

Methodology article

Open Access

Congruence of tissue expression profiles from Gene Expression Atlas, SAGEmap and TissueInfo databases

Lukasz Huminiecki*, Andrew T Lloyd and Kenneth H Wolfe

Address: Department of Genetics, Smurfit Institute, University of Dublin Trinity College, Dublin 2, Ireland

Email: Lukasz Huminiecki* - huminiel@tcd.ie; Andrew T Lloyd - atlloyd@tcd.ie; Kenneth H Wolfe - khwolfe@tcd.ie

* Corresponding author

Published: 29 July 2003

Received: 20 January 2003

BMC Genomics 2003, 4:31

Accepted: 29 July 2003

This article is available from: <http://www.biomedcentral.com/1471-2164/4/31>

© 2003 Huminiecki et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Abstract

Background: Extracting biological knowledge from large amounts of gene expression information deposited in public databases is a major challenge of the postgenomic era. Additional insights may be derived by data integration and cross-platform comparisons of expression profiles. However, database meta-analysis is complicated by differences in experimental technologies, data post-processing, database formats, and inconsistent gene and sample annotation.

Results: We have analysed expression profiles from three public databases: Gene Expression Atlas, SAGEmap and TissueInfo. These are repositories of oligonucleotide microarray, Serial Analysis of Gene Expression and Expressed Sequence Tag human gene expression data respectively. We devised a method, Preferential Expression Measure, to identify genes that are significantly over- or under-expressed in any given tissue. We examined intra- and inter-database consistency of Preferential Expression Measures. There was good correlation between replicate experiments of oligonucleotide microarray data, but there was less coherence in expression profiles as measured by Serial Analysis of Gene Expression and Expressed Sequence Tag counts. We investigated inter-database correlations for six tissue categories, for which data were present in the three databases. Significant positive correlations were found for brain, prostate and vascular endothelium but not for ovary, kidney, and pancreas.

Conclusion: We show that data from Gene Expression Atlas, SAGEmap and TissueInfo can be integrated using the UniGene gene index, and that expression profiles correlate relatively well when large numbers of tags are available or when tissue cellular composition is simple. Finally, in the case of brain, we demonstrate that when PEM values show good correlation, predictions of tissue-specific expression based on integrated data are very accurate.

Background

High-throughput expression profiling is a major tool for functional annotation of sequenced genomes. SAGEmap [1] and NCI60 cDNA microarray [2] datasets have been incorporated into the Ensembl and UCSC human genome browsers. Large-scale expression data have also been used to construct a transcriptional profile of adult skeletal mus-

cle [3]; to identify genes preferentially expressed in prostate [4], granulocytes [5] and thyroid [6]; to discover markers of pathological states such as cancer [7], and for whole genome structure analysis such as identification of clusters of similarly expressed genes [8,9].

Numerous public databases have been created to store large-scale expression data. These databases are, however, heterogeneous in format, sample annotation, and statistical post-processing of experimental results. Furthermore, the different databases store data from different experimental technologies. A cross-platform comparison between cDNA and oligonucleotide microarrays has demonstrated that data generated by different platforms may have poor correlation [10]. We investigated whether three public gene expression databases: the Gene Expression Atlas published by Su et al. [11], SAGEmap [1] and TissueInfo [12] (representing respectively oligonucleotide microarray, Serial Analysis of Gene Expression – SAGE, and Expressed Sequence Tags – ESTs) are internally consistent and complementary when used to compare expression profiles of human genes within six normal tissue categories, which were represented in the three databases.

The scope of this study is integration of expression data and meta-analysis of datasets. Our aim is to evaluate the congruence of publicly available expression datasets. It was not our intention to provide a technical benchmarking of different expression platforms, because such a benchmarking exercise would require controlled microarray, SAGE and EST experiments starting with identical RNA samples derived from the same tissue specimen, and would be beyond the resources of most laboratories.

Results and discussion

Preferential Expression Measure

We devised a Preferential Expression Measure (PEM) to score differential expression of genes in tissues. PEM describes the expression of a gene in a given tissue in relation to its expression in all tissues. Therefore, PEM controls for the fact that some genes are highly expressed across many tissues (housekeeping genes), and has the virtue of reporting a negative value for under-expressed genes, and a positive value for over-expressed genes. Large positive PEM scores for a gene in a particular tissue indicate that the gene is unusually highly expressed in that tissue, relative to its expression in other tissues. Large negative PEM scores indicate repression of a gene in a tissue.

For SAGEmap and TissueInfo we define PEM as $\log_{10}(o/e)$, where o is the observed SAGE or EST tag count for a gene in a given tissue, and e the expected tag count under the null hypothesis of uniform expression in all tissues. e is calculated as $(G * N/T)$ where G is the total number of tags ascribed to the gene, N is the total number of tags for the tissue, and T is the total number of tags in the dataset.

For example, carbonic anhydrase XI (UniGene cluster Hs.22777) is linked to a SAGE tag GTCGCTGAGA. This tag occurs 132 times in all normal tissue SAGEmap libraries.

The total number of SAGE tags in the normal brain tissue category is 326,481, and the total number of tags in all normal tissue categories is 1,077,231. Therefore, if the distribution of this tag was uniform across all libraries, it would be expected to occur ~ 40 times ($132 * 326,481 / 1,077,231$) in brain libraries (expected count). The actual count in brain libraries is 124 (observed count) bringing the PEM value for this tag in brain to $\log_{10}(o/e) = \log_{10}(132/40) = 0.49$.

For microarray experiments, where the raw data is a continuous variable (a relative intensity of a gene-specific fluorescent signal) as opposed to a tag count, we defined PEM as $\log_{10}(S/A)$, where S is the specific tissue signal for a gene and A is the arithmetic mean signal for the gene across all tissues.

Reproducibility and internal consistency

A major concern with all types of high-throughput expression data is reproducibility and internal consistency. To investigate this, for each of the three databases, we compared replicate experiments (where available) or equivalent tissues and measured the Pearson correlation coefficient for relative fluorescent signal intensity values assigned to the same probe in a pair of experiments. Gene Expression Atlas oligonucleotide microarray data had good correlation between repeat hybridizations of the same RNA sample (mean $r^2 = 0.94 \pm 0.04$; $N = 45$ pairwise comparisons) and, as expected, slightly lower values for repeat hybridizations of different RNA preparations from the same tissue type (mean $r^2 = 0.87 \pm 0.06$; $N = 17$ pairwise comparisons).

Internal correlations for SAGEmap and TissueInfo were lower. Counts of identical tags were correlated in pairwise SAGEmap library comparisons yielding mean r^2 values of 0.51 ± 0.25 , $N = 15$ for brain; 0.26 ± 0.08 , $N = 6$ for prostate; 0.89 , $N = 2$ for vascular endothelium; 0.27 , $N = 2$ for ovary; 0.83 , $N = 2$ for kidney; and 0.97 , $N = 2$ for pancreas.

