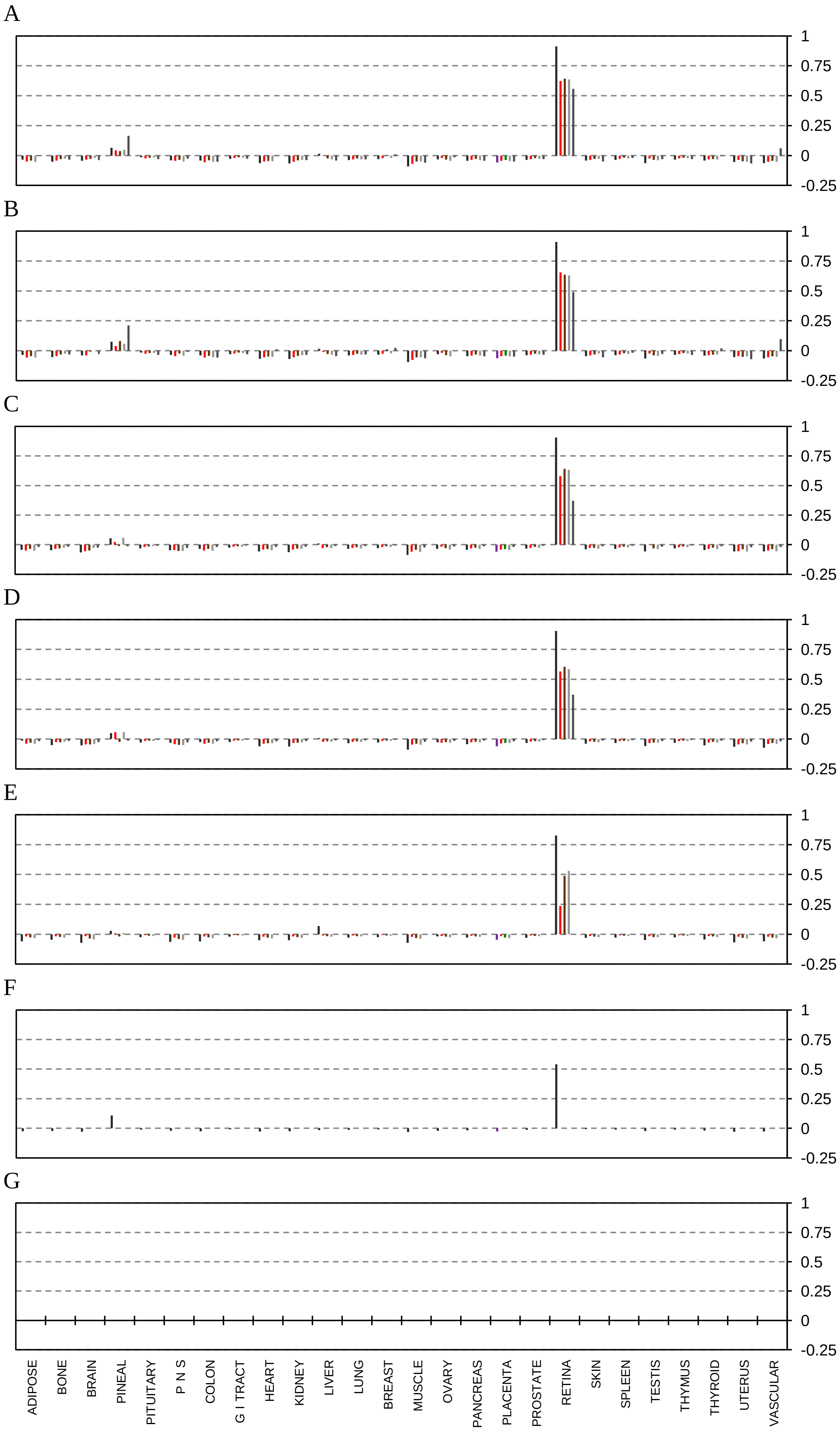

A
B
C
D
E
F
G
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 17. 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 12.

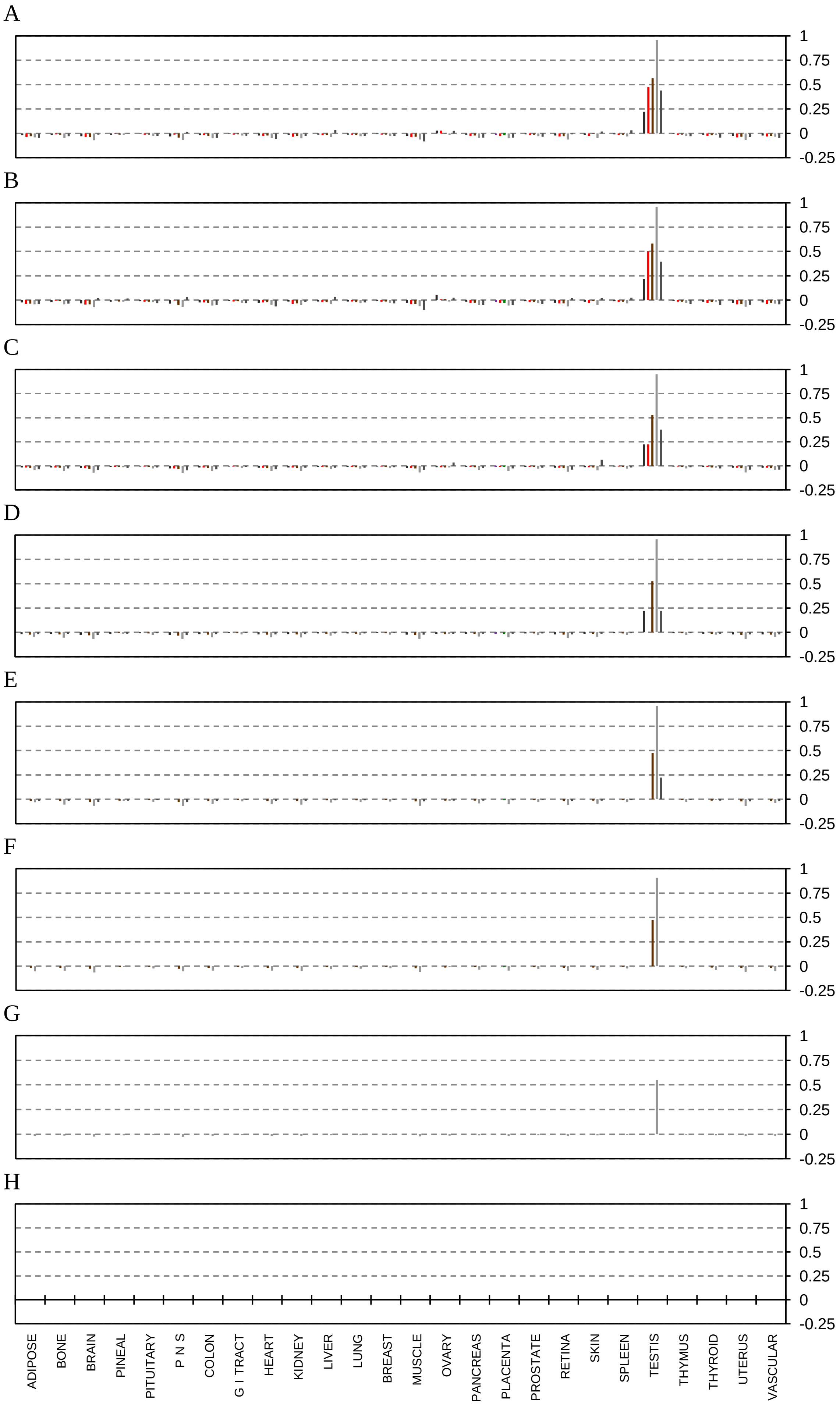

A
B
C
D
E
F
G
H
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 18. 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 13.

Table 9. 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

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

[2] Using the matrices and as described at the beginning of this section.

[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 10. 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

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

[2] Using the matrices and as described at the beginning of this section.

[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 11. 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_8to9weeks_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

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

[2] Using the matrices and as described at the beginning of this section.

[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 12. 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 at the beginning of this section.

[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 13. 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 at the beginning of this section.

[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.

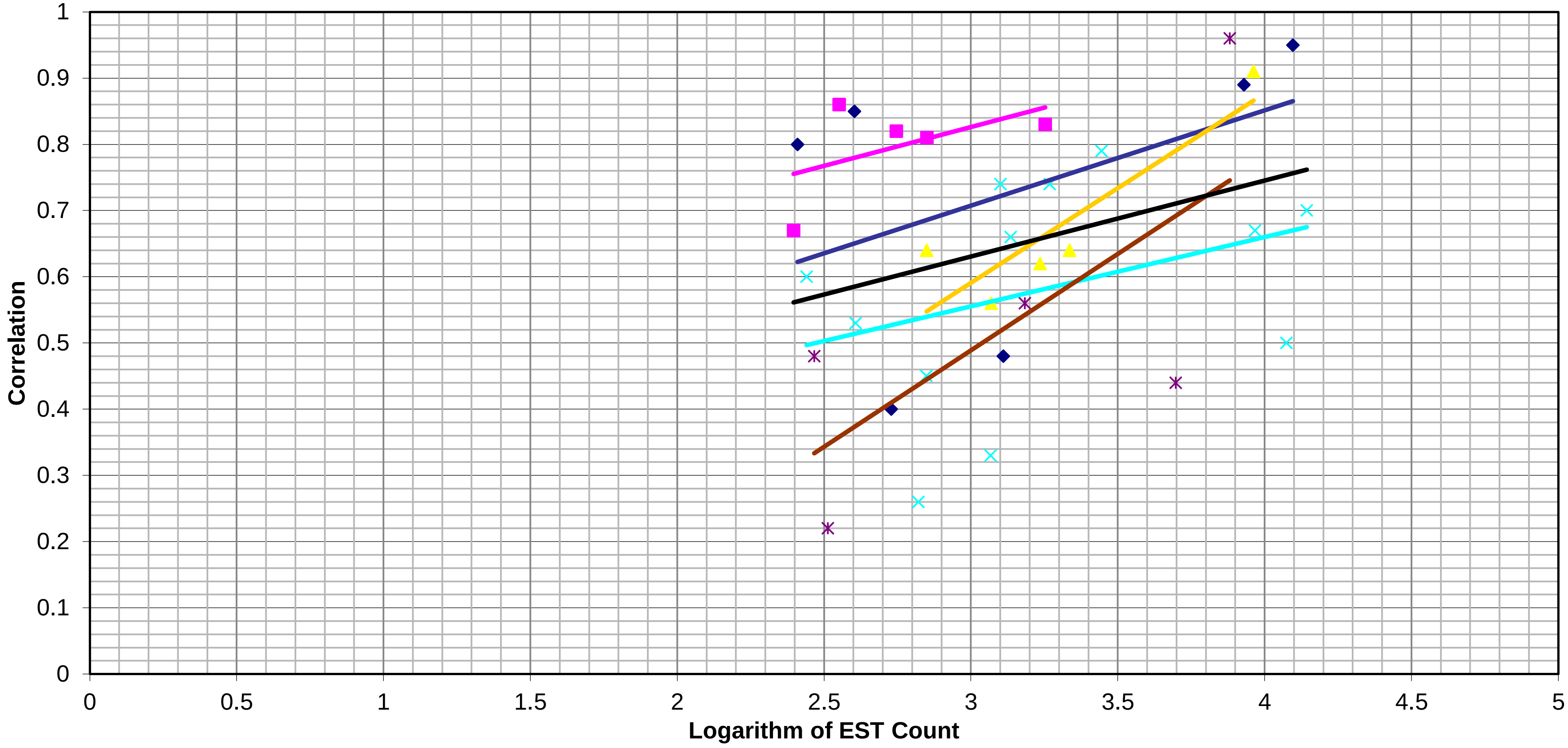


Figure 19. Pearson correlation values of original EST libraries (y-axis) vs. EST count (x-axis). The black trendline is fitted to all of the data points shown (all tissues), while the other trendlines are fitted to the individual tissues: lung (dark blue), pancreas (pink), placenta (light blue), retina (yellow) and testis (brown).

