Pacific Symposium on Biocomputing 5:427-439 (2000)

EST databases as multi-conditional gene expression datasets

R.M. EWING
Carnegie Institution of Washington, Department of Plant Biology, 260 Panama Street, Stanford, California 94305. (ewing@genome.stanford.edu)

J-M. CLAVERIE
Structural and Genetic Information Laboratory, CNRS UMR-1889, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France. (jmc@igs.cnrs-mrs.fr)

Large-scale expression data, such as that generated by hybridization to microarrays, is potentially a rich source of information on gene function and regulation. By clustering genes according to their expression profiles, groups of genes involved in the same pathways or sharing common regulatory mechanisms may be identified. Publicly-available EST collections are a largely unexplored source of expression data. We previously used a sample of rice ESTs to generate 'digital expression profiles' by counting the frequency of tags for different genes sequenced from different cDNA libraries. A simple statistical test was used to associate genes or cDNA libraries having similar expression profiles. Here we further validate this approach using larger samples of ESTs from the UniGene projects (clustered human, mouse and rat ESTs). Our results show that genes clustered on the basis of expression profile may represent genes implicated in similar pathways or coding for different subunits of multi-component enzyme complexes. In addition we suggest that comparison of clusters from different species, may be useful for confirmation or prediction of orthologs.

1 Introduction

Techniques for monitoring in parallel the expression of 1000s of genes, such as microarray hybridization, EST generation and SAGE, are providing biologists with huge amounts of expression information. A notion in common to many studies is that genes that function in the same pathways are likely to be co-expressed and possibly co-regulated. We subscribe to this 'guilt-by-association' [1] viewpoint and believe that associating genes based upon their expression patterns is a powerful means of annotating 'anonymous' genes, and assigning genes to pathways or cellular roles.

Several approaches have been presented for the analysis of large-gene expression datasets (recently reviewed by Claverie [2]). The expression of a set of 112 genes was assayed using RT-PCR in developing rat spinal cord [3]. Both Euclidean distance and information theoretic approaches were used to assign genes to one of 6 basic 'waves' of expression [3 4]. Eisen *et al* [5] developed a method in which linear correlation coefficients are calculated for each pair of expression profiles followed by hierarchical clustering and display of the pri-

mary expression data as a colormap. This technique was used to analyse both human and yeast gene expression data [5,6]. In many cases, genes implicated in the same processes or encoding subunits of macro-molecular complexes (e.g proteasome, ribosome and histone components) were found to have correlated expression profiles.

Tag-sampling approaches, such as SAGE [7] and ESTs [8] have also been used for large-scale gene expression analyses [2]. Instances of differential expression of genes in different tissues may be identified by counting tags occurring in different cDNA libraries [9]. Alternatively, genes may be grouped according to the similarity of their expression profiles. We recently used a publicly-available rice EST dataset to derive digital expression profiles of genes and cDNA libraries [10]. Correlations between profiles were identified and genes or libraries organised into clusters. Our results showed that despite the small sample size used (707 genes represented in 10 cDNA libraries), interesting and biologically-relevant correlations between libraries or genes were revealed.

Here we present an exploratory study of human, mouse and rat gene expression using EST datasets as the primary source of data. By using EST frequencies from different cDNA libraries as a rough gene expression measurement, and associating genes with similar expression profiles, we show that genes that would be expected to be co-expressed, based upon known function, and groups of genes encoding subunits from multi-subunit complexes cluster together. We also show how cDNA library expression profiles may be derived from EST data and used to compare the overall patterns of gene expression in different tissues or organs. In addition, by integrating sequence-based information with gene expression data and comparing results between species (human, mouse and rat), we show that 'conserved correlations' may be identified. These observations may be useful for identification or confirmation of gene orthologs between species.

We argue that EST databases are a valid and reliable, but as yet relatively unexplored source of gene expression data. We focus on the biological implications of clustering expression data and illustrate our results with several examples.

2 Results

2.1 Preparation of initial data

The starting point for our analysis is the classification generated by the human, mouse and rat UniGene projects, in which ESTs, full-length mRNA sequences and extracted genomic coding regions are organised into 'gene-

oriented' clusters (http://www.ncbi.nlm.nih.gov/UniGene/index.html). The cDNA libraries used for generating the ESTs are themselves derived from a multitude of different organs/tissues/cell-types and developmental stages. This information is exploited in our study by analysing the representation of cDNA libraries in unigenes*, and thereby deriving rough expression profiles for each gene in UniGene.