For TissueInfo, we included tumour tissues to maximize the size of datasets and, therefore, minimize the sampling error. Counts of EST tags linked to the same UniGene cluster in pools of libraries were correlated. Comparison of brain TissueInfo libraries comprising 154,214 EST tags from normal tissue and 111,317 from tumour tissue, revealed an r^2 of only 0.27.

Correlations among tissue profiles from Gene Expression Atlas, SAGEmap and TissueInfo

To compare different databases we grouped libraries into higher-level tissue categories, and calculated Pearson correlation coefficients for PEM scores for categories that were represented in all three databases. Tumour tissue libraries were not used in any of the interdatabase

Table 1: r^2 values for Pearson correlations between PEM scores computed from Gene Expression Atlas (GEA), SAGEmap and TissueInfo expression profiles for six human tissue categories, for which data were available in the three databases.

Tissue	TissueInfo vs. SAGEmap				GEA vs. SAGEmap				GEA vs. TissueInfo			
	Genes ^a	r^2	P ^b	Signs ^c	Genes ^a	r^2	P ^b	Signs ^c	Genes ^a	r^2	P ^b	Signs ^c
Brain	346	0.59	<0.01	92%	479	0.49	<0.01	84%	1431	0.43	<0.01	77%
Prostate	58	0.36	<0.01	76%	237	0.09	<0.01	59%	266	0.19	<0.01	63%
Vascular endothelium	13	0.29	0.6	92%	150	0.37	<0.01	73%	23	0.05	0.29	39%
Ovary	10	0.11	0.35	30%	175	0.03	0.2	49%	36	0.03	0.28	69%
Kidney	103	0.01	0.36	34%	140	0.00	0.78	50%	508	0.11	<0.01	55%
Pancreas	21	0.02	0.55	62%	141	0.00	0.59	48%	82	0.02	0.19	55%

^a Number of genes compared. SAGEmap and TissueInfo data were only used for genes whose PEM scores were significant by χ^2 test ($P = 0.05$), and only tags that could be mapped unambiguously to UniGene were used. ^b Two-tailed significance values for Pearson correlation coefficients. ^c Percentage of genes whose PEM scores have the same sign for the two methods being compared.

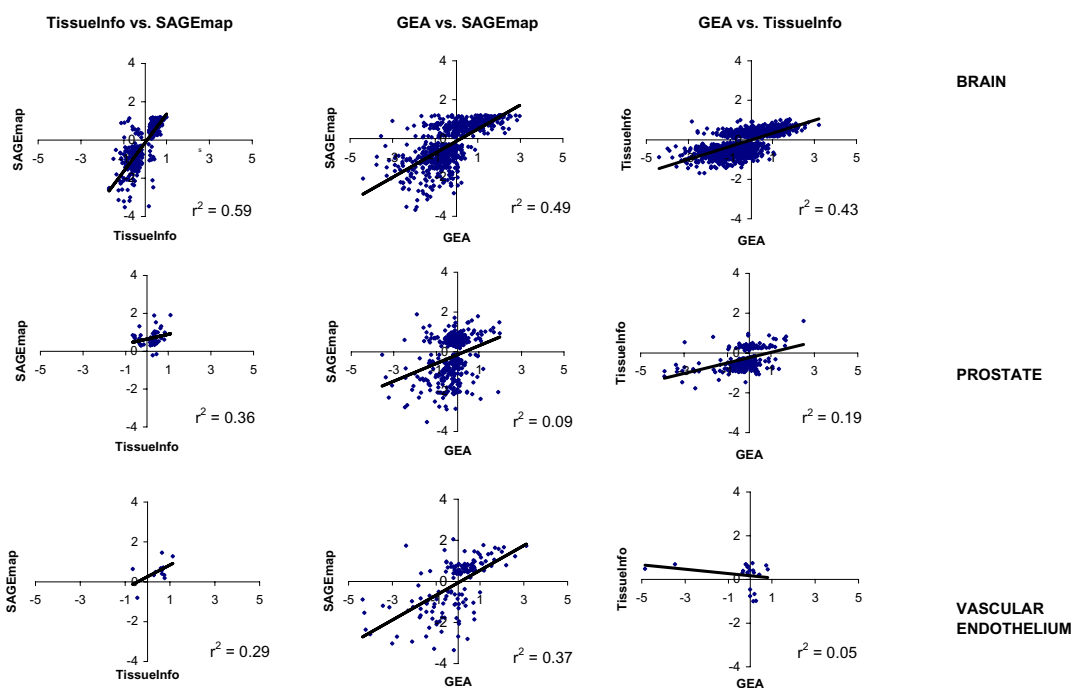


Figure 1
Correlations between the three databases Correlations between Gene Expression Atlas (GEA), SAGEmap, and Tissue-Info preferential expression measures (PEM) for brain, prostate and vascular endothelium. Trend lines for linear regression and the corresponding r^2 values are shown for each pairwise comparison.

Table 2: Number of SAGE and EST tags in tissue categories derived from SAGEmap and TissueInfo.

	Brain	Prostate	Kidney	Vascular endothelium	Ovary	Pancreas
SAGE	326,481	150,116	67,923	110,460	96,911	64,577
EST	154,214	51,299	60,188	8,435	12,587	23,545
SAGE+EST	480,695	201,415	128,111	118,895	109,498	88,122

comparisons. UniGene was used as the common gene index to link entries from the three databases. There were six tissues available for comparison (Table 1). Significant positive correlations were found for brain, prostate and vascular endothelium (Figure 1) but not for ovary, kidney, and pancreas.

Because of sampling error, the total number of sequenced tags is a key determinant of reliability of SAGE and EST expression profiles. In our analysis, brain is the tissue category with the highest number of both SAGE and EST tags (Table 2) and accordingly shows the strongest correlations of PEM values among oligonucleotide microarray, SAGEmap, and TissueInfo (Table 1). Prostate, the tissue category with the second highest total number of tags, showed good correlation between SAGEmap and TissueInfo, but the correlation between the tag-based databases and oligonucleotide microarray data was weaker (r^2 of 0.09 and 0.19 for SAGEmap and TissueInfo respectively).

We suggest that the relative complexity of a tissue is another important factor for comparison of high-throughput expression profiles deposited in public databases. Complex tissues consist of many cell types that may be mixed in different proportions in different samples. Expression profiles obtained from these tissues will be heterogeneous and consequently require a large number of tags and repeated hybridizations to detect signal in the noise. This is shown by the relatively strong correlations seen in SAGEmap/TissueInfo and SAGEmap/microarray correlations for vascular endothelium despite the small number of tags available (in comparison to brain, about 3-fold fewer SAGE tags and 18-fold fewer EST tags). SAGEmap and dbEST/TissueInfo vascular endothelium libraries are derived from *in vitro* cultures or homogenous primary isolates of vascular endothelial cells [13].