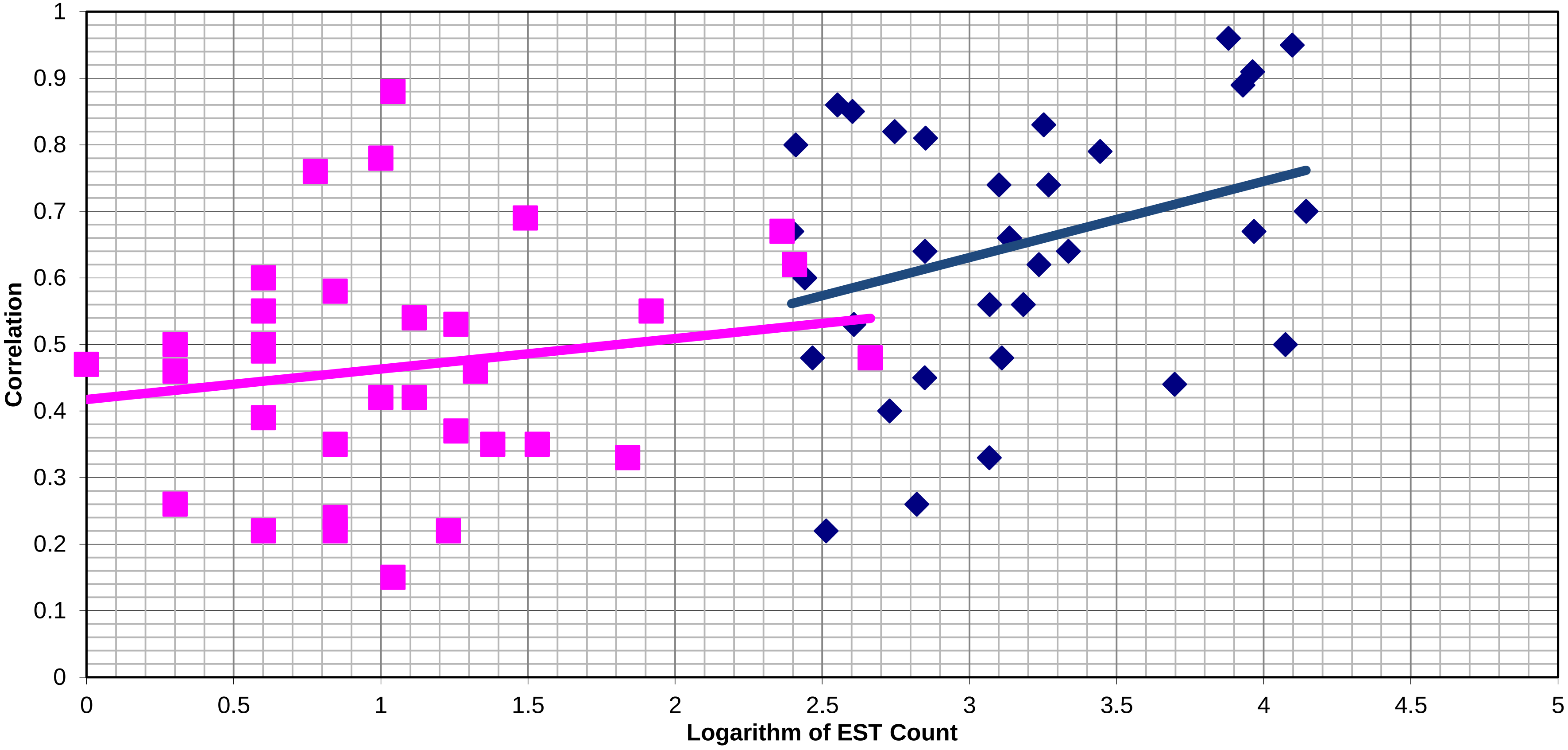


Figure 20. Pearson correlation values of original and scaled down libraries (y-axis vs. EST count (x-axis). Data points corresponding to the original libraries, as shown in Figure 19 (blue). Data points representing the modelled scaled down libraries, although the modelling involved non-linear transformation of the data, the graph shows similar degree of positive correlation between 0.15 and 0.88 (pink).

4.4.3. EST libraries from mixed, uncharacterised or poorly defined tissue preparations

It was further decided to apply the EST expression matrix to the identification of unknown or mixed tissue libraries. Small number of EST libraries annotated as being produced from uncharacterised tissues and therefore not included in the EST selection procedure, but for which their tissue origins are identifiable, were used. Figure 21A shows correlation results for one such library (NCI_CGAP_HN5), derived from gum tissue. This library shows clear positive correlation with the skin tissue type, which is the most related tissue type from the 26 tissue types included in the EST matrix, proving the accuracy of tissue typing using the matrix. Another example of uncharacterised library is the umbilical cord library (Stratagene endothelial cell 937223) which showed positive correlation with vascular tissue type and to a lesser degree with ovary and peripheral nervous system tissue types (Figure 21B). Whilst high positive correlation with vascular tissue and a degree of correlation with the ovary are likely, correlation with peripheral nervous system was unexpected because nervous fibres are only present in the proximal part of the umbilical cord (Marzioni et al, *2004*). However, since ovaries are innervated, the matching of both ovary and peripheral nervous system tissue types might be easily explained if the original preparation of umbilical cord contained some ovary tissue. In the absence of further independent information on that library source it would be reasonable to assume that the tissue could have contained some ovary tissue or was obtained from the proximal part of the umbilical cord. However, the highest positive correlation for this EST library is with vascular tissue which is the best match from the tissues available in the matrix. These examples show that the EST expression matrix can help to identify tissue origins of EST libraries. Figure 21C shows an example of correlations obtained for a pooled library (NIH_MGC_184). The correlations indicate the presence of a mixed (lung + thymus) tissues. Such a particular tissue mixture is not impossible, since these two tissues are normally situated in very close proximity to each other and the library may indeed have been made from such a mixed tissue preparation (the library annotation is "pooled tissue"). Another example of mixed tissue library "NCI_CGAP_HN20" is shown in Figure 21D. Correlations indicate the presence of ovary and thymus, the combination which is unlikely to have occurred by accidental tissue mixing, since the two organs are normally located far apart, but the library description does not specify the tissue origins and therefore no means exist to prove or disprove this tissue matching. A conclusion from this particular result would be to avoid using such a library for quantitative expression analysis. Figure 21E and Figure 21F exemplify correlation values obtained for embryonic libraries ("Embryo, 8 week I" and "Embryo, 12 week II" respectively). If these annotations are correct, and both libraries are made from the unfractionated embryonic tissue, the data would suggest that bone and brain tissue markers should have been more prominent at the earlier stages of development whilst towards week 12 muscle specific markers dominate. Such changes do indeed reflect the high prominence of the brain over the rest of the embryo at early gestation stages and the forming of bone around weeks 5 and 10 of gestation (Brakus et al, *2010*), followed by the development of muscle tissues and heart at later developmental stages (Allan, *2010*; Tanaka et al, *1995)* thus validating the interpretation. The stronger correlation with vascular tissue in the 12-week library is consistent with increasing vascularisation following the development of the heart.

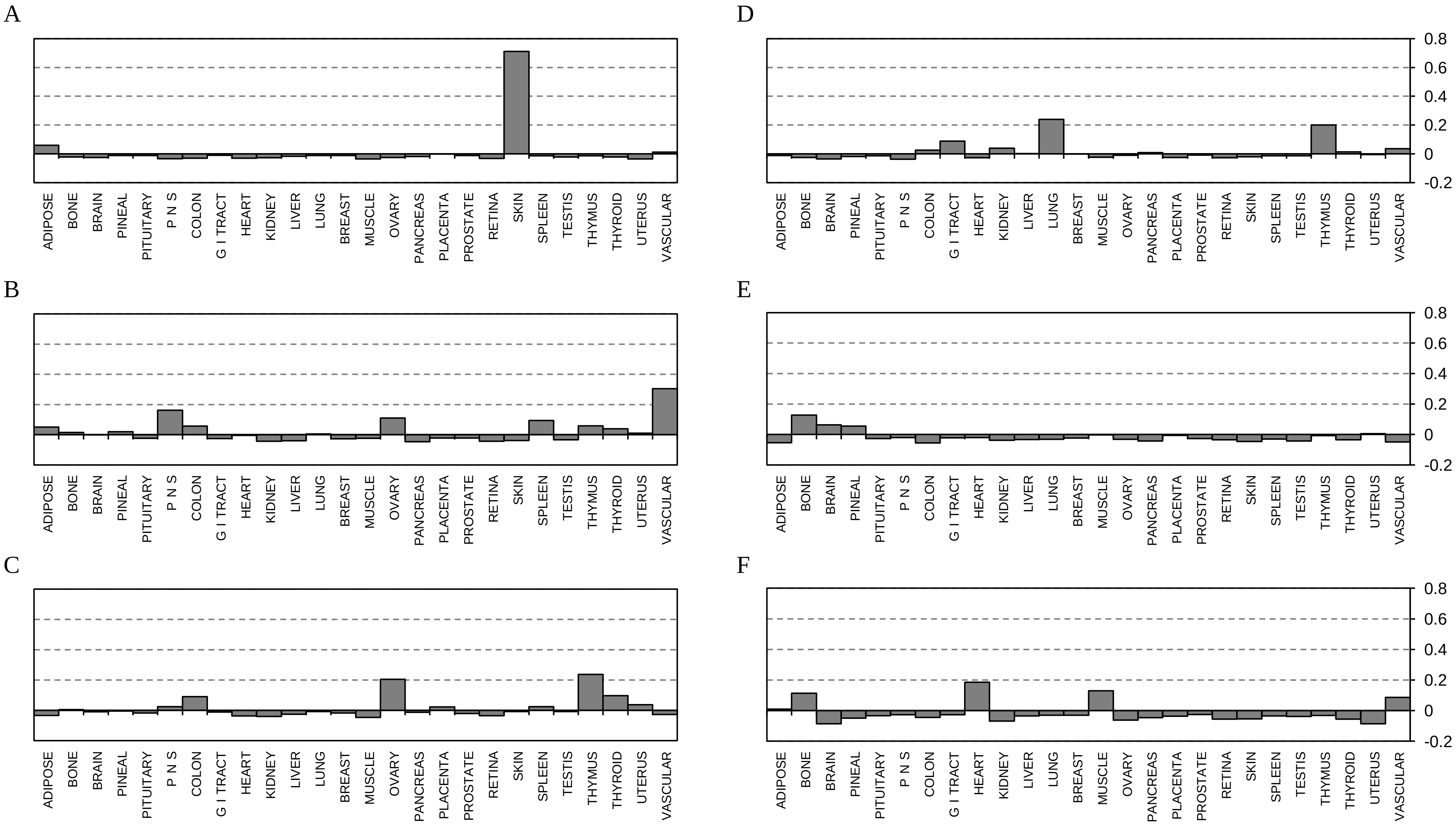

A
B
C
D
E
F
0.8
0.6
0.4
0.2
0
-0.2
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 21. Correlation of the EST matrix with individual libraries from uncharacterised or poorly defined tissue preparations. Pearson correlation coefficients (vertical axes) calculated between the individual EST libraries and the EST expression matrix. **A:** "Uncharacterised" library NCI_CGAP_HN5 derived from gum tissue. **B:** "Uncharacterised" Stratagene endothelial cell 937223 library. **C and D:** pooled libraries NIH_MGC_184 and NCI_CGAP_HN20 respectively. **E:** "Embryo, 8 week I" library. **F:** "Embryo, 12 week II" library.