Data from large-scale gene expression experiments is best summarized as as a gene by condition (g X c) data matrix, whereby each cell in the matrix contains the expression measurement of gene, g, in condition c. In our study, genes (rows) are represented by unigenes, and conditions (columns) by cDNA libraries. Each cell, $g_i c_j$, is populated by counting the number of ESTs in unigene, g_i, derived from cDNA library, c_j. Matrix rows are therefore gene expression profiles and columns cDNA library expression profiles.

To generate expression data matrices of manageable size (the total human UniGene dataset consists of approximately 1 000 000 ESTs in 65 000 clusters), an initial filtration step was performed. This consisted of rejecting clusters with fewer than 10 constituent sequences (not including non-EST entries such as full-length mRNAs) and rejecting cDNA libraries for which fewer than 1000 ESTs have been sampled. For the human set, we also reduced the dataset by excluding 'anonymous' unigenes - those without significant database matches. The resulting expression data matrices are detailed in Table 1.

Table 1: Summary of initial data and complete-link clustering

	HUMAN	MOUSE	RAT
(a) Expression data matrices			
Initial ESTs	425451	216517	22767
unigenes	5624	3889	1295
cDNA Libraries	129	48	22
(b) Complete-link clustering of unigenes			
Complete-link clusters	746	602	341
unigenes in clusters	2394	2206	956
Cluster size range (largest:smallest)	109:2	88:2	27:2

2.2 Methods overview

The heart of our analysis is measurement of the correlation between each pair of gene or library expression profiles. In common with other studies [5 11], we

*To avoid confusion, the UniGene project will hereafter be referred to as 'UniGene', and the UniGene clusters as 'unigenes'.

found that the Pearson correlation coefficient is both a simple and appropriate measure of expression profile similarity. Pearson correlation coefficients are calculated and stored as c X c library or g X g gene similarity matrices, which are then the basis for clustering of genes or libraries.

The data was clustered in two different ways; discrete, 'complete-link' clustering and hierarchical clustering. Complete-link clusters, in which clusters are joined only if all members of both clusters match, were generated using a threshold Pearson correlation value (r>=0.75) (see Table 1). Hierarchical 'phylogenetic-type' clustering was performed by first calculating Euclidean distances and then using the UPGMA algorithm (average-linkage cluster analysis [12]), implemented in the Phylip package [13], to generate a dendrogram representing all objects. Note that we calculate the Euclidean distance between two objects, X and Y, from the respective correlation coefficients with all other objects (x_1, x_2, ... , x_N) and (y_1, y_2, ... , y_N), rather than calculating the distance solely from the correlation value for X and Y.

Since the starting point for both the complete-link clustering and hierarchical clustering are the same gene or library similarity matrices, the associations found in the data are generally very similar (e.g genes found in the same complete-link clusters will be in close proximity on the appropriate dendrogram). We found, however, that the relatively small number of cDNA libraries (129 for human) were best represented as a dendrogram, whereas the much larger number of genes (5624 for human) were more easily manipulated as discrete, complete-link clusters. Results from both approaches are therefore presented in the following sections.

In common with other reports [14 5 3] we find that colormaps are a good visual way of representing expression data. In all colormaps presented here, the data has been reordered such that objects along both axes are grouped according to similarity (reordered according to the order present in the appropriate dendrogram).

2.3 cDNA Library Analysis

To characterise the overall similarities between transcriptomes of different tissues/organs, cDNA libraries were clustered using the methods described in the preceding section.

An unrooted dendrogram of the human cDNA libraries is shown in Figure 1. The 129 libraries have been classified into one of 30 different tissue/organ classifications, based upon the existing UniGene classification (see http://www.ncbi.nlm.nih.gov/UniGene/Hs.Home.html). In addition, the 34 normalized libraries are marked 'n', and brain libraries further classified ac-

cording to the key.

Several conclusions can be drawn from the tree. First, libraries from the same tissues do not consistently cluster together (notable exceptions are brain, muscle and prostate clades - marked with arrows). Both biological and methodological factors likely contribute to this observation. It may be that expression profiles of some tissues are composed mainly of ubiquitously expressed genes ('housekeeping genes') - libraries derived from these tissues would be expected to have overlapping expression profiles. The 'tissue-specific clades' - brain, muscle and prostate for example, imply that expression profiles from these tissues are sufficiently distinct to form outlying groups - perhaps an indication that there are significant numbers of genes specifically expressed in brain, muscle or prostate tissues.