Correlations of PEM scores obtained from the three databases for kidney, pancreas, and ovary had very low r^2 values (Table 1). However, measuring the correlation coefficient of PEM scores is a very rigorous test of congruence, because it not only requires that two datasets being compared identify the same sets of genes as being over- or under-expressed in the tissue, but also that the genes deviate from uniform expression to a similar degree. This may

be an unrealistic expectation; for example, it is apparent from Figure 1 that PEM scores for SAGEmap in brain never exceed +1.19 (i.e., approximately 15-fold over-representation), which is probably due to the limited size of the SAGE database. A simpler test is to divide the plots into quadrants and measure the fraction of genes for which the two methods being compared give PEM scores with the same sign. For example, for brain tissue the PEM scores from SAGEmap and TissueInfo data are both positive for 134 genes, both negative for 184 genes, and in disagreement for only 28 genes (i.e., one method identifies the gene as significantly over-expressed in brain, and the other identifies it as significantly under-expressed) (Figure 1). Across all six tissues and three datasets, the PEM signs are in agreement for more than 50% of genes (the null expectation) in 13 of the 18 comparisons made (Table 1).

Tissue-specific genes

We showed previously that integrating two expression databases (SAGEmap and dbEST) provided accurate predictions of vascular endothelium-specific expression that were verified experimentally, whereas neither method alone was reliable [13]. Similarly, genes for which Gene Expression Atlas, SAGEmap and TissueInfo analyses all yield positive PEM scores should be very strong candidates for being tissue-specific. To test this hypothesis, we integrated PEM scores from the three types of expression profile to search for putative brain-specific genes. The 50 top-scoring genes were investigated through LocusLink and OMIM and of these 47 proved to be previously known brain-specific or preferentially brain expressed genes (Table 3). Three genes were novel, one of which was highly similar to a known brain-specific inorganic phosphate co-transporter. This confirms the very high reliability of integrated predictions of tissue-specific expression. The bottom-scoring ten genes for brain are also listed in Table 1. They have negative PEM values indicating under-representation in the brain transcriptome. Four of these genes are epithelial markers (EPCAM, KRT5, KRT18, and KRT19).

The top ten prostate-specific genes are listed in Table 4. As with the analysis for brain, combining Gene Expression Atlas, SAGEmap and TissueInfo data appears to identify prostate-specific genes successfully. Five genes were previ-

Table 3: Top 50 brain-specific and bottom ten brain underexpressed genes (in italics) identified by integrated PEM scores from Gene Expression Atlas, SAGEmap and TissueInfo.

%PEM ^a	UniGene cluster	LocusLink symbol	Description and references
76%	Hs.6535	n/a	novel cDNA clone, FLJ33742 fis, highly similar to Homo sapiens BNPI mRNA for brain-specific Na-dependent inorganic phosphate cotransporter (GenBank AK091061)
76%	Hs.143535	CAMK2A	calcium/calmodulin-dependent protein kinase, implied in control of fear and aggression [18]
75%	Hs.78854	ATP1B2	ATPase, Na ⁺ /K ⁺ transporting [19,20]
74%	Hs.194301	MAPIA	microtubule-associated protein, expressed strongly in brain, spinal cord and weakly in muscle [21]
73%	Hs.84389	SNAP25	synaptosomal associated protein; one of SNARE proteins required for neuronal exocytosis [22]
72%	Hs.2288	VSNLI	visinin like 1; known to be expressed in neuronal cells in retina [23]
72%	Hs.1787	PLPI	lipophilin – the primary constituent of myelin [24]
72%	Hs.296184	GNAOI	G protein alpha activating activity polypeptide O; multiple neurological abnormalities in knock-out mice [25]; cloned from brain [26]
71%	Hs.75819	GPM6A	neuronal membrane glycoprotein M6A; similar to myelin proteolipid protein [27]
71%	Hs.69547	MBP	myelin basic protein [28]
70%	Hs.74565	APLP1	amyloid beta (A4) precursor-like protein 1 [29]
69%	Hs.6139	SYNGR1	synaptogyrin 1 [30]
69%	Hs.154679	SYTI	synaptotagmin 1 [31]
68%	Hs.20912	APCL	brain-specific adenomatous polyposis coli (APC) homologue [32]
66%	Hs.5422	GPM6B	neuronal membrane glycoprotein M6B, highly similar to the myelin proteolipid protein [27]
66%	Hs.79000	GAP43	neuromodulin, implied in axon guidance [33]
65%	Hs.22777	CA11	carbonic anhydrase XI; brain-specific carbonic anhydrase isosyme [34]
65%	Hs.239356	STXB1	syntaxin binding protein 1 [35]
65%	Hs.182859	LFG	lifeguard (alias neuromembrane protein 35) [36]
65%	Hs.7979	SV2	synaptic vesicle glycoprotein 2 [37]
64%	Hs.322430	NDRG4	ndrg family member 4, expressed in brain and heart [38]
64%	Hs.74554	SYTI1	synaptotagmin 11; identified by genomic screen with highly conserved synaptotagmin motif [39]
64%	Hs.82749	TM4SF2	transmembrane 4 superfamily member 2; contains a trinucleotide repeat [40], implied in X linked mental retardation [41], implied in axon guidance
64%	Hs.169401	APOE	apolipoprotein E; implied in Alzheimer's disease [42]
63%	Hs.166161	DNMI	dynamitin 1 – brain expression at least 30-fold higher than in other tissues [43]
62%	Hs.6164	NECLI	nectin-like protein 1; alias brain immunoglobulin receptor precursor (unpublished, GenBank NM_021189)
62%	Hs.76888	INA	internexin – neuronal intermediate filament protein alpha [44]
61%	Hs.146580	ENO2	enolase 2 – neuronal-specific enolase [45]
61%	Hs.90005	STMN2	stathmin-like 2; expressed in neuronal growth cones [46]
61%	Hs.226133	GAS7	growth arrest-specific 7; expressed in neurons and growth-arrested fibroblasts [47]
60%	Hs.155524	PNUTL2	septin 4; human orthologue of the mouse h5 brain protein [48]; involved in cell-division control; expressed in brain and, an alternatively spliced form, in heart [49]
60%	Hs.117546	NNAT	neuronatin; known to contain a 5' neurospecific silencer element which controls brain-specific expression [50]; possible role in brain development
60%	Hs.155247	ALDOC	brain-type aldolase; fructose-bisphosphate aldolase C [51]
60%	Hs.7357	CLIPR-59	microtubule-binding protein involved in Golgi targeting; known to be preferentially expressed in brain [52]
60%	Hs.74583	SPOCK2	sparc/osteonectin like; testican 2; calcium-binding proteoglycan specific to brain; expressed in many neuronal cell types in olfactory bulb, cerebral cortex, thalamus, hippocampus, cerebellum, and medulla [53]
60%	Hs.80395	MAL	hydrophobic integral membrane lipoprotein; a component of myelin [54]
60%	Hs.6349	BC008967	novel protein BC008967
60%	Hs.6755	RPIP8	RaP2 interacting protein; belongs to Ras family; expressed principally in brain [55]
60%	Hs.75090	SPRING	involved in synaptic exocytosis; expressed only in brain [56]
60%	Hs.90063	NCALD	neurocalcin delta; calcium-binding protein; expressed preferentially in brain [57]
58%	Hs.83384	S100B	calcium-binding protein S100; expressed in glia, astrocytes and neurons [58]
58%	Hs.104925	ENC1	ectodermal-neural cortex; p53 induced gene; expressed preferentially in brain, particularly in amygdala and hippocampus [59]
58%	Hs.7782	PNMA2	onconeural antigen MA2; expressed in brain and testis [60]
58%	Hs.8526	B3GNT6	beta-1,3-N-acetylglucosaminyltransferase 6; expressed preferentially in foetal and adult brain [61]
56%	Hs.125359	THY1	Thy-1 cell surface antigen; expressed in brain and on some T cells; neuronal expression pattern changes in development suggesting role in neurogenesis [62]