4.4.4. EST libraries from cancer preparations

Although initial cluster selection procedure relied on both normal and cancer libraries, about 95% of all the transcripts found were present in normal tissues. The optimisation procedures relied on the normal EST libraries only. It was therefore interesting to see how the EST matrix would score cancer library preparations, which are expected to reflect aberrations in gene expression as well as genomic abnormalities which characterise cancers. Figure 22 shows a few typical examples of correlations obtained for a number of EST libraries from non-normalised bulk cancer tissues; these can be divided in two main categories. The first group represent cancer libraries which correlate well with the stated tissues of origins (Figure 22A – Figure 22C). One exception is a colon cancer library “NCI_CGAP_Co12”, where the "Gastrointestinal tract" EST profile scored nearly as well as the "Colon" profile. This is likely because of the close relation between the two tissue type definitions (as Figure 11B shows, the gastrointestinal tract library is annotated as originating from the neighbouring small intestine) or because a mixed tissue preparation was used, or both. The second group of libraries produced unexpected correlation results (Figure 22D – Figure 22F). The tissue of origin did not score in any of these, and the matching, at least numerically, was with apparently irrelevant tissues (liver instead of brain in “NCI_CGAP_Brn64”, thymus instead of kidney in “NCI_CGAP_Kid13” and no single tissue scored in brain cancer library “NCI_CGAP_Brn53” (Figure 22F). Clear tissue type matching in some cases of cancer derived libraries, but not in others is probably due to differences in cancer progression. It is reasonable to expect that gene expression changes will increase with the progression of cancer and the progressive deregulation of normal cellular processes. The decreasing accuracy of tissue matching for cancer samples using the EST expression matrix is an indication that the analysis should be capable, in principle, of accurate cancer staging.

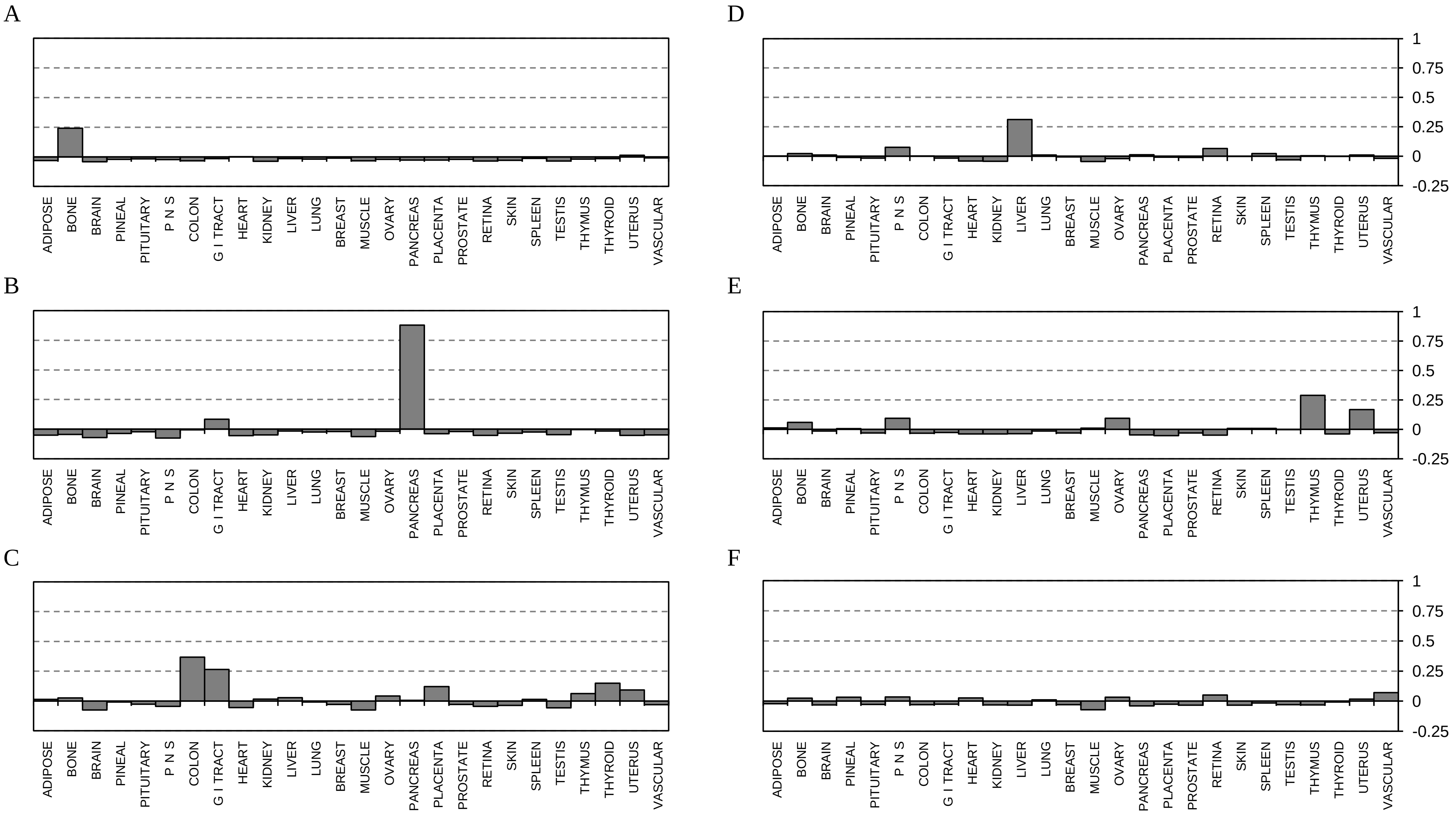

A
B
C
D
E
F
1
0.75
0.5
0.25
0
-0.25
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 22. Correlation of the EST expression matrix with individual EST libraries from cancer preparations. Pearson correlation coefficients (vertical axes) calculated between the individual EST libraries and the EST expression matrix. **A:** Bone cancer library NCI_CGAP_Ch1. **B:** Pancreatic library "Pancreas tumor III". **C:** Colon cancer library NCI_CGAP_Co12. **D:** Brain cancer library NCI_CGAP_Brn64. **E:** Kidney cancer library NCI_CGAP_Kid13. **F:** Brain cancer library NCI_CGAP_Brn53.

4.4.5. Normalised EST libraries

Normalising a cDNA library changes the apparent expression levels in that library and should ultimately remove any differences in the gene expression (in normalised libraries) or leave only differentially expressed cDNA transcripts (in subtracted libraries). The progressive disappearance of gene expression differences will depend on the degree of normalisation. It might be reasonable to assume that unless the library is completely normalised the genes which were highly over expressed originally may still have high EST counts, albeit reduced to some degree. For example if a hypothetical library containing three genes with relative abundances 1, 10, 100 is partially normalised to yield e.g. 11, 12 and 13 ESTs or e.g. 1, 2 and 3 ESTs, such three datasets would still correlate well with the original counts (for the above example the correlation would be +0.904 in both cases), and both such "normalised" libraries might both score reasonably well if correlated to EST expression matrix such as created in this work. Although normalisation and subtraction are in essence non-linear transformations we continued using Pearson product-moment correlation coefficient and did not calculate Spearman's and Kendall's correlation coefficients in order to keep the results comparable with all the previous calculations. The correlation data for a number of normalised libraries are shown in Figure 23. Normalised placenta library “NIH_MGC_148” correlated well with placental tissue array from the EST expression matrix scoring (+0.69) despite being normalised (Figure 23A). Two different normalised lung libraries “UI-CF-EC1” and “UI-CF-FN0” both had lung as the most highly positively scored tissue, but had different levels of unanticipated cross-tissues correlation (Figure 23B and Figure 23C). The data in Figure 23C show a degree of positive correlation with heart, muscle and spleen. Such unexpected cross-tissue relations probably arise from gradual loss of lung gene expression specificity following normalisation. This is clearly seen in Figure 23D, where normalised thymus library

“Soares_thymus_NHFTh” is scored using the EST matrix. That library correlated with none of the 26 tissue types in our EST matrix.

Using normalised libraries for the selection and optimisation the EST matrix was not feasible (with the degree of normalisation unknown no such optimisation was practically achievable). Therefore, an alternative approach was used to validate the lack of tissue correlations found in normalised library such as in Figure 23D. An artificial "normalised" EST matrix was created where all the 244 different transcript expression levels were set to "1" (except one value set to 0.999 to avoid a divide by zero error in calculating the Pearson correlation coefficient). This model "normalised" dataset was then correlated to the EST expression matrix. Similarly to the normalised thymus library “Soares_thymus_NHFTh”, the artificially "normalised" library did not correlate with any of the other tissues (Figure 23E). Such lack of any correlation between the model "normalised" dataset and any of the tissues confirms that high degree of library normalisation will yield zero correlations if compared with the EST matrix. To further test the robustness of the matrix another artificial dataset was created by assigning random values to each of the 244 transcripts. Such an artificially arbitrary array did not show positive correlation with any of the 26 tissues from the EST expression matrix. A representative graph is shown in Figure 23F. Thus only tissue-specific non-normalised cDNA libraries (such as in Figures 10 – Figure 12) are expected and have yielded positive correlations, proving the functionality of the new approach.