Second, there are methodological issues. Much variation between libraries derived from the same tissues may be attributed to cDNA library preparation - for example how the tissue was initially dissected. The 7 prostate libraries in Figure 1 are distributed between a prostate-specific clade and other tissue-non-specific clades. It could be that those prostate libraries clustering with other tissue types were prepared from prostate contaminated with surrounding tissues. This type of analysis may be useful for selection of cDNA libraries for further sequencing. For example, it may be more productive to search for prostate-specific genes in the 'clean' prostate libraries (those in the prostate-specific clade), than in those prostate libraries which cluster with libraries from other tissues.

Normalization of cDNA libraries is possibly the most important methodological factor - normalized cDNA libraries are marked 'n' in Figure 1 and there is evidently some segregation of normalized and non-normalized libraries, regardless of tissue-type. Since normalization of cDNA libraries reduces the quantitative differences between abundant and rare cDNAs, it may be that a significant portion of the variablility between tissue expression profiles is due to quantitative (levels of expression) rather than qualitative (the expression of distinct genes) differences in gene expression. It could also be argued that normalized cDNA libraries should not be used for tag-based expression profiling since tag counts are no longer a true reflection of transcript abundances (see concluding remarks section for further discussion of this).

Third, it should be borne in mind that the library classifications in Figure 1 are relatively broad. The 'brain' category, for example, contains libraries derived from many different brain tissues and developmental stages. Finer subclassification of tissue types reveals some clustering of 'infant' brain for example - perhaps indicating the presence of gene expression patterns specific to this period of development.

2.4 Clustering of gene expression profiles

Many interesting correlations of genes with related functions were observed in the clustered data.

In several cases, clusters contained unigenes encoding multiple subunits of multi-protein enzyme complexes. An example , drawn from the mouse data, is shown in Figure 2(A). Cytochrome c oxidase is the terminal oxidase in mitochondrial electron transport chain and in eukaryotes is comprised of 7-11 subunits; the largest three are encoded on the mitochondrial genome and the remainder in the nuclear genome. Figure 2(A), shows all complete-link clusters found to contain at least one cytochrome c oxidase subunit. Ten of the total of 15 unigenes encoding cytochrome c oxidase components are represented in 5 clusters. The 5 clusters shown are also rich in other nuclear-encoded mitochondrial proteins; 4 subunits of NADH-ubiquinone oxidoreductase (another electron transport chain component) show correlated expression profiles.

Figure 2(B) shows a discrete cluster of 8 human unigenes, all encoding commonly found muscle-related proteins. Several genes are involved in energy transduction (ATPase, AMP deaminase, creatine kinase and beta-enolase), whereas others (troponin, myosin) are structural components of muscle fibres.

We also sought to integrate our findings with data from other large-scale gene expression experiments. As a preliminary investigation, we compared our results to results obtained in a study in which the responses of 8600 genes were analysed following treatment of human fibroblasts with serum [6]. Comparisons between the studies were facilitated by the fact that Iyer *et al* [6] used the UniGene database to select cDNA clones for inclusion on microarrays, making it easy to cross-reference genes between studies.

We identified genes that were correlated both during the serum-stimulation time-course [6], and across the 129 human cDNA libraries. The best example of our findings, and one that is supported by existing literature, is shown in Figure 2(C). During the serum stimulation time course the kinetics of induction of Early growth response 1 (EGR1) and P55-c-fos proto-oncogene (C-FOS) are very similar, and the genes are clustered together [6]. Similarly, in our own study, hierarchical clustering of gene expression profiles places the two genes on adjacent positions on the tree, suggesting that for these two genes, correlated expression extends beyond the specific cell-type (fibroblast) and condition (serum-stimulation) to many different cell-types, tissues and organs. This conclusion is supported by other reports, in which the expression kinetics of EGR1 and C-FOS have been shown to be remarkably similar, suggesting that the genes are co-regulated[15,16]. Note that the absolute numbers of ESTs in the EGR1 and C-FOS expression profiles are relatively low (65 C-FOS ESTs in

the aorta library is the maximum, all other counts are 15 or below), suggesting that tag-based expression profiles may be accurate for even relatively weakly-expressed transcripts.

2.5 Conservation of correlations between species

The examples cited above indicate that biologically-relevant correlations between expression profiles of genes or libraries can be identified from EST data. We wished to extend these results and explore other ways of using the data.

One interesting possibility is the identification of 'conserved correlations' in data from different species. If correlations between gene expression profiles are indeed functionally relevant, it should be possible to identify genes which show the same associations or correlations in different species. Observing 'conserved correlations' between different organisms would firstly add confidence to correlations observed in a single species, and secondly may be a powerful method of confirming or identifying orthologous relationships between genes.