Table 3: Top 50 brain-specific and bottom ten brain underexpressed genes (in italics) identified by integrated PEM scores from Gene Expression Atlas, SAGEmap and TissueInfo. (Continued)

56%	Hs.194534	VAMP2	synaptobrevin 2; one of proteins involved in fusion of synaptic vesicles with the presynaptic membrane [63]
56%	Hs.31463	ELMO1	engulfment and cell motility 1; orthologue of <i>C. elegans</i> gene <i>ced-12</i> ; has two isoforms – 4.4 kb expressed ubiquitously and 2.4 kb expressed only in brain [64]
55%	Hs.3763	APBB1	amyloid beta (A4) precursor protein-binding; expressed preferentially in brain [65]
55%	Hs.286055	CHN2	chimerin 2; expressed preferentially in cerebellum in granule cells [66]
55%	Hs.12305	DKFZP566B183	novel protein DKFZP566B183
-50%	Hs.80988	COL6A3	<i>collagen type VI alpha 3 chain; type 3 alpha chain of a beaded filament collagen found in most connective tissues [67]</i>
-50%	Hs.79914	LUM	<i>lumican, a keratan sulfate proteoglycan known to be abundant in the corneal stroma and collagenous matrices of the heart, skeletal muscle, aorta, and intervertebral discs [68]</i>
-51%	Hs.38972	TSPAN-1	<i>tetraspan 1 [69]</i>
-53%	Hs.692	EPCAM	<i>epithelial cellular adhesion molecule; expressed only on epithelial cells [70]</i>
-56%	Hs.297681	SERPINA1	<i>antitrypsin, primarily expressed in liver [71]</i>
-57%	Hs.5372	CLDN4	<i>claudin 4; expressed in kidney, small intestine, lung, heart, liver, and skeletal muscle; no expression was found in brain or spleen [72]</i>
-57%	Hs.81892	KIAA0101	<i>identified in the Kazusa large-scale cDNA sequencing project [73]; northern blot revealed expression in many tissues but signal is completely absent from brain http://www.kazusa.or.jp/huge/gfpage/KIAA0101/</i>
-58%	Hs.195850	KRT5	<i>keratin 5; expressed in many epithelia including mammary epithelial cells and squamous epithelia lining the upper digestive tract [74]</i>
-59%	Hs.65114	KRT18	<i>keratin 18 [75]; expressed in many epithelia</i>
-60%	Hs.182265	KRT19	<i>keratin 19; found predominantly in the periderm – the transient layer that surrounds the developing epidermis [76]</i>

^a Integrated PEM score, calculated as the average of $PEM_{GEA}/PEM_{GEA_{max}}$, $PEM_{SAGEMAP}/PEM_{SAGEMAP_{max}}$, $PEM_{TISSUEINFO}/PEM_{TISSUEINFO_{max}}$, where "max" refers to the maximum PEM value encountered in the tissue. This weighting scheme was used because brain microarray PEM values were several fold higher than those for SAGE and EST ($PEM_{GEA_{max}} = 5.21$, $PEM_{SAGEMAP_{max}} = 1.19$, $PEM_{TISSUEINFO_{max}} = 1.15$)

ously known to be expressed specifically by prostate epithelium (MSMB, ACPP, KLK3, AMD1 and RDH11). The other five are muscle-specific genes. This is not unexpected since the fibromuscular stroma surrounding prostatic glands is known to account for about half of the volume of the prostate [14].

Conclusions

We show here that data from Gene Expression Atlas, SAGEmap and TissueInfo can be integrated when libraries are grouped into higher-level tissue categories and genes are mapped between datasets using the UniGene gene index. After integration, expression profiles between tissue categories represented in multiple databases can be compared using a measure of differential expression, PEM. Internal consistency of PEM scores is high for Gene Expression Atlas but poorer in the tissue categories derived from SAGEmap and TissueInfo. Between datasets, PEM values correlate relatively well when large numbers of SAGE and EST tags are available (brain, prostate) or when tissue cellular composition is simple (vascular endothelium), but in the other tissue categories examined the correlation is low. The usefulness of the integrated dataset is demonstrated by the accuracy with which brain- and prostate-specific genes could be identified. This suggests that similar accuracy could be achieved for other tis-

sues if oligonucleotide microarray, SAGE, and EST experiments were performed on RNAs from the same well-annotated tissue samples of relatively simple cellular composition. This would also enable direct comparison and benchmarking of expression profiles obtained using different technologies. The integrated dataset, code and database schema are available on request from L.H.

Methods

Gene Expression Atlas

Affymetrix U95A oligonucleotide microarray [15] data for 12,587 consensus probes in 101 human tissue samples were downloaded from the Gene Expression Atlas website http://expression.gnf.org/data_public_U95.gz. Image analysis and normalization has been performed using Genechip 3.2 software (Affymetrix, Santa Clara, CA) by Su et al. [11]. The libraries were grouped into 47 higher-level tissue categories (averaging across duplicates and triplicates). Tumour libraries were grouped as separate to normal tissues and not used for further analysis.

SAGEmap

Tag frequencies for 132 SAGEmap libraries [1] were obtained from the project ftp site <ftp://ftp.ncbi.nih.gov/pub/sage>. After downloading, the libraries were grouped into 15 higher-level tissue categories (brain, prostate, kid-

Table 4: Top ten prostate-specific genes identified by integrated PEM scores from Gene Expression Atlas, SAGEmap and TissueInfo.