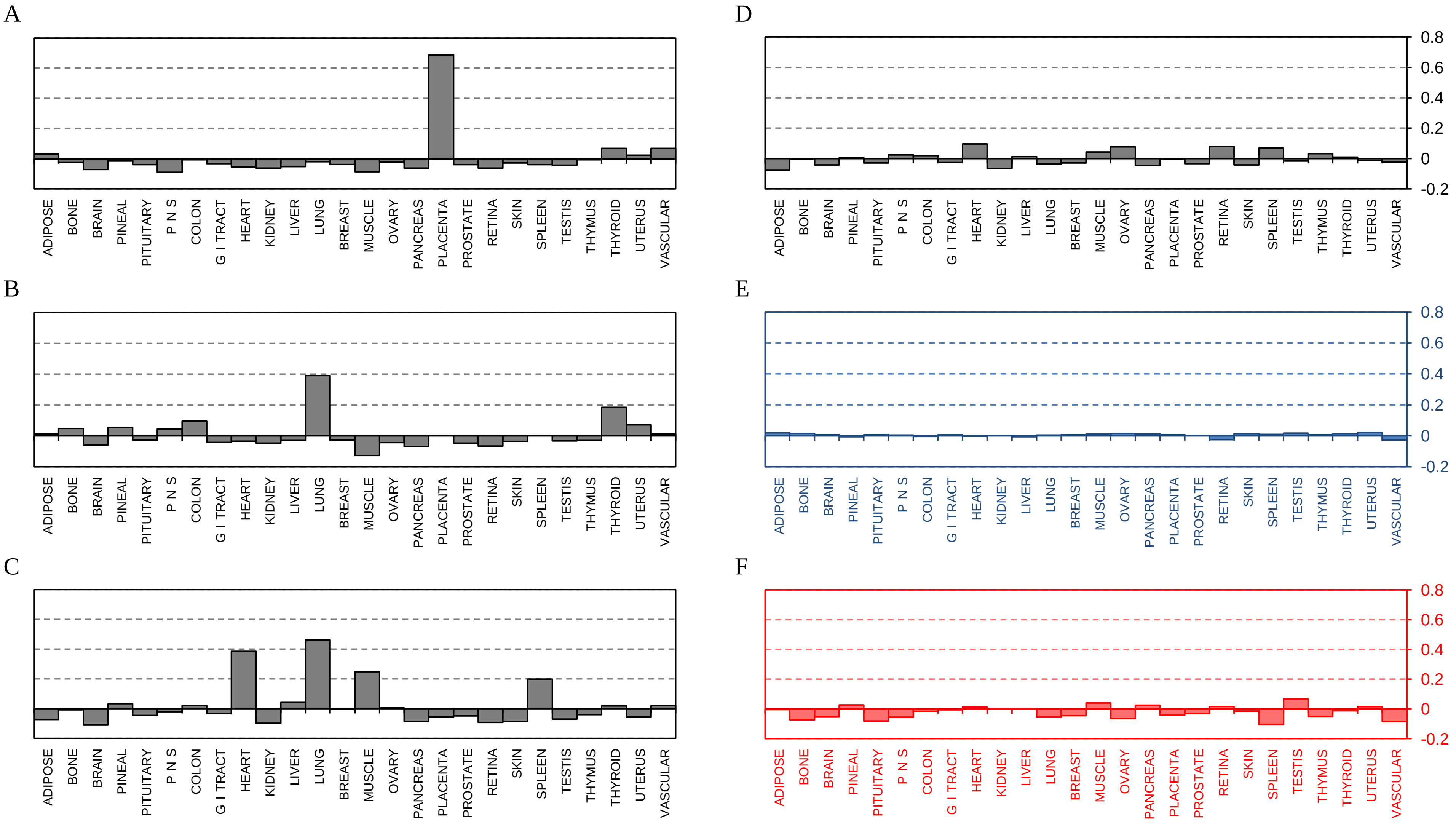

A
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
D
0.8
0.6
0.4
0.2
0
-0.2
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
B
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
E
0.8
0.6
0.4
0.2
0
-0.2
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
C
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR
F
0.8
0.6
0.4
0.2
0
-0.2
ADIPOSE
BONE
BRAIN
PINEAL
PITUITARY
P N S
COLON
G I TRACT
HEART
KIDNEY
LIVER
LUNG
BREAST
MUSCLE
OVARY
PANCREAS
PLACENTA
PROSTATE
RETINA
SKIN
SPLEEN
TESTIS
THYMUS
THYROID
UTERUS
VASCULAR

Figure 23. Correlation of the EST expression matrix with individual EST libraries from cancer preparations. Pearson correlation coefficients (vertical axes) calculated between the individual EST libraries and the EST expression matrix. **A:** Bone cancer library NCI_CGAP_Ch1. **B:** Pancreatic library "Pancreas tumor III". **C:** Colon cancer library NCI_CGAP_Co12. **D:** Brain cancer library NCI_CGAP_Brn64. **E:** Kidney cancer library NCI_CGAP_Kid13. **F:** Brain cancer library NCI_CGAP_Brn53.

5. Discussion

5.1. Errors in CGAP tools and database, and solutions to these problems

5.1.1. Errors in library search algorithm

The existing CGAP library parsing algorithm used by CGAP until 12 May 2010 appears to search for libraries which contain the required tissue name in their "keywords" field regardless of whether that tissue name (a text string) is part of a longer phrase (part of a longer text string) such as "clear cell ovarian tumor" and regardless of whether their "unique tissue" field states a relevant or irrelevant tissue origin. This resulted in a massively inaccurate choice of libraries and could easily lead to the selection of completely irrelevant libraries and yield artificial differences in gene expression and false disease markers. This is a major problem, which went undetected for many years and which require re-evaluation of all previously reported results where NCBI CGAP expression data and tools were used. CGAP creators allowed for the additional manual control of the choice of libraries before the gene expression data are obtained. But even this feature might not be practical for larger library collections, such as e.g. brain (over 1,000 libraries), or e.g. "uncharacterised tissue" (over 2,000 libraries, of which over half actually contain detailed descriptions with sufficient data for library classification).

As Table 1 shows, the CGAP hierarchical classification system also appears to consider libraries made from secondary tumours which have formed by metastasis of the primary tumour in the tissue in question, as belonging to that tissue. When brain tissue is selected, libraries are included from nine irrelevant tissues, including bone and bone marrow. The bone library in question was created from a Ewing’s sarcoma sample. Its inclusion under brain tissue is erroneous because this is known to be a bone condition, although it has been discovered that Ewing's sarcoma will metastasise to the right front parietal scalp, which is adjacent to the frontal and parietal lobes of the right cerebral hemisphere. Generally speaking, when a secondary tumour forms it will present a significantly different gene expression profile from the primary tumour due to the different gene expression profile of the secondary tumour's location. Hence, for the purpose of gene expression analysis, secondary tumours should not be considered as belonging to the tissues they metastasised from. Similarly, the inclusion of the bone marrow libraries under brain is erroneous because they are made from secondary metastases of the primary neuroblastoma, which is located in brain tissue (Ootsaka et al, *2008*; Yip et al, *2009*).

The suggested amendments implemented in our algorithm solve the problems by searching the contents of each library's "unique tissue" field, with the result that our tool groups only the correct libraries to the chosen tissue type. The effect of this is that any genes found to be differentially expressed between normal and cancerous libraries from that tissue will be genuine tumour markers because they are differentially expressed only in the specified tissue, and are not false positive results that are due to the impact of libraries from other tissues on the expression data, as could be the case with the CGAP results. If those results are further investigated, these errors may give rise to incorrectly designed diagnostic tests or treatments.

Once these findings were reported to NCBI by Andrew Milnthorpe on 12 May 2010 in a face-to-face discussion with Carl Schaefer at the US National Cancer Institute, this error in CGAP's library parsing algorithm was corrected. When last checked for this on 10 January 2011, both xProfiler and DGED algorithms search for libraries which contain the phrase for the chosen tissue in their “unique tissue” field and ignore libraries which contain this string as part of a longer string within this field. However, as of 25 January 2012, the CGAP parsing tools would still erroneously include libraries created from mixed tissue samples if their “keywords” annotations contain the required text phrase for the chosen tissue.

Furthermore, CGAP’s parsing algorithms show an inconsistency in whether libraries from genuinely dependent tissues are presented in the results for their parent tissue and do not show bone marrow libraries under bone tissue. The suggested amendments in our algorithm solve both of these problems through the availability of “pooled tissue” as a separate user-selectable tissue type instead of the inclusion of these libraries with tissues for which they contain one or more of the desired keywords, and the consistent presentation of libraries from dependent tissues in the results for their parent, if the user so-chooses (the recreated algorithm also gives the user control over whether dependent tissues are required).

5.1.2. Errors in CGAP’s gene search algorithm

The reasons were investigated as to why the list of cDNA xProfiler transcripts differ from the list obtained by DGED, as illustrated for adipose tissue in Figure 6. The internally available flat file database accessed by the cDNA xProfiler was found to show differences in the transcripts present in each library when compared with the publicly available CGAP relational database. It was not possible to find an explanation as to why this is so, given that not all transcripts which are reported by only one tool are related to those which are reported by both tools as shown in Table 14. The analysis of transcript annotations revealed that the cDNA xProfiler incorrectly lists cDNAs which are absent from the library pool, but which have names or functions similar to the transcripts present in the designed library pool. The effect of this is that, even if the list of libraries for the chosen tissue is correct (as they are for tissues such as adipose, as Table 1 shows), the transcript list could still include false positive differentially expressed genes or omit valid tumour markers which could otherwise warrant further investigation for use in cancer diagnosis or as a novel target for anticancer therapy.

Table 14. Data from UniGene relational database for α-actinin transcripts reported by CGAP xProfiler and/or cDNA DGED tools for a comparison of a pool containing normal adipose libraries with a pool containing cancerous adipose libraries.

Tool that reported transcript in either or both pools	**Symbol**	**Title**	**UniGene Cluster ID**
cDNA DGED only	ACTN4	Actinin, alpha 4	270291
cDNA DGED and cDNA xProfiler	ACTN1	Actinin, alpha 1	509765

The recreated transcript parsing algorithms solve the problems associated with CGAP's cDNA xProfiler algorithm by reporting only the UniGene transcripts which ESTs in each library map on to, thus reporting the same transcript regardless of whether the output format is Boolean or includes the EST odds ratios, as Table 3 shows. Now that the library parsing algorithm has also been corrected, this will ensure that the reported transcripts do not include false positive differentially expressed genes or omit genuine tumour markers which could otherwise be investigated further. Since these findings were reported to NCBI on 12 May 2010 this error has been corrected. Both tools (last accessed 10 January 2011) show identical numbers of genes when all the same parameters are used.

However, as of 17 November 2011 the inclusion of certain libraries in a cDNA xProfiler search in which transcripts are reported in all columns of the results table (known unique, unknown unique, known non-unique and unknown non-unique) causes at least one non-unique transcript found in both pools which is non-unique to be incorrectly reported as unique in the rows representing both pools, despite the fact that the same transcript is correctly presented as non-unique in the list for each individual pool. As Table 4 shows for brain this causes the number of transcripts in some boxes of cDNA xProfiler’s results table to be incorrect, with the correct numbers presented in Table 5.

This suggests that the cDNA xProfiler is incorrectly processing the internally flat file database it accesses, in which the transcripts associated with each library are divided into known, unknown, unique and non-unique groups. Although most of the problems had been corrected by 27 June 2012, some results continue to be incorrectly reported. This could lead to the discovery of false positive results or the omission of genuine candidate tumour biomarkers or targets from further study.

The Excel-based algorithm created in this investigation avoids this problem by searching the CGAP relational database instead, which does not report this information. However this algorithm can compare more than two pools in one search, so it is possible to ascertain whether a transcript is unique or non-unique to the chosen libraries by configuring an additional pool with every library except those in the other pools and elucidating which transcripts are also reported in that pool.