By taking the dendrograms derived from the hierarchical clustering of human and mouse genes, and cross-referencing objects on those dendrograms (known human/mouse orthologs, defined in a previous study of human/rodent orthologous genes[17] or human and mouse genes with significant sequence alignment scores (gapped-TBLASTx [18], default scoring matrix, score > 380)), we were able to identify associations between genes that are conserved between human and mouse.

An example is shown in Figure 3. A fragment of the human gene dendrogram is shown opposite two fragments from the mouse gene dendrogram. Sequence-based relationships between human and mouse genes (confirmed orthologs or significant sequence alignments) are overlaid on the gene expression dendrograms; solid boxes/lines represent human/mouse ortholog pairs, hashed boxes/dotted lines represent significant sequence alignments.

Interestingly, several of the genes featured in Figure 3 (Thrombospondins, B94 , LDL-related receptor and tissue inhibitors of metalloproteinases) are involved in vascularisation/angiogenesis [19 20 21]. Furthermore, it is known that the LDL-receptor related protein mediates the cellular-internalisation of thrombospondin and its subsequent degradation [22 23]. The associations identified in Figure 3 therefore concur with existing data and suggest that the mouse and human orthologs have maintained their functions since the divergence of the two species.

3 Concluding remarks

We have derived digital expression profiles from publicly-available EST data and shown how correlations between gene or library expression profiles may be identified. The principal aim of our study is to argue that existing EST databases are a valuable source of expression data, which can be integrated with expression data from other sources and used to identify many interesting biological relationships. Our approach differs from other large-scale studies of gene expression [3 5 24] in that we are not examining gene expression within a specific developmental window or specific cell-types. Rather, we are identifying overall correlations - for example genes whose expression is correlated across many different conditions (e.g 129 human cDNA libraries). Our results show (and we predict) that genes with correlated profiles are frequently genes whose products either physically interact (as in a multi-subunit complex) or function in the same pathway.

Potential drawbacks of expression analysis based upon EST counts are as follows. First, EST data is derived from cDNA libraries prepared using different techniques. The classification of cDNA libraries (see Figure 1) suggests that normalization of cDNA libraries has a significant effect on the resulting expression profile. Finer 'filtration' of the initial data - e.g exclusion of normalized libraries may allow better correlation of libraries based upon tissue-type. Alternatively, we have explored the possibility of transforming 'continuous' EST count data into a discrete binary representation (presence and absence of ESTs in a given library represented by 1 and 0 respectively) (results not shown). Expression profiles are thereby represented as strings of 1s and 0s, and amenable to analysis using information theoretic approaches, as explored by Michaels *et al* [4] in this context. Although quantitative information is lost from the expression profiles, this approach may be more appropriate for comparisons between normalized and non-normalized cDNA libraries, since quantitative information in normalized libraries (i.e the absolute numbers of EST counts) is not truly representative of actual transcript abundances.

Second, sequence-based clustering of ESTs is not an unambiguous process, and the results vary according to the clustering stategy and parameters. A recent comparison of three publicly-accessible 'gene-indexing' projects showed how the relatively non-stringent parameters used in the UniGene project affect the resulting clusters [25]. Results from the other gene-indexing projects examined, HGI (http://www.tigr.org/tdb/tdb.html) and STACK (http://www.sanbi.ac.za/), are relatively redundant, but allow alternative transcript forms, such as alternatively spliced transcripts, to be represented by different clusters. Clearly, in the context of digital expression profiling, the

clustering parameters are of importance.

We envisage that observations of correlated gene expression will be most useful when combined with other molecular data (especially sequence data). As illustrated here (Figure 3), this may take the form of identification of conserved correlations between different species. Genes with correlated expression may be co-regulated - screening regulatory regions of genes with correlated expression may lead to identification of regulatory elements. This approach has already been applied in yeast, whereby putative 5' regulatory elements were identified in yeast genes with correlated profiles [26].

Finally, integration of physical mapping data may lead to discovery of correlations between chromosomal position and gene expression. This approach was taken with yeast SAGE data [27], and was also briefly discussed by Michaels *et al* [4], who noted that three genes with correlated expression profiles also mapped to the same cytogenetic band. In the context of EST data, the UniGene project is already integrated with mapping data (e.g see Bortoluzzi et al [28]), which will provide a convenient means of linking expression and physical mapping data.