%PEM ^a	UniGene cluster	LocusLink symbol	Description and references
95%	Hs.183752	MSMB	beta microseminoprotein; synthesized by the epithelial cells of the prostate gland and secreted into the seminal plasma [77]
94%	Hs.1852	ACPP	prostatic acid phosphatase [78]
82%	Hs.171995	KLK3	kallikrein 3; prostate specific antigen (PSA) [79]
73%	Hs.78344	MYH11	smooth muscle-specific myosin heavy polypeptide 11 [80]
70%	Hs.300772	TPM2	tropomyosin 2 (beta); binds actin [81]
55%	Hs.195464	FLNA	filamin A; binds actin [82]
54%	Hs.75777	TAGLN	transgelin; smooth muscle-specific actin binding protein [83]
53%	Hs.262476	AMD1	S-adenosylmethionine decarboxylase 1; key enzyme in the synthesis of seminal polyamines such as spermine, spermidine [84]
53%	Hs.9615	MYL9	myosin, smooth muscle-specific regulatory light polypeptide 9 [85]
48%	Hs.179817	RDH11	androgen-regulated short-chain dehydrogenase/reductase 1 [86]

^a Prostate PEMGEA_{max} = 4.30, PEMSAGEMAP_{max} = 1.97, and PEMTISSUEINFO_{max} = 1.61

ney, colon, pancreas, ovary, breast, skin, peritoneum, stomach, blood, lung, liver, heart and vascular endothelium) and annotated according to their disease status (91 tumour tissues, 37 normal tissue libraries and four immortalised cell lines). Unless stated otherwise, only normal tissue libraries (total of 1,077,231 tags) were used for further analysis.

TissueInfo

The TissueInfo [12] database links ESTs from dbEST to the tissue where they are expressed by assigning a tissue-key to each EST entry. The dataset was downloaded from the TissueInfo website http://icb.mssm.edu/tissueinfo/local_inst.xml, which is updated daily, on 27 February 2002. Our aim was to group ESTs into major tissues with sufficient numbers of ESTs for statistically meaningful analysis. Therefore, some TissueInfo categories were grouped together into higher-level tissue categories. For example hypothalamus was also annotated as brain. We excluded ESTs that were annotated as mitochondrial libraries, tumour libraries, or multiple tissues. This procedure resulted in 48 higher-level tissue categories, 27 of which had EST counts in excess of 10,000 (the number of ESTs we accepted as the threshold for dataset size). We also included vascular endothelium (8,435 ESTs) because of the high number of SAGE tags available for this category (110,460) and its simple cellular composition.

Mapping

Six higher-level tissue categories were available for comparison between Gene Expression Atlas, SAGEmap, and TissueInfo (Table 1). An integrated MySQL database was set up and gene identifiers from the UniGene gene index were used to link expression data from the three datasets. Only UniGene entries with more than one EST (62,008 entries) were considered as valid clusters. UniGene mapping was provided as part of both SAGEmap and Tissue-

Info datasets. To map oligonucleotide microarray probes, probe consensus sequences were searched against longest representative sequence of each UniGene cluster using BLASTN with the cutoff parameter $E = 10e-20$. Alignments were accepted if the percentage identity was higher than 94% and length was at least 99 bp or 90% of the length of the query, or when percentage identity was 100% and length more than 49 bp [11].

When linking UniGene clusters to SAGEmap data, UniGenes mapping to more than SAGE tag sequence were excluded (25% of clusters). The size of the UniGene dataset linked to SAGEmap was, therefore, reduced from 21,224 to 15,872 clusters. Similarly, UniGene clusters mapping to more than one Affymetrix probe were excluded (22% of clusters). The size of the UniGene dataset linked to Gene Expression Atlas data was, therefore, reduced from 9,402 to 7,375 clusters. The overlaps between UniGenes linked to GEA and SAGEmap, GEA and TissueInfo, and SAGEmap and TissueInfo were 2742, 6758 and 15872 clusters respectively. Therefore:

- 63% and 8% of UniGene clusters linked to GEA could not be linked to SAGEmap and TissueInfo respectively;
- 83% and 0% of UniGene clusters linked to SAGEmap could not be linked to GEA and TissueInfo respectively;
- 74% and 89% of UniGene clusters linked to TissueInfo could not be linked to SAGEmap and GEA respectively.

Statistics

Gene expression can be estimated by counting gene-specific SAGE and EST tags but, as a stochastic process, these estimates are subject to sampling error. The size of the sampled dataset can have a dramatic influence on reliability of the data. Two independent groups have established

χ^2 as the best statistical test to use in tag sampling experiments [16,17]. We excluded from subsequent comparative analysis all tissue/gene data for which the expected number of tags (e) was less than 5, which is the lower limit of reliability for the χ^2 statistic. We used χ^2 to filter the PEM score for both tag-based methods: unless significantly different from expectation (with $P = 0.05$) we concluded that we had no confidence in either the sign or the magnitude of the apparent deviation from uniform expression. A relatively low significance threshold of 0.05 was used to maximize the number of values available for inter-database comparison. Because χ^2 test is only used on a case-by-case basis for preliminary selection of informative datapoints, a multiple testing correction such as Bonferroni correction was not applicable.

To compare the results from any two databases for a given tissue, we plotted the PEM scores for all genes for which data was available from the two experiments, but excluding SAGE and EST data with non-significant χ^2 . We calculated the correlation coefficient (r) and report r^2 , which represents the proportion of the variability in the data that is explained by the correlation.

Database versions

Gene Expression Atlas, Affymetrix human Genechip U95A <http://expression.gnf.org>

SAGEmap, January 2002 <http://www.ncbi.nlm.nih.gov/SAGE/>

TissueInfo, 27 February 2002 <http://icb.mssm.edu/crt/tissueinfowebservice.xml>

UniGene, Build 148 <http://www.ncbi.nlm.nih.gov/UniGene/>

List of abbreviations

NCI60 – set of 60 human cancer cell lines used by the National Cancer Institute for chemotherapeutic drug screens

UCSC – University of California Santa Cruz

cDNA – complementary DNA

SAGE – Serial Analysis of Gene Expression

EST – Expressed Sequence Tag

PEM – Preferential Expression Measure

BLAST – Basic Local Alignment Search Tool

OMIM – Online Mendelian Inheritance in Man

MySQL – open source relational database

Authors' contributions

LH study design, SAGE database, Affymetrix database, correlations, manuscript draft, preparation of figures and tables

AL statistical methods, TissueInfo database, Affymetrix database, manuscript draft

KW study design, coordination, manuscript review

Acknowledgements

This study was supported by Science Foundation Ireland. We would like to thank Dr. Keith Ching, Genome Informatics Group, Novartis Genomics Institute for providing us with mapping of Affymetrix microarray probes to UniGene clusters, and Dr. Greg Singer for helpful discussions.