5.1.3. Problem with CGAP statistics

The statistics used by cDNA DGED are calculated using the Fisher Exact Test (Equation (1)) and the Benjamini-Hochberg False Discovery Rate (Equation (2)). Although the former does not depend on the display cut-off value for “F”, CGAP include “F” in the latter and therefore make the calculated false discovery rate “Q” value dependent on the display setting. We believe this is an error, contributing to the false discovery rate of tumour markers or the omission of potentially valid markers.

Although it was not possible to reproduce exactly the Benjamini-Hochberg implemented by cDNA DGED (Benjamini and Hochberg, *1995*), the Fisher Exact Test was implemented, on which these statistics are supposedly based. This approach is based on Equation (1) and it allowed the elucidation of where the observed expression difference of a given transcript between the two pools is due to chance. The output is given on a scale of zero to one, such that “P” value close to zero for a transcript indicates that the observed expression difference for that transcript is not due to chance. As Table 6 shows, our method yields the same “P” values regardless of the chosen display cut-off values.

5.1.4. Problem with number of sequences reported and inclusion of empty or missing database entries

The reason for the number of ESTs as annotated in the library database being greater than the ESTs of sequences which map onto the transcripts in the library was also studied. This is illustrated for adipose tissue in Table 7. This difference in the "sequences" annotation when compared to the "number of ESTs mapping onto transcripts in library" annotation could not explain the differences in the transcript lists produced by the CGAP tools and was thought to arise from the fact that some of the ESTs in the library did not map onto transcripts when the library was originally sequenced.

Also, although the user can filter the libraries by size (by setting the minimum number of ESTs per library), the CGAP tools use the “sequences” annotation in the library database (see Figure 1) to implement such a cut-off, rather than the number of ESTs which map onto the transcripts in the library. The CGAP approach produces results which are less reliable than they initially appear because, although the “sequences” annotation in the library database may be greater than the chosen cut-off value, the number of ESTs mapping onto the transcripts in the library may actually be below the cut-off.

The actual number of ESTs which map onto a library’s transcripts was also calculated and the library parsing algorithm was programmed to apply the ESTs display cut-off to this value rather than the “sequences” annotation of each library, which includes ESTs which do not map onto transcripts. As Table 1 shows this recreated algorithm reports the same total number of sequences in the library list as it does for the transcript list if the chosen output format shows the EST odds ratios, which in turn is the same as the value reported by CGAP’s cDNA DGED for the same libraries. The effect of this is that the user can more accurately apply this to determine the reliability of the reported libraries, for a library in which few ESTs map onto genes is less likely to provide a representative profile of gene expression in the sample from which it was created than a library in which many ESTs map onto transcripts. Furthermore, the display cut-off will not take into account any ESTs which do not map onto transcripts, so it can be used reliably to determine the quality of the results.

NCBI have not yet implemented a solution to this problem in the CGAP library and “gene” parsing algorithms (last checked on 12 January 2011). The sum of the number of ESTs per library annotations (in the “sequences” field, as reported by CGAP’s library parsing algorithm) is still greater than the number of ESTs the transcript parsing algorithm of cDNA DGED reports to be mapped onto all the transcripts in each pool at the top of the expression table. Furthermore, it has also been discovered that the library named “SARS-Cov infected lung” is incorrectly annotated as containing no ESTs and the CGAP tools erroneously report this. However, when the number of ESTs which map onto transcripts was calculated for all the libraries, this library was found to contain 1,083 ESTs which mapped onto 1,023 transcripts. This error could lead to the user deselecting this library due to its perceived non-contribution to the results, the consequences of which could be false positive discoveries or potentially valid tumour markers or targets not being further investigated. The recreated algorithm solves this problem by presenting the number of ESTs mapping onto transcripts and using this for the ESTs display cut-off.

The CGAP library database also appears to omit some of the libraries for which expression data is present in the expression database. This could result in an inaccurate choice of libraries for certain settings (however, it is impossible to know which combinations of settings this would apply to because of these libraries not being present in the library list). As with the problems identified in CGAP's library search algorithm, this could result in falsely discovered disease markers or the non-discovery of potentially genuine biomarkers or targets.

Due to the lack of information available on these additional libraries it was not possible to enter them into the library database used by the recreated algorithm, so these libraries were therefore omitted from the expression database. This has the advantage of increasing the number of searches that can be performed within a given time period.

In addition to excluding libraries containing expression information as explained above, the CGAP library database also appears to include libraries which do not contain any transcript-mapping ESTs, as illustrated in Figure 7. Similarly, the transcript database was found to contain transcripts which were not found to map onto any ESTs in any libraries, as Figure 8 shows.

These are errors because the purpose of the database is for gene expression profiling and not to serve as a catalogue. The libraries and transcripts concerned will not make any contribution to any results and will therefore not suggest any candidate biomarkers or targets for further investigation. Furthermore, the number of libraries reported for a particular tissue can give an indication of tissue coverage (if more libraries are present for a tissue it is more likely that the results will be representative of *in vivo* expression levels), so the inclusion of empty libraries will mislead. For this reason the number of transcript-mapping ESTs in those libraries must also be taken into account when assessing tissue coverage.

5.2. Further features of the created Excel-based tools

5.2.1. Features included in Excel-based tools to address these shortcomings

Two new algorithms have been created in Excel to mimic the capabilities of CGAP's cDNA xProfiler and cDNA DGED tools. In addition to solving the problems with the CGAP tools identified above, the new algorithms allow the user to compare three or more groups of samples to be compared side-by-side, a feature limited only by the size of an Excel worksheet, which permits a comparison of 8,192 groups of samples. The new tools also allow mixed tissue libraries to be selected for inclusion in a pool separately from libraries from other tissues, unlike the CGAP tools which include them if one or more of the tissues listed in their "keywords" field is chosen, despite the fact that their gene expression levels are likely to be different from those in the chosen tissue.

The new Excel-tools also make available lists of developmental stages, genders and pregnancy states that the user can choose to filter the list of reported libraries according to which of these they are annotated with. This would enable results to be displayed which are solely caused by or a consequence of the cancer of interest, and not due to gene expression in the other situations detailed above.

Finally, the new tools allow the user to choose whether the specified library name search string is used to present partial matches along with exact matches, or to report exact matches only. This would avoid the problem of a library whose name is a partial match presenting a different expression profile and providing false positive results or leading to the omission of genuine candidate markers from the results.

5.2.2. Shortcomings of CGAP tools that are solved by the inclusion of the new features in the Excel-based tools

Even if existing tools did not contain the errors detailed above, there would still be omissions made by the CGAP algorithms which are sorted with the inclusion of the features described above in the new Excel-based algorithms. Currently the accepted gene expression profiling practice is to compare two groups of samples (EST libraries in the case of the CGAP tools) side-by-side (National Cancer Institute, *n.d.e*) (also reviewed in (Murray et al, *2007*)). However, there are situations where more than two pools would be useful, such as for studying the gene expression in tissues related or proximal to the tissue containing a tumour as well as the tissue to which the tumour is localised, at the same time as comparing that expression with the tumour and its tissue. This would reveal genes which are overexpressed in the tumour compared to all the chosen tissues, as well as enabling the separate study of those tissues. Furthermore, one type of cancer could be studied separately from and compared with another in the same way. For example, genes could be identified which are only expressed in papillary thyroid carcinoma and not in medullary thyroid carcinoma, thyroid or any related tissues (see Equation (7)).

The existing CGAP tools do not allow libraries created from mixed or uncharacterised samples to be selected separately except those annotated as “uncharacterised tissue”. This could be useful, for example, in extending the quality control investigation by searching for expression data in such libraries and then comparing the values for the 244 marker genes with those in the EST matrix, to determine if such libraries can, for example, be annotated with the correct tissue histology or library protocol.

The CGAP tools also do not enable the user to select which libraries should be displayed according to developmental stage, gender or pregnancy state. This must be accomplished by selecting and deselecting libraries as appropriate once the list of matching libraries is presented. Filtering the library list to present settings that the user desires would avoid genes being further investigated whose expression levels are due to the developmental stage, gender or pregnancy state of the individual rather than as a cause or consequence of the cancer.

For example, selecting only adult libraries would mean that any observed expression of fibroblast growth factor signalling factors is due to the cancer of interest. While overexpression of these proteins can lead to cancer, these proteins also play a major role in embryogenesis (Dailey et al, *2005*). Similarly, selecting just one gender will result in gene expression levels which are not dependent on gender, perhaps due to different lifestyles undertaken by each gender. This would avoid the detection of KI-67, BLC-2 and CD-44 if the focus is on non-small lung cancer in women, for these genes have been found to be more highly expressed in men. While this has been thought to be due to the increased incidence of smoking amongst men (D'Amico et al, *2000*), it has also been discovered that men have a higher risk of mortality even after smoking and treatment histories are taken into account (Visbal et al, *2004*). Filtering by gender would eliminate the risk of false positive results or the exclusion of genuine markers from further investigation due to differences in expression levels being due to different patients' genders. Furthermore, excluding libraries from pregnant individuals would ensure that observed overexpression of genes such as pregnancy-specific β1 glycoprotein 9, which has been found to be overexpressed during the early stages of colorectal cancer (Salahshor et al, *2005*), is a result of the cancer of interest and not due to libraries being included from pregnant individuals. In a similar manner, cancers also utilise the mechanisms which support pregnancy through the invasion of the uterus by the placenta and the evasion of the host immune system by the developing foetus (Holtan et al, *2009*). Therefore, not showing libraries from pregnant individuals would lead to results solely due to the development of the cancer.

Finally, the CGAP tools at present will present partial matches for a library name search string. However, two libraries which are from the same source but are named differently in this way may present different gene expression profiles. For example, "a library named "Aorta endothelial cells" may present different expression information than one named "Aorta endothelial cells, TNF alpha-treated". In this case, it has previously been reported that platelet endothelial cell-adhesion molecule 1 (PECAM-1, CD31) expression significantly decreases when the endothelial cells are TNF-alpha treated (Stewart et al, *1996*). This is because along with interferon-γ, TNF-α induces the expression of many genes during the inflammatory response (Ohmori et al, *1997*).

5.3. Creation of a procedure for the quality control of gene expression data

5.3.1. Optimisation of the list of tissue specific transcripts and creation of the EST expression matrix

Along with the corrections to the errors in the CGAP tools and the implementation of improvements to those algorithms, the need was also identified for a quality control method for expression data based purely on the data itself.

In the quest to create this method, the CGAP algorithms were used to select the initial list of 1,437 transcripts, which was subjected to two rounds of optimising 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), and present a high normalised EST count in the target tissue.

An EST expression matrix of these markers in 26 tissues was created and used as the control against which other libraries were compared. As mentioned earlier, tissue-specific gene expression has previously been used as a quality control method to assess three SAGE databases (Huminiecki et al, *2003*), but it has not previously been used as a quality control method within a single database.

5.3.2. Tissue typing of EST libraries using EST expression matrix

The EST expression matrix was correlated with 113 libraries of known identity. The correlations presented in Figure 10 – Figure 12 show that the EST expression matrix is more versatile than had been anticipated, for it was not only possible to correctly confirm the tissue origin of the libraries presented, the matrix could also identify distant but related tissue types, as illustrated by Figure 12, which also proves that the matrix can identify possible contamination from tissues which are in close proximity to the one of interest. Figure 13, which summarises all the correlations from Figures 10 – Figure 12, shows that the inter-tissue correlations presented in Figure 12 were significantly greater than those between non-matching tissues. Figure 13 also reveals that all the inter-tissue correlations were significantly smaller than the intra-tissue correlations.

Further to this the matrix was correlated with small EST model libraries generated from the original libraries. The findings presented in Figure 14 – Figure 18 and Table 9 – Table 13 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). This is despite a clear relationship between the size of a library and the quality of the tissue match, as presented in Figure 19, and the positive intra-tissue correlations reported for the scaled down libraries were still significant, as reported in Figure 20. These findings show that tissue-specific gene expression can be used as a robust quality control method because it can be used to correctly identify small libraries, which are likely to be less representative than large libraries due to the increased likelihood of only the more abundant transcripts being included.

The results presented in Figure 21 further confirm the potential use of the EST expression matrix as a means to elucidate the tissue of origin of libraries whose tissue identity is unknown or not listed in the database record. Six libraries were used whose tissue origins can be identified but which were not well characterised enough for selection and optimisation. It was possible to identify the tissue origin of all six libraries, and in all but one case the identity matched the annotation, except for the pooled tissue library entitled "NIH_MGC_184". As Figure 21C shows, this library correlates best with lung and thymus, suggesting it was created from a mixture of these tissues, which are in close proximity to one another. However, this library is annotated as originating from adrenal gland, parathyroid, thyroid and pineal gland. Because all other libraries were correctly annotated (or, as with the other pooled tissue library entitled "NCI_CGAP_HN20 (Figure 21D), were not annotated in enough detail to be incorrectly labelled), this shows that the annotation of this library was incorrect. This shows that the EST expression matrix can be used to identify incorrect annotations as well as verify the identity of correctly annotated libraries or characterise those which are not sufficiently annotated.

Cancer libraries were excluded from all but the initial cluster selection procedure, and around 95%of the transcripts detected were contained within libraries from normal tissues. Cancer libraries were excluded from most of this work because they are known to show changes in gene expression as well as mutations to the genome (and therefore the transcriptome) which are characteristic of the type of cancer they were created from. As the disease progresses, gene expression in cancer is known to increasingly no longer resemble normal gene expression in the tissue in which the primary tumour arose. As Figure 22 shows, six cancer libraries were compared with the EST matrix to elucidate its potential in cancer staging. Figure 22A – Figure 22C present libraries which correlate well with the tissue with which they are annotated. The exception is “NCI_CGAP_Co12”, which is shown in Figure 22C. It is believed that the stronger correlation with colon than with gastrointestinal tract (as Figure 11B shows this is the adjacent small intestine) arose because these two tissues are closely related as part of the same organ system, or because the two tissues were pooled together during library preparation. Figure 22D – Figure 22F show three libraries whose expression profiles did not match that of the stated tissue of origin. It is though that this is due to these libraries being from later stages of cancer, in which the disease has metastasised to other tissues and the resulting secondary tumours present gene expression profiles resembling that of their new location. This shows that the EST expression matrix can also be used for accurate cancer staging.

Normalised libraries were not used in any of the methods used to prepare the EST expression matrix. This is because the process of normalising a cDNA library alters the relative differences in transcript abundance levels. The resemblance of the resulting gene expression profile to that in the state tissue of origin should depend on the extent to which the library has been normalised. This was confirmed by the six libraries presented in Figure 23which show different degrees of correlation with their annotated tissue of origin. Figure 23A shows “NIH_MGC_148”, a placental library which still correlates highly with placenta, indicating a low degree of normalisation. Figure 23B and Figure 23C present two lung libraries which increased correlation with other tissues, suggesting a greater degree of normalisation, the correlations arising from a gradual loss of gene expression matching that of the annotated tissue due to increased normalisation. This is confirmed by Figure 23D, where “Soares_thymus_NHFTh” does not correlate with any of the tissues in the matrix, suggesting that this library is almost completely normalised. This was confirmed through the creation of an artificial normalised library where all the expression levels were given an equal value of one, which, as Figure 23E shows, showed no correlation at all with any tissue. Similarly, no positive correlation was presented by the artificial library of random values presented in Figure 23F. These results show that it is possible to use the EST matrix to show the degree of normalisation of normalised libraries, for this is indicated by the degree of correlation with the annotated tissue of origin.

5.3.3. Shortcomings of existing research the EST matrix has the potential to solve

These findings show that tissue-specific gene expression can be used as a quality control method, an idea not examined by previous studies. Other investigations focussed on the whole genome (Liang et al, *2006*), studied aspects such as GC content (Arhondakis et al, *2006*) or, even if they focussed on tissue-specific gene expression, as a few did (Russ and Futschik, *2010*), did not use such 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*). Furthermore, tissue-specific genes have been identified in this investigation which are also highly expressed in their target tissues, unlike the genes reported previously in (Li et al, *2011*).This study is also an improvement on many existing search tools and secondary database, including those hosted by CGAP, which are merely information repositories and retrieval algorithms with no 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*).

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. Furthermore, this approach could be used to correctly identify very small libraries, which will have a lower depth of sequencing and will therefore not provide as good a quantitative estimate of gene expression than larger libraries (Simon et al, 2009) due to the reduced likelihood of rare transcripts being included (Bashir et at al, *2010*). The effect of library size has been included previously in statistical tests, which have been used to study gene expression levels in a range of cancers (Abba et al, *2004*; Baggerly et al, *2003*, *2004*; Robinson and Smyth, 2007; Ruijter et al, *2002*; Silveira et al, *2008*; Thygesen, *2006*), but this study is different from previous investigations in that its effect on inter-library correlations was studied.

Although the correlation with the original tissue was reduced, the scaled down libraries still presented an extremely good match for the tissue of origin, confirming the matrix as an extremely robust method of quality control.

The next step is to adapt and apply this method to other publicly available gene expression data. It is envisaged that with the increasing amounts of EST expression data, the optimised EST marker set could be improved and the tissue range might be expanded. The use of other expression information, which could be obtained from SAGE data (Leyritz et al, *2008*), DNA microarray data (Baron et al, *2011*) and northern blots (Schlamp et al, *2008*) and the merging of this data could improve the selection even further. It is also envisaged that increasing amounts of available data could further decrease the number of transcripts in the expression matrix and may allow accurate analysis and tissue typing of the related and dependent tissues.

6. Conclusions

CGAP's cDNA xProfiler and cDNA DGED have been used by the scientific community to find genes which are differentially expressed in cancer for over ten years. It is known that such genes could be used as indicative biomarkers in diagnostic or prognostic tests or as therapeutic targets in novel treatments. However, the currently accepted practice is to compare two pools of EST libraries, preventing the cancer and normal tissue in which the cancer is located from being compared with related or proximal tissues in the same search. Providing 7 pools would enable genes which are preferentially expressed in the cancer compared to related or proximal tissues as well as the local tissue to be discovered and investigated much more efficiently than is currently possible.

The provision of multiple pools makes a greater range of investigations possible in the same time frame. However, everything depends on the algorithms themselves being correctly written. The libraries reported must be the ones which originate from the specified tissue and the transcripts presented must be exactly those which are reported in the chosen libraries. Furthermore, the statistics used must indicate the significance of differential expression between the two groups of libraries and not be dependent on the proportion of the results displayed. Finally the data used must be archived and annotated correctly. The CGAP algorithms and databases were found to contain significant errors which impact investigations carried out using those tools in all of these ways. These problems have the potential to lead to incorrect diagnostic tests or treatments. These sources of error would be eliminated by reconfiguring the database and recoding the tools in the ways suggested here.

Correction of the above problems would ensure that the results from investigations into differentially expressed genes in cancer would not be affected by errors in the algorithms. However, the results are still dependent on the EST counts themselves being correct, which existing algorithms assume to be the case. It has been shown here that the tissue type annotations of EST libraries could be verified by using an EST expression matrix based on tissue-specific markers, showing this method to be a suitable means of quality control. 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. Moreover, cancer staging can be performed by correlating the expression levels in a cancer library with those in the matrix to assess the degree of similarity with the stated tissue location of the tumour. Another possible use of the matrix could be to indicate the amount of normalisation a library has undergone from its degree of resemblance to the tissue with which it is annotated.

Together, these findings increase the reliability of the results of differential gene expression studies for cancer, eliminating the possibility of such errors leading to misdiagnosis of cancer patients and incorrectly applied therapy.

Bibliography

Abba, M.C., Drake, J.A., Hawkins, K.A., Hu, Y., Sun, H., Notcovich, C., Gaddis, S., Sahin, A., Baggerly, K., Aldaz, C.M. (2004) Transcriptomic changes in human breast cancer progression as determined by serial analysis of gene expression. *Breast Cancer Research.* **6** (5) pp. R499 – R513.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubrick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., Venter, J.C. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science.* **252** (5,013) pp. 1,651 – 1,656.

Ahmed, F.E., Vos, P., iJames, S., Lysle, D.T., Allison, R.R., Flake, G., Sinar, D.R., Naziri, W., Marcuard, S.P., Pennington, R. (2007) Transcriptomic molecular markers for screening human colon cancer in stool and tissue. *Cancer Genomics & Proteomics.* **4** (1) pp.1 – 20.

Ahmed, F.E., Vos, P.W., Ijames, S., Lysle, D.T., Flake, G., Sinar, D.R., Naziri, W., Marcuard, S.P. (2007) Standardization for transcriptomic molecular markers to screen human colon cancer. *Cancer Genomics & Proteomics.* **4** (6) pp. 419 – 431.

Alaiya, A., Roblick, U., Egevad, L., Carlsson, A., Franzén, B., Volz, D., Huwendiek, S., Linder, S., Auer, G. (2000) Polypeptide expression in prostate hyperplasia and prostate adenocarcinoma. *Analytical Cellular Pathology.* **21** (1) pp. 1 – 9.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O., Staudt, L.M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* **403** (6,769) pp. 503 – 511.

Allan, L. (2010) Fetal cardiac scanning today. *Prenatal Diagnosis.* **30** (7) pp. 639 – 643.

D'Amico, T.A., Aloia, T.A., Moore, M.B., Herndon, J.E. 2nd., Brooks, K.R., Lau, C.L., Harpole, D.H. Jr. (2000) Molecular biologic substaging of stage I lung cancer according to gender and histology. *The Annals of Thoracic Surgery.* **69** (3) pp. 882 – 886.

Andrews-Pfannkoch, C., Fadrosh, D.W., Thorpe, J., Williamson, S.J. (2010) Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Applied and Environmental Microbiology.* **76** (15) pp. 5,039 – 5,045.

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

Arsanious, A., Bjarnason, G.A., Yousef, G.M. (2009) From bench to bedside: current and future applications of molecular profiling in renal cell carcinoma. *Molecular Cancer.* **8** (20).

Audic, S., Claverie, J.M. (1997) The significance of digital gene expression profiles. *Genome Research***. 7** (10) pp. 986 – 995.

Baggerly, K.A., Deng, L., Morris, J.S., Aldaz, C.M. (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics.* **19** (12) pp. 1,477 – 1,483.

Baggerly, K.A., Deng, L., Morris, J.S., Aldaz, C.M. (2004) Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. *BMC Bioinformatics.* **5** (144).

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).

Bashir, A., Bansal, V., Bafna, V. (2010) Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. *BMC Genomics.* **11** (385).