References

- Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ and Altschul SF: **SAGEmap: a public gene expression resource** *Genome Res* 2000, **10**:1051-1060.
- Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D and Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines** *Nat Genet* 2000, **24**:227-235.
- Bortoluzzi S, d'Alessi F, Romualdi C and Danieli GA: **The human adult skeletal muscle transcriptional profile reconstructed by a novel computational approach** *Genome Res* 2000, **10**:344-349.
- Vasmataz G, Essand M, Brinkmann U, Lee B and Pastan I: **Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis** *Proc Natl Acad Sci U S A* 1998, **95**:300-304.
- Itoh K, Okubo K, Utiyama H, Hirano T, Yoshii J and Matsubara K: **Expression profile of active genes in granulocytes** *Blood* 1998, **92**:1432-1441.
- Moreno JC, Pauws E, van Kampen AH, Jedlickova M, de Vijlder JJ and Ris-Stalpers C: **Cloning of tissue-specific genes using serial analysis of gene expression and a novel computational sub-structure approach** *Genomics* 2001, **75**:70-76.
- St Croix B, Rago C, Velculescu V, Traverso G, Romans KE, Montgomery E, Lal A, Riggins GJ, Lengauer C, Vogelstein B and Kinzler KW: **Genes expressed in human tumor endothelium** *Science* 2000, **289**:1197-1202.
- Caron H, van Schaik B, van der Mee M, Baas F, Riggins G, van Sluis P, Hermus MC, van Asperen R, Boon K, Voute PA, Heisterkamp S, van Kampen A and Versteeg R: **The human transcriptome map: clustering of highly expressed genes in chromosomal domains** *Science* 2001, **291**:1289-1292.
- Lercher MJ, Urrutia AO and Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome** *Nat Genet* 2002, **31**:180-183.
- Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L and Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies** *Bioinformatics* 2002, **18**:405-412.
- Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, Patapoutian A, Hampton GM, Schultz PG and Hogenesch JB: **Large-scale analysis of the human and mouse transcriptomes** *Proc Natl Acad Sci U S A* 2002, **99**:4465-4470.
- Skrabaneck L and Campagne F: **TissueInfo: high-throughput identification of tissue expression profiles and specificity** *Nucleic Acids Res* 2001, **29**:E102-2.
- Huminiecki L and Bicknell R: **In silico cloning of novel endothelial-specific genes** *Genome Res* 2000, **10**:1796-1806.
- Farnsworth WE: **Prostate stroma: physiology** *Prostate* 1999, **38**:60-72.

15. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H and Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays** *Nat Biotechnol* 1996, **14**:1675-1680.
16. Man MZ, Wang X and Wang Y: **POWER_SAGE: comparing statistical tests for SAGE experiments** *Bioinformatics* 2000, **16**:953-959.
17. Romualdi C, Bortoluzzi S and Danieli GA: **Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests** *Hum Mol Genet* 2001, **10**:2133-2141.
18. Chen C, Rainnie DG, Greene RW and Toneyawa S: **Abnormal fear response and aggressive behavior in mutant mice deficient for alpha-calcium-calmodulin kinase II** *Science* 1994, **266**:291-294.
19. Martin-Vasallo P, Dackowski W, Emanuel JR and Levenson R: **Identification of a putative isoform of the Na,K-ATPase beta subunit. Primary structure and tissue-specific expression** *J Biol Chem* 1989, **264**:4613-4618.
20. Pagliusi S, Antonicek H, Gloor S, Frank R, Moos M and Schachner M: **Identification of a cDNA clone specific for the neural cell adhesion molecule AMOG** *J Neurosci Res* 1989, **22**:113-119.
21. Lien LL, Feener CA, Fischbach N and Kunkel LM: **Cloning of human microtubule-associated protein 1B and the identification of a related gene on chromosome 15** *Genomics* 1994, **22**:273-280.
22. Hu K, Carroll J, Fedorovich S, Rickman C, Sukhodub A and Davletov B: **Vesicular restriction of synaptobrevin suggests a role for calcium in membrane fusion** *Nature* 2002, **415**:646-650.
23. Polymeropoulos MH, Ide S, Soares MB and Lennon GG: **Sequence characterization and genetic mapping of the human VSNL1 gene, a homologue of the rat visinin-like peptide RNVP1** *Genomics* 1995, **29**:273-275.
24. Diehl HJ, Schaich M, Budzinski RM and Stoffel W: **Individual exons encode the integral membrane domains of human myelin proteolipid protein** *Proc Natl Acad Sci U S A* 1986, **83**:9807-9811.
25. Jiang M, Gold MS, Boulay G, Spicher K, Peyton M, Brabet P, Srinivasan Y, Rudolph U, Ellison G and Birnbaumer L: **Multiple neurological abnormalities in mice deficient in the G protein Go** *Proc Natl Acad Sci U S A* 1998, **95**:3269-3274.
26. Strathmann M, Wilkie TM and Simon MI: **Alternative splicing produces transcripts encoding two forms of the alpha subunit of GTP-binding protein Go** *Proc Natl Acad Sci U S A* 1990, **87**:6477-6481.
27. Yan Y, Lagenaur C and Narayanan V: **Molecular cloning of M6: identification of a PLP/DM20 gene family** *Neuron* 1993, **11**:423-431.
28. Eylar EH, Brostoff S, Hashim G, Caccam J and Burnett P: **Basic A1 protein of the myelin membrane. The complete amino acid sequence** *J Biol Chem* 1971, **246**:5770-5784.
29. Wasco W, Bupp K, Magendantz M, Gusella JF, Tanzi RE and Solomon F: **Identification of a mouse brain cDNA that encodes a protein related to the Alzheimer disease-associated amyloid beta protein precursor** *Proc Natl Acad Sci U S A* 1992, **89**:10758-10762.
30. Kedra D, Pan HQ, Seroussi E, Fransson I, Guilbaud C, Collins JE, Dunham I, Blennow E, Roe BA, Piehl F and Dumanski JP: **Characterization of the human synaptogyrin gene family** *Hum Genet* 1998, **103**:131-141.
31. Fernandez-Chacon R, Konigstorfer A, Gerber SH, Garcia J, Matos MF, Stevens CF, Brose N, Rizo J, Rosenmund C and Südhof TC: **Synaptotagmin I functions as a calcium regulator of release probability** *Nature* 2001, **410**:41-49.
32. Nakagawa H, Koyama K, Murata Y, Morito M, Akiyama T and Nakamura Y: **APCL, a central nervous system-specific homologue of adenomatous polyposis coli tumor suppressor, binds to p53-binding protein 2 and translocates it to the perinucleus** *Cancer Res* 2000, **60**:101-105.
33. Strittmatter SM, Fankhauser C, Huang PL, Mashimo H and Fishman MC: **Neuronal pathfinding is abnormal in mice lacking the neuronal growth cone protein GAP-43** *Cell* 1995, **80**:445-452.
34. Bellingham J, Gregory-Evans K and Gregory-Evans CY: **Sequence and tissue expression of a novel human carbonic anhydrase-related protein, CARP-2, mapping to chromosome 19q13.3** *Biochem Biophys Res Commun* 1998, **253**:364-367.
35. Pevsner J, Hsu SC and Scheller RH: **n-Sec1: a neural-specific syntaxin-binding protein** *Proc Natl Acad Sci U S A* 1994, **91**:1445-1449.
36. Schweitzer B, Taylor V, Welcher AA, McClelland M and Suter U: **Neural membrane protein 35 (NMP35): a novel member of a gene family which is highly expressed in the adult nervous system** *Mol Cell Neurosci* 1998, **11**:260-273.
37. Bajjalieh SM, Peterson K, Linial M and Scheller RH: **Brain contains two forms of synaptic vesicle protein 2** *Proc Natl Acad Sci U S A* 1993, **90**:2150-2154.
38. Zhou RH, Kokame K, Tsukamoto Y, Yutani C, Kato H and Miyata T: **Characterization of the human NDRG gene family: a newly identified member, NDRG4, is specifically expressed in brain and heart** *Genomics* 2001, **73**:86-97.
39. Craxton M: **Genomic analysis of synaptotagmin genes** *Genomics* 2001, **77**:43-49.
40. Li SH, McInnis MG, Margolis RL, Antonarakis SE and Ross CA: **Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms** *Genomics* 1993, **16**:572-579.
41. Zemni R, Bienvenu T, Vinet MC, Sefiani A, Carrie A, Billuart P, McDonnell N, Couvert P, Francis F, Chafey P, Fauchereau F, Riocourt G, Portes V, Cardona A, Frints S, Meindl A, Brandau O, Ronce N, Moraine C, Bokhoven H, Ropers HH, Sudbrak R, Kahn A, Fryns JP, Beldjord C and Chelly J: **A new gene involved in X-linked mental retardation identified by analysis of an X;2 balanced translocation** *Nat Genet* 2000, **24**:167-170.
42. Rall S. C., Jr., Weisgraber KH and Mahley RW: **Human apolipoprotein E. The complete amino acid sequence** *J Biol Chem* 1982, **257**:4171-4178.
43. van der Bliek AM, Redelmeier TE, Damke H, Tisdale EJ, Meyerowitz EM and Schmid SL: **Mutations in human dynamin block an intermediate stage in coated vesicle formation** *J Cell Biol* 1993, **122**:553-563.
44. Reddy TR, Li X, Jones Y, Ellisman MH, Ching GY, Liem RK and Wong-Staal F: **Specific interaction of HTLV tax protein and a human type IV neuronal intermediate filament protein** *Proc Natl Acad Sci U S A* 1998, **95**:702-707.
45. Oliva D, Cali L, Feo S and Giallongo A: **Complete structure of the human gene encoding neuron-specific enolase** *Genomics* 1991, **10**:157-165.
46. Stein R, Mori N, Matthews K, Lo LC and Anderson DJ: **The NGF-inducible SCG10 mRNA encodes a novel membrane-bound protein present in growth cones and abundant in developing neurons** *Neuron* 1988, **1**:463-476.
47. Ju YT, Chang AC, She BR, Tsauro ML, Hwang HM, Chao CC, Cohen SN and Lin-Chao S: **gas7: A gene expressed preferentially in growth-arrested fibroblasts and terminally differentiated Purkinje neurons affects neurite formation** *Proc Natl Acad Sci U S A* 1998, **95**:11423-11428.
48. Zieger B, Tran H, Hainmann I, Wunderle D, Zgaga-Griesz A, Blaser S and Ware J: **Characterization and expression analysis of two human septin genes, PNUTL1 and PNUTL2** *Gene* 2000, **261**:197-203.
49. Tanaka M, Tanaka T, Kijima H, Itoh J, Matsuda T, Hori S and Yamamoto M: **Characterization of tissue- and cell-type-specific expression of a novel human septin family gene, Bradeion** *Biochem Biophys Res Commun* 2001, **286**:547-553.
50. Dou D and Joseph R: **Cloning of human neuronatin gene and its localization to chromosome-20q 11.2-12: the deduced protein is a novel "proteolipid"** *Brain Res* 1996, **723**:8-22.
51. Kukita A, Mukai T, Miyata T and Hori K: **The structure of brain-specific rat aldolase C mRNA and the evolution of aldolase isozyme genes** *Eur J Biochem* 1988, **171**:471-478.
52. Perez F, Pernet-Gallay K, Nizak C, Goodson HV, Kreis TE and Goud B: **CLIPR-59, a new trans-Golgi/TGN cytoplasmic linker protein belonging to the CLIP-170 family** *J Cell Biol* 2002, **156**:631-642.
53. Vannahme C, Schubel S, Herud M, Gosling S, Hulsmann H, Paulsson M, Hartmann U and Maurer P: **Molecular cloning of testican-2: defining a novel calcium-binding proteoglycan family expressed in brain** *J Neurochem* 1999, **73**:12-20.
54. Frank M: **MAL, a proteolipid in glycosphingolipid enriched domains: functional implications in myelin and beyond** *Prog Neurobiol* 2000, **60**:531-544.
55. Janoueix-Lerosey I, Pasheva E, de Tand MF, Tavittian A and de Gunzburg J: **Identification of a specific effector of the small GTP-binding protein Rap2** *Eur J Biochem* 1998, **252**:290-298.