Beaty, R.M., Edwards, J.B., Boon, K., Siu, I.M., Conway, J.E., Riggins, G.J. (2007) PLXDC1 (TEM7) is identified in a genome-wide expression screen of glioblastoma endothelium. *Journal of Neuro-oncology.* **81** (3) pp. 241 – 248.

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** (7) pp. 1,001 – 1,010.

Benjamini, Y. Y. Hochberg, Y. (1995) Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological).* **57** (1) pp. 289 – 300.

Bidon, N., Brichory, F., Hanash, S., Bourguet, P., Dazord, L., Le Pennec, J.P. (2001) Two messenger RNAs and five isoforms for Po66-CBP, a galectin-8 homolog in a human lung carcinoma cell line. *Gene.* **274** (1 – 2) pp. 253 – 262.

Bonaldo, M.F., Lennon, G., Soares, M.B. (1996) Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Research.* **6** (9) pp. 791 – 806.

Bracken, P. (2003) Properties of Certain Sequences Related to Stirling's Approximation for the Gamma Function. *Expositiones Mathematicae.* **21** (2) pp. 171 – 178.

Brakus, S.M., Govorko, D.K., Vukojevic, K., Jakus, I.A., Carev, D., Petricevic, J., Saraga-Babic, M. (2010) Apoptotic and anti-apoptotic factors in early human mandible development. *European Journal of Oral Sciences.* **118** (6) pp. 537 – 546.

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.

Bumcrot, D., Manoharan, M., Koteliansky, V., Sah, D.W. (2006) RNAi therapeutics: a potential new class of pharmaceutical drugs. *Nature Chemical Biology.* **2** (12) pp. 711 – 719.

Chen, X., Ji, Z.L., Chen, Y.Z. (2002) TTD: Therapeutic Target Database. *Nucleic Acids Research.* **30** (1) pp. 412 – 415.

Chen, Z., Wang, W., Ling, X.B., Liu, J.J., Chen, L. (2006) GO-Diff: mining functional differentiation between EST-based transcriptomes. *BMC Bioinformatics.* **7** (72).

Coutelle, C., Speer, A., Rogers, J., Kalsheker, N., Humphries, S., Williamson, R. (1985) Construction and partial characterization of a human liver cDNA library. *Biomedica Biochimica Acta.* **44** (3) pp. 421 – 431.

Dailey, L., Ambrosetti, D., Mansukhani, A., Basilico, C. (2005) Mechanisms underlying differential responses to FGF signalling. *Cytokine & Growth Factor Reviews.* **16** (2) pp. 233 – 247.

Davis, M.E., Zuckerman, J.E., Choi, C.H., Seligson, D., Tolcher, A., Alabi, C.A., Yen, Y., Heidel, J.D., Ribas, A. (2010) Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles. *Nature.* **464** (7,291) pp. 1,067 – 1,070.

Daya, S. (2002) Fisher Exact Test. *Evidence-based Obstetrics and Gynecology.* **4**(1) pp. 3 – 4.

De Young, M.P., Damania, H., Scheurle, D., Zylberberg, C., Narayanan, R. (2002) Bioinformatics-based discovery of a novel factor with apparent specificity to colon cancer. *In Vivo.* **16** (4) pp. 239 – 248.

Delaval, B., Birnbaum, D. (2007) A cell cycle hypothesis of cooperative oncogenesis (Review). *International Journal of Oncology.* **30** (5) pp. 1,051 – 1,058.

Deyoung, M.P., Scheurle, D., Damania, H., Zylberberg, C., Narayanan, R. (2002) Down's syndrome-associated single minded gene as a novel tumor marker. *Anticancer Research.* **22** (6A) pp. 3,149 – 3,157.

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) pp. D613 – D616.

Elek, J., Pinzon, W., Park, K.H., Narayanan, R. (2000) Relevant genomics of neurotensin receptor in cancer. *Anticancer Research.* **20** (1A) pp. 53 – 58.

European Bioinformatics Institute (2012) *ArrayExpress Archive* [Online]. Available: http://www.ebi.ac.uk/arrayexpress/ [Accessed 20 July 2012].

Genetech Inc (n.d.) *GeneHub-GEPIS* [Online]. Available: http://research-public.gene.com/Research/genentech/genehub-gepis/index.html [Accessed 21 July 2012].

Gudas, J.M., Payton, M., Thukral, S., Chen, E., Bass, M., Robinson, M.O., Coats, S. (1999) Cyclin E2, a novel G1 cyclin that binds Cdk2 and is aberrantly expressed in human cancers. *Molecular and Cellular Biology.* **19** (1) pp. 612 – 622.

Holtan, S.G., Creedon, D.J., Haluska, P., Markovic, S.N. (2009) Cancer and pregnancy: parallels in growth, invasion, and immune modulation and implications for cancer therapeutic agents. *Mayo Clinic Proceedings***. 84** (11) pp. 985 – 1,000.

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., Jiang, C.L., 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.

Huang, Z.G., Ran, Z.H., Lu, W., Xiao, S.D. (2006) Analysis of gene expression profile in colon cancer using the Cancer Genome Anatomy Project and RNA interference. *Chinese Journal of Digestive Diseases.* **7** (2) pp. 97 – 102.

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** (1) p. 31.

Ichikawa, Y., Hirokawa, M., Aiba, N., Fujishima, N., Komatsuda, A., Saitoh, H., Kume, M., Miura, I., Sawada, K. (2004) Monitoring the expression profiles of doxorubicin-resistant K562 human leukemia cells by serial analysis of gene expression. *International Journal of Hermatology.* **79** (3) pp. 276 – 282.

Israeli, R.S., Powell, C.T., Fair, W.R., Heston, W.D. (1993) Molecular cloning of a complementary DNA encoding a prostate-specific membrane antigen. *Cancer Research.* **53** (2) pp. 227 – 230.

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

Kawamoto, S., Yoshii, J., Mizuno, K., Ito, K., Miyamoto, Y., Ohnishi, T., Matoba, R., Hori, N., Matsumoto, Y., Okumura, T., Nakao, Y., Yoshii, H., Arimoto, J., Ohashi, H., Nakanishi, H., Ohno, I., Hashimoto, J., Shimizu, K., Maeda, K., Kuriyama, H., Nishida, K., Shimizu-Matsumoto, A., Adachi, W., Ito, R., Kawasaki, S., Chae, K.S. (2000) BodyMap: a collection of 3' ESTs for analysis of human gene expression information. *Genome Research.* **10** (11) pp. 1,817 – 1,827.

Ke, X., Wang, J., Gao, Z., Zhao, L., Li, M., Jing, H., Wang, J., Zhao, W., Gilbert, H., Yang, X.F. (2010) Clinical characteristics and prognostic analysis of Chinese patients with diffuse large B-cell lymphoma. *Blood Cells, Molecules & Diseases.* **44** (1) pp. 55 – 61.

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) pp. 36,592 – 36,600.

Lal, A., Lash, A.E., Altschul, S.F., Velculescu, V., Zhang, L., McLendon, R.E., Marra, M.A., Prange, C., Morin, P.J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K.W., Strausberg, R.L., Riggins, G.J. (1999) A public database for gene expression in human cancers. *Cancer Research.* **59** (21) pp. 5,403 – 5,407.

Lapointe, J., Li, C., Higgins, J.P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A.M., Tibshirani, R., Botstein, D., Brown, P.O., Brooks, J.D., Pollack, J.R. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America.* **101** (3) pp. 811 – 816.

Larsson, S.C., Orsini, N., Wolk, A. (2010) Vitamin B6 and risk of colorectal cancer: a meta-analysis of prospective studies. *JAMA.* **303** (11) pp. 1,077 – 1,083.

Lee, D.K., Nguyen, T., Lynch, K.R., Cheng, R., Vanti, W.B., Arkhitko, O., Lewis, T., Evans, J.F., George, S.R., O'Dowd, B.F. (2001) Discovery and mapping of ten novel G protein-coupled receptor genes. *Gene.* **275** (1) pp. 83 – 91.

Leyritz, J., Schicklin, S., Blachon, S., Keime, C., Robardet, C., Boulicaut, J.F., Besson, J., Pensa, R.G., 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.

Livingston, D.M., Shivdasani, R. (2001) Toward mechanism-based cancer care. *JAMA.* **285** (5) pp. 588 – 593.

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

Loging, W.T., Lal, A., Siu, I.M., Loney, T.L., Wikstrand, C.J., Marra, M.A., Prange, C., Bigner, D.D., Strausberg, R.L., Riggins, G.J. (2000) Identifying potential tumor markers and antigens by database mining and rapid expression screening. *Genome Research.* **10** (9) pp. 1,393 – 1,402.

Marzioni, D., Tamagnone, L., Capparuccia, L., Marchini, C., Amici, A., Todros, T., Bischof, P., Neidhart, S., Grenningloh, G., Castellucci, M. (2004) Restricted innervation of uterus and placenta during pregnancy: evidence for a role of the repelling signal Semaphorin 3A. *Developmental Dynamics.* **231** (4) pp. 839 – 848.

Meng, L.X., Li, Q., Xue, Y.J., Guo, R.D., Zhang, Y.Q., Song, X.Y. (2007) [Identification of gastric cancer-related genes by multiple high throughput analysis and data mining] [Article in Chinese]. *Zhonghua Wei Chang Wai Ke Za Zhi.* **10** (2) pp. 169 – 172.

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.

Mitas, M., Mikhitarian, K., Hoover, L., Lockett, M.A., Kelley, L., Hill, A., Gillanders, W.E., Cole, D.J. (2002) Prostate-Specific Ets (PSE) factor: a novel marker for detection of metastatic breast cancer in axillary lymph nodes. *British Journal of Cancer* **86** (6) pp. 899 – 904.

Morgan, D. (2002) 'The Cell Cycle and Programmed Cell Death.' In *Molecular Biology of the Cell 4th Edition*. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter P. eds. Garland Science: New York, U.SA. pp. 983 – 1026.

Mortici, C. (2011) A New Stirling series as continued fraction. *Numerical Algorithms.* **56** (1) pp. 17 – 26.

Murray, D., Doran, P., MacMathuna, P., Moss, A.C. (2007) In silico gene expression analysis--an overview. *Molecular Cancer.* **6** (50).

Nam, M.J., Kee, M.K., Kuick, R., Hanash, S.M. (2005) Identification of defensin alpha6 as a potential biomarker in colon adenocarcinoma. *The Journal of Biological Chemistry*. **280** (9) pp. 8,260 – 8,265.

National Cancer Institute (n.d.a) *Cancer Genome Anatomy Project (CGAP)* [Online]. Available: http://cgap.nci.nih.gov/cgap.html [Accessed 20 July 2012].

National Cancer Institute (n.d.b) *cDNA xProfiler* [Online]. Available: http://cgap.nci.nih.gov/Tissues/xProfiler [Accessed 21 July 2012].

National Cancer Institute (n.d.c) *cDNA Digital Gene Expression Displayer (DGED)* [Online]. Available: http://cgap.nci.nih.gov/Tissues/GXS [Accessed 21 July 2012].

National Cancer Institute (n.d.d) *Download CGAP Data* [Online]. Available: http://cgap.nci.nih.gov/Info/CGAPDownload [Accessed 21 July 2012].

National Cancer Institute (n.d.e) *All About the cDNA xProfiler Tool* [Online]. Available: http://cgap.nci.nih.gov/Tissues/cDNAxProfilerHowTo [Accessed 21 July 2012].

National Center for Biotechnology Information (n.d.a) *UniGene* [Online]. Available: http://www.ncbi.nlm.nih.gov/unigene [Accessed 20 July 2012].

National Center for Biotechnology Information (n.d.b) *Probe* [Online]. Available: http://www.ncbi.nlm.nih.gov/probe [Accessed 20 July 2012].

National Center for Biotechnology Information (n.d.c) *Gene Expression Omnibus* [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/ [Accessed 20 July 2012].

National Center for Biotechnology Information (n.d.d) *dbEST* [Online]. Available: http://www.ncbi.nlm.nih.gov/dbEST/ [Accessed 21 July 2012].

National Center for Biotechnology Information (n.d.e) *Digital Differential Display* [Online] Available: http://www.ncbi.nlm.nih.gov/UniGene/ddd.cgi [Accessed 21 July 2012].

Neufeld, G., Cohen, T., Gengrinovitch, S., Poltorak, Z. (1999) Vascular endothelial growth factor (VEGF) and its receptors. *FASEB Journal.* **13** (1) pp. 9 – 22.

Ohmori, Y., Schreiber, R.D., Hamilton, T.A. (1997) Synergy between interferon-gamma and tumor necrosis factor-alpha in transcriptional activation is mediated by cooperation between signal transducer and activator of transcription 1 and nuclear factor kappaB. *The Journal of Biological Chemistry.* **272** (23) pp. 14,899 – 14,907.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, 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.

Ootsuka, S., Asami, S., Sasaki, T., Yoshida, Y., Nemoto, N., Shichino, H., Chin, M., Mugishima, H., Suzuki, T. (2008) Useful markers for detecting minimal residual disease in cases of neuroblastoma. *Biological & Pharmaceutical Bulletin.* **31** (6) pp. 1,071 – 1,074.

Panja, S., Saha, S., Jana, B., Basu, T. (2006) Role of membrane potential on artificial transformation of *E. coli* with plasmid DNA. *Journal of Biotechnology.* **127** (1) pp. 14 – 20.

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

Pariset, L., Chillemi, G., Bongiorni, S., Romano Spica, V., Valentini, A. (2009) Microarrays and high-throughput transcriptomic analysis in species with incomplete availability of genomic sequences. *New Biotechnology.* **25** (5) pp. 272 – 279.

Peterson, L.A., Brown, M.R., Carlisle, A.J., Kohn, E.C., Liotta, L.A., Emmert-Buck, M.R., Krizman, D.B. (1998) An improved method for construction of directionally cloned cDNA libraries from microdissected cells. *Cancer Research.* **58** (23) pp. 5,326 – 5,328.

Petersson, S., Shubbar, E., Yhr, M., Kovacs, A., Enerbäck, C (2011) Loss of ICAM-1 signaling induces psoriasin (S100A7) and MUC1 in mammary epithelial cells. *Breast Cancer Research and Treatment.* **125** (1) pp. 13 – 25.

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** (1) p. 22.

Riggins, G.J., Strausberg, R.L. (2001) Genome and genetic resources from the Cancer Genome Anatomy Project. *Human Molecular Genetics.* **10** (7) pp. 663 – 667.

Robinson, M.D., Smyth, G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics.* **23** (21) pp. 2,881 – 2,887.

Ruijter, J.M., Van Kampen, A.H., Baas, F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiological Genomics.* **11** (2) pp. 37 – 44.

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

Salahshor S, Goncalves J, Chetty R, Gallinger S, Woodgett JR (2005) Differential gene expression profile reveals deregulation of pregnancy specific beta1 glycoprotein 9 early during colorectal carcinogenesis. *BMC Cancer.* **27** (5) p. 66.

Salama, P., Platell, C. (2009) Colorectal cancer stem cells. *ANZ Journal of Surgery.* **79** (10) pp. 697 – 702.

Sasaki, Y.F., Ayusawa, D., Oishi, M. (1994) Construction of a normalized cDNA library by introduction of a semi-solid mRNA-cDNA hybridization system. *Nucleic Acids Research.* **22** (6) pp. 987 – 992.

Schaaf, G.J., van Ruissen, F., van Kampen, A., Kool, M., Ruijter, J.M. (2008) Statistical comparison of two or more SAGE libraries: one tag at a time. *Methods in Molecular Biology.* **387** pp. 151 – 168.

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.

Shen, D., He, J., Chang, H.R. (2005) In silico identification of breast cancer genes by combined multiple high throughput analyses. *International Journal of Molecular Medicine.* **15** (2) pp. 205 – 212.

Sher, Y.P., Shih, J.Y., Yang, P.C., Roffler, S.R., Chu, Y.W., Wu, C.W., Yu, C.L., Peck, K. (2005) Prognosis of non-small cell lung cancer patients by detecting circulating cancer cells in the peripheral blood with multiple marker genes. *Clinical Cancer Research.* **11** (1) pp. 173 – 179.

Shostak, K., Labunskyy, V., Dmitrenko, V., Malisheva, T., Shamayev, M., Rozumenko, V., Zozulya, Y., Zehetner, G., Kavsan, V. (2003) HC gp-39 gene is upregulated in glioblastomas. *Cancer Letters.* **198** (2) pp. 203 – 210.

Shmulevich, I., Zhang, W. (2002) Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics.* **18** (4) pp. 555 – 565.

Silveira, N.J., Varuzza, L., Machado-Lima, A., Lauretto, M.S., Pinheiro, D.G., Rodrigues, R.V., Severino, P., Nobrega, F.G. Head and Neck Genome Project GENCAPO, Silva, W.A. Jr., de B Pereira, C.A., Tajara, E.H. (2008) Searching for molecular markers in head and neck squamous cell carcinomas (HNSCC) by statistical and bioinformatic analysis of larynx-derived SAGE libraries. *BMC Medical Genomics.* **1** (56).

Simon, S.A., Zhai, J., Nandety, R.S., McCormick, K.P., Zeng, J., Mejia, D., Meyers, B.C. (2009) Short-read sequencing technologies for transcriptional analyses *Annual Review of Plant Biology.* **60** pp. 305 – 333.

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

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., Efstratiadis, A. (1994) Construction and characterization of a normalized cDNA library. *Proceedings of the National Academy of Sciences of the United States of America* **91** (20) pp. 9,228 – 9,232.

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

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., Van de Vijver, M.J., Bergh, J., Piccart, M., Delorenzi, M (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*. **98** (4) pp. 262 – 272.

Stewart, R.J., Kashour, T.S., Marsden, P.A. (1996) Vascular endothelial platelet endothelial adhesion molecule-1 (PECAM-1) expression is decreased by TNF-alpha and IFN-gamma. Evidence for cytokine-induced destabilization of messenger ribonucleic acid transcripts in bovine endothelial cells. *Journal of Immunology.* **156** (3) pp. 1,221 – 1,228.

Strausberg, R.L. (2001) The Cancer Genome Anatomy Project: new resources for reading the molecular signatures of cancer. *The Journal of Pathology.* **195** (1) pp. 31 – 40.

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** pp. 156 – 164.

Takei, Y., Kadomatsu, K., Yuzawa, Y., Matsuo, S., Muramatsu, T. (2004) A small interfering RNA targeting vascular endothelial growth factor as cancer therapeutics. *Cancer Research.* **64** (10) pp. 3,365 – 3,370.

Tanaka, O., Shinohara, H., Oguni, M., Yoshioka, T. (1995) Ultrastructure of developing muscle in the upper limbs of the human embryo and fetus. *The Anatomical Record.* **241** (3) pp. 417 – 424.

Thygesen, H.H., Zwinderman, A.H. (2006) Modeling Sage data with a truncated gamma-Poisson model. *BMC Bioinformatics.* **7** (157).

Troncone, G., Malapelle, U., Cozzolino, I., Palombini, L.. (2010) KRAS mutation analysis on cytological specimens of metastatic colo-rectal cancer. *Diagnostic Cytopathy.* **38** (12) pp. 869 – 873.

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 osteoblast differentiation resulting in the identification of novel markers for bone development. *Journal of Bone and Mineral Research.* **17** (12) pp. 2,106 – 2,118.

van Eijk, R., van Puijenbroek, M., Chhatta, A.R., Gupta, N., Vossen, R.H., Lips, E.H., Cleton-Jansen, A.M., Morreau, H., van Wezel, T. (2010) Sensitive and specific KRAS somatic mutation analysis on whole-genome amplified DNA from archival tissues. *The Journal of Molecular Diagnostics.* **12** (1) pp. 27 – 34.

van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* **415** (6,871) pp. 530 – 536.

Visbal, A.L., Williams, B.A., Nichols, F.C. 3rd., Marks, R.S., Jett. J.R., Aubry, M.C., Edell, E.S., Wampfler, J.A., Molina, J.R., Yang, P. (2004) Gender differences in non-small-cell lung cancer survival: an analysis of 4,618 patients diagnosed between 1997 and 2002 *The Annals of Thoracic Surgery.* **78** (1) pp. 209 – 215.

Wang, J., Liang, P. (2003) DigiNorthern, digital expression analysis of query genes based on ESTs. *Bioinformatics.* **19** (5) pp. 653 – 654.

Weinberg, R.A. (2007) *The Biology of Cancer.* Garland Science: New York, U.S.A.

Wu, G., Peng, L., Jin, F.S., Li, Q.S. (2006) [High throughput screening and analysis of prostate cancer-related genes through mining databases] [Article in Chinese]. *Ai Zheng.* **25** (5) pp. 645 – 650.

Yang, J., Zhao, J.J., Zhu, Y., Xiong, W., Lin, J.Y., Ma, X. (2008) Identification of candidate cancer genes involved in human retinoblastoma by data mining. *Child's Nervous System.* **24** (8) pp. 893 – 900.

Yip, C.M., Hsu, S.S., Chang, N.J., Wang, J.S., Liao, W.C., Chen, J.Y., Liu, S.H., Chen, C.H. (2009) Primary vaginal extraosseous Ewing sarcoma/primitive neuroectodermal tumor with cranial metastasis. *Journal of the Chinese Medical Association.* **72** (6) pp. 332 – 335.

Yousef, G.M., Yacoub, G.M., Polymeris, M.E., Popalis, C., Soosaipillai, A., Diamandis, E.P Diamandis (2004a) Kallikrein gene downregulation in breast cancer. *British Journal of Cancer.* **90** (1) pp. 167 – 172.

Yousef, G.M., Borgoño, C.A., Popalis, C., Yacoub, G.M., Polymeris, M.E., Soosaipillai, A., Diamandis, E.P. (2004b) In-silico analysis of kallikrein gene expression in pancreatic and colon cancers. *Anticancer Research.* **24** (1) pp. 43 – 51.

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

Zhou, A., Zhang, F., Chen, J.Y. (2010) PEPPI: a peptidomic database of human protein isoforms for proteomics experiments. *BMC Bioinformatics.* **11** (Supplement 6) p. S7.

Zou, K.H., Tuncali, K., Silverman, S.G. (2003) Correlation and simple linear regression. *Radiology.* **227** (3) pp. 617 – 622.