56. Li Y, Chin LS, Weigel C and Li L: **Spring, a novel RING finger protein that regulates synaptic vesicle exocytosis** *J Biol Chem* 2001, **276**:40824-40833.
57. Wang W, Zhou Z, Zhao W, Huang Y, Tang R, Ying K, Xie Y and Mao Y: **Molecular cloning, mapping and characterization of the human neurocalcin delta gene (NCALD)** *Biochim Biophys Acta* 2001, **1518**:162-167.
58. Vives V, Alonso G, Solal AC, Joubert D and Legraverend C: **Visualization of S100B-positive neurons and glia in the central nervous system of EGFP transgenic mice** *J Comp Neurol* 2003, **457**:404-419.
59. Hernandez MC, Andres-Barquin PJ, Holt I and Israel MA: **Cloning of human ENC-1 and evaluation of its expression and regulation in nervous system tumors** *Exp Cell Res* 1998, **242**:470-477.
60. Dalmau J, Gultekin SH, Voltz R, Hoard R, DesChamps T, Balmaceda C, Batchelor T, Gerstner E, Eichen J, Frennier J, Posner JB and Rosenfeld MR: **Ma1, a novel neuron- and testis-specific protein, is recognized by the serum of patients with paraneoplastic neurological disorders** *Brain* 1999, **122 (Pt 1)**:27-39.
61. Sasaki K, Kurata-Miura K, Ujita M, Angata K, Nakagawa S, Sekine S, Nishi T and Fukuda M: **Expression cloning of cDNA encoding a human beta-1,3-N-acetylglucosaminyltransferase that is essential for poly-N-acetylglucosamine synthesis** *Proc Natl Acad Sci U S A* 1997, **94**:14294-14299.
62. Morris R: **Thy-1 in developing nervous tissue** *Dev Neurosci* 1985, **7**:133-160.
63. Hunt JM, Bommert K, Charlton MP, Kistner A, Habermann E, Augustine GJ and Betz H: **A post-docking role for synaptobrevin in synaptic vesicle fusion** *Neuron* 1994, **12**:1269-1279.
64. Gumienny TL, Brugnera E, Tosello-Tramont AC, Kinchen JM, Haney LB, Nishiwaki K, Walk SF, Nemergut ME, Macara IG, Francis R, Schedl T, Qin Y, Van Aelst L, Hengartner MO and Ravichandran KS: **CED-12/ELMO, a novel member of the Crkl/Dock180/Rac pathway, is required for phagocytosis and cell migration** *Cell* 2001, **107**:27-41.
65. Bressler SL, Gray MD, Sopher BL, Hu Q, Hearn MG, Pham DG, Dinulos MB, Fukuchi K, Sisodia SS, Miller MA, Distèche CM and Martin GM: **cDNA cloning and chromosomal mapping of the human Fe65 gene: interaction of the conserved cytoplasmic domains of the human beta-amyloid precursor protein and its homologues with the mouse Fe65 protein** *Hum Mol Genet* 1996, **5**:1589-1598.
66. Leung T, How BE, Manser E and Lim L: **Cerebellar beta 2-chimaerin, a GTPase-activating protein for p21 ras-related rac is specifically expressed in granule cells and has a unique N-terminal SH2 domain** *J Biol Chem* 1994, **269**:12888-12892.
67. Jobsis GJ, Keizers H, Vreijling JP, de Visser M, Speer MC, Wolterman RA, Baas F and Bolhuis PA: **Type VI collagen mutations in Bethlem myopathy, an autosomal dominant myopathy with contractures** *Nat Genet* 1996, **14**:113-115.
68. Grover J, Chen XN, Korenberg JR and Roughley PJ: **The human lumican gene. Organization, chromosomal location, and expression in articular cartilage** *J Biol Chem* 1995, **270**:21942-21949.
69. Todd SC, Doctor VS and Levy S: **Sequences and expression of six new members of the tetraspanin/TM4SF family** *Biochim Biophys Acta* 1998, **1399**:101-104.
70. Spurr NK, Durbin H, Sheer D, Parkar M, Bobrow L and Bodmer WF: **Characterization and chromosomal assignment of a human cell surface antigen defined by the monoclonal antibody AUAI** *Int J Cancer* 1986, **38**:631-636.
71. Lai EC, Kao FT, Law ML and Woo SL: **Assignment of the alpha 1-antitrypsin gene and a sequence-related gene to human chromosome 14 by molecular hybridization** *Am J Hum Genet* 1983, **35**:385-392.
72. Katahira J, Sugiyama H, Inoue N, Horiguchi Y, Matsuda M and Sugimoto N: **Clostridium perfringens enterotoxin utilizes two structurally related membrane proteins as functional receptors in vivo** *J Biol Chem* 1997, **272**:26652-26658.
73. Nagase T, Miyajima N, Tanaka A, Sazuka T, Seki N, Sato S, Tabata S, Ishikawa K, Kawarabayashi Y and Kotani H: **Prediction of the coding sequences of unidentified human genes. III. The coding sequences of 40 new genes (KIAA0081-KIAA0120) deduced by analysis of cDNA clones from human cell line KG-1 (supplement)** *DNA Res* 1995, **2**:51-59.
74. Trask DK, Band V, Zajchowski DA, Yaswen P, Suh T and Sager R: **Keratins as markers that distinguish normal and tumor-derived mammary epithelial cells** *Proc Natl Acad Sci U S A* 1990, **87**:2319-2323.
75. Kulesh DA and Oshima RG: **Complete structure of the gene for human keratin 18** *Genomics* 1989, **4**:339-347.
76. Bader BL, Magin TM, Hatzfeld M and Franke WW: **Amino acid sequence and gene organization of cytokeratin no. 19, an exceptional tail-less intermediate filament protein** *Embo J* 1986, **5**:1865-1875.
77. Mbikay M, Nolet S, Fournier S, Benjannet S, Chapdelaine P, Paradis G, Dube JY, Tremblay R, Lazure C and Seidah NG: **Molecular cloning and sequence of the cDNA for a 94-amino-acid seminal plasma protein secreted by the human prostate** *DNA* 1987, **6**:23-29.
78. Sharief FS, Lee H, Leuderman MM, Lundwall A, Deaven LL, Lee CL and Li SS: **Human prostatic acid phosphatase: cDNA cloning, gene mapping and protein sequence homology with lysosomal acid phosphatase** *Biochem Biophys Res Commun* 1989, **160**:79-86.
79. Lundwall A and Lilja H: **Molecular cloning of human prostate specific antigen cDNA** *FEBS Lett* 1987, **214**:317-322.
80. Matsuoka R, Yoshida MC, Furutani Y, Imamura S, Kanda N, Yanagisawa M, Masaki T and Takao A: **Human smooth muscle myosin heavy chain gene mapped to chromosomal region 16q12** *Am J Med Genet* 1993, **46**:61-67.
81. Laing NG, Wilton SD, Akkari PA, Dorosz S, Boundy K, Kneebone C, Blumbergs P, White S, Watkins H and Love DR: **A mutation in the alpha tropomyosin gene TPM3 associated with autosomal dominant nemaline myopathy** *Nat Genet* 1995, **9**:75-79.
82. Maestrini E, Patrosso C, Mancini M, Rivella S, Rocchi M, Repetto M, Villa A, Frattini A, Zoppe M and Vezzoni P: **Mapping of two genes encoding isoforms of the actin binding protein ABP-280, a dystrophin like protein, to Xq28 and to chromosome 7** *Hum Mol Genet* 1993, **2**:761-766.
83. Solway J, Seltzer J, Samaha FF, Kim S, Alger LE, Niu Q, Morrisey EE, Ip HS and Parmacek MS: **Structure and expression of a smooth muscle cell-specific gene, SM22 alpha** *J Biol Chem* 1995, **270**:13460-13469.
84. Maric SC, Crozat A and Janne OA: **Structure and organization of the human S-adenosylmethionine decarboxylase gene** *J Biol Chem* 1992, **267**:18915-18923.
85. Kumar CC and Chang C: **Human smooth muscle myosin light chain-2 gene expression is repressed in ras transformed fibroblast cells** *Cell Growth Differ* 1992, **3**:1-10.
86. Lin B, White JT, Ferguson C, Wang S, Vessella R, Bumgarner R, True LD, Hood L and Nelson PS: **Prostate short-chain dehydrogenase reductase I (PSDR1): a new member of the short-chain steroid dehydrogenase/reductase family highly expressed in normal and neoplastic prostate epithelium** *Cancer Res* 2001, **61**:1611-1618.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