4 References

1. R Brent. *Curr Biol*, 9(9):338–41, 1999.
2. J-M. Claverie. *Hum Mol Gen*, 8(10):1821–1832, 1999.
3. X Wen, S Fuhrman, GS Michaels, DB Carr, S Smith, JL Barker, and R Somogyi. *Proc Natl Acad Sci U S A*, 95(1):334–9, 1998.
4. GS Michaels, DB Carr, M Askenazi, S Fuhrman, X Wen, and R Somogyi. In *Pacific Symposium on Biocomputing*, volume 3, pages 42–53, 1998.
5. MB Eisen, PT Spellman, PO Brown, and D Botstein. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.
6. VR Iyer, MB Eisen, DT Ross, G Schuler, T Moore, JCF Lee, JM Trent, LM Staudt, J Hudson Jr, MS Boguski, D Lashkari, D Shalon, D Botstein, and Brown PO. *Science*, 283(5398):83–87, 1999.
7. VE Velculescu, L Zhang, B Vogelstein, and KW Kinzler. *Science*, 270(5235):484–7, 1995.
8. K Okubo, N Hori, R Matoba, T Niiyama, A Fukushima, Y Kojima, and K Matsubara. *Nat Genet*, 2(3):173–9, 1992.
9. S Audic and J-M Claverie. *Genome Research*, 7:986–995, 1997.
10. R Ewing, A Ben Kahla, O Poirot, F Lopez, S Audic, and J-M Claverie. *Genome Research*, 9(10):000–000, 1999.
11. P D'haeseleer, X Wen, S Fuhrman, and R Somogyi. *Information Processing in cells and tissues*, pages 203–212. Plenum Publishing, 1998.

12. RR Sokal and CD Michener. *University of Kansas Scientific Bulletin*, 28:1409–1438, 1958.
13. J Felsenstein. Distributed by the author. Department of Genetics, University of Washington, Seattle, 1993.
14. JN Weinstein, TG Myers, PM O'Connor, SH Friend, KW Ko hn AJ Fornace Jr, T Fojo, SE Bates, LV Rubinstein, NL Anderson, JK Buolamwini, WW van Osdol, AP Monks, DA Scudiero, EA Sausville, DW Zaharevitz, B Bunow, VN Viswanadhan, GS Johnson, RE Wittes, and KD Paull. *Science*, 275(5298):343–9, 1997.
15. P Lemaire, O Revelantand R Bravo, and P Charnay. *Proc Natl Acad Sci U S A*, 85(13):4691–5, 1988.
16. AP McMahon, JE Champion, JA McMahon, and VP Sukhatme. *Development*, 108(2):281–7, 1990.
17. W Makalowski and MS Boguski. *Proc Natl Acad Sci U S A*, 95(16):9407–12, 1998.
18. SF Altschul, TL Madden, AS Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. *Nucleic Acids Research*, 25(17):3389–3402, 1997.
19. V Sarma, FW Wolf, RM Marks, TB Shows, and VM Dixit. *J Immunol*, 148(10):3302–12, 1992.
20. FW Wolf, V Sarma, M Seldin, S Drake, SJ Suchard, H Shao, KS O'Shea, and VM Dixit. *J Biol Chem*, 269(5):3633–40, 1994.
21. DE Gomez, DF Alonso, H Yoshiji, and UP Thorgeirsson. *Eur J Cell Biol*, 74(2):111–22, 1997.
22. I Mikhailenko, MZ Kounnas, and DK Strickland. *J Biol Chem*, 270(16):9543–9, 1995.
23. DK Strickland, MZ Kounnas, and WS Argraves. *FASEB J*, 9(10):890–8, 1995.
24. P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, D Dmitrovsky, ES Lander, and TR Golub. *Proc Natl Acad Sci U S A*, 96(6):2907–2912, 1999.
25. J Bouck, W Yu, R Gibbs, and K Worley. *Trend Genet*, 15(4):159–62, 1999.
26. A Brzma, I Jonassen, J Vilo, and E Ukkonen. *Genome Res*, 8(11):1202–15, 1998.
27. VE Velculescu, L Zhang, W Zhou, B Vogelstein, MA Basrai, DE Bassett Jr, P Hieter, B Vogelstein, and KW Kinzler. *Cell*, 88(2):243–51, 1997.
28. S Bortoluzzi, L Rampoldi, B Simionati, R Zimbello, A Barbon, F d'Alessi, N Tiso, A Pallavicini, S Toppo, N Cannata, G Valle, G Lanfranchi, and GA Danieli. *Genome Res*, 8(8):817–25, 1998.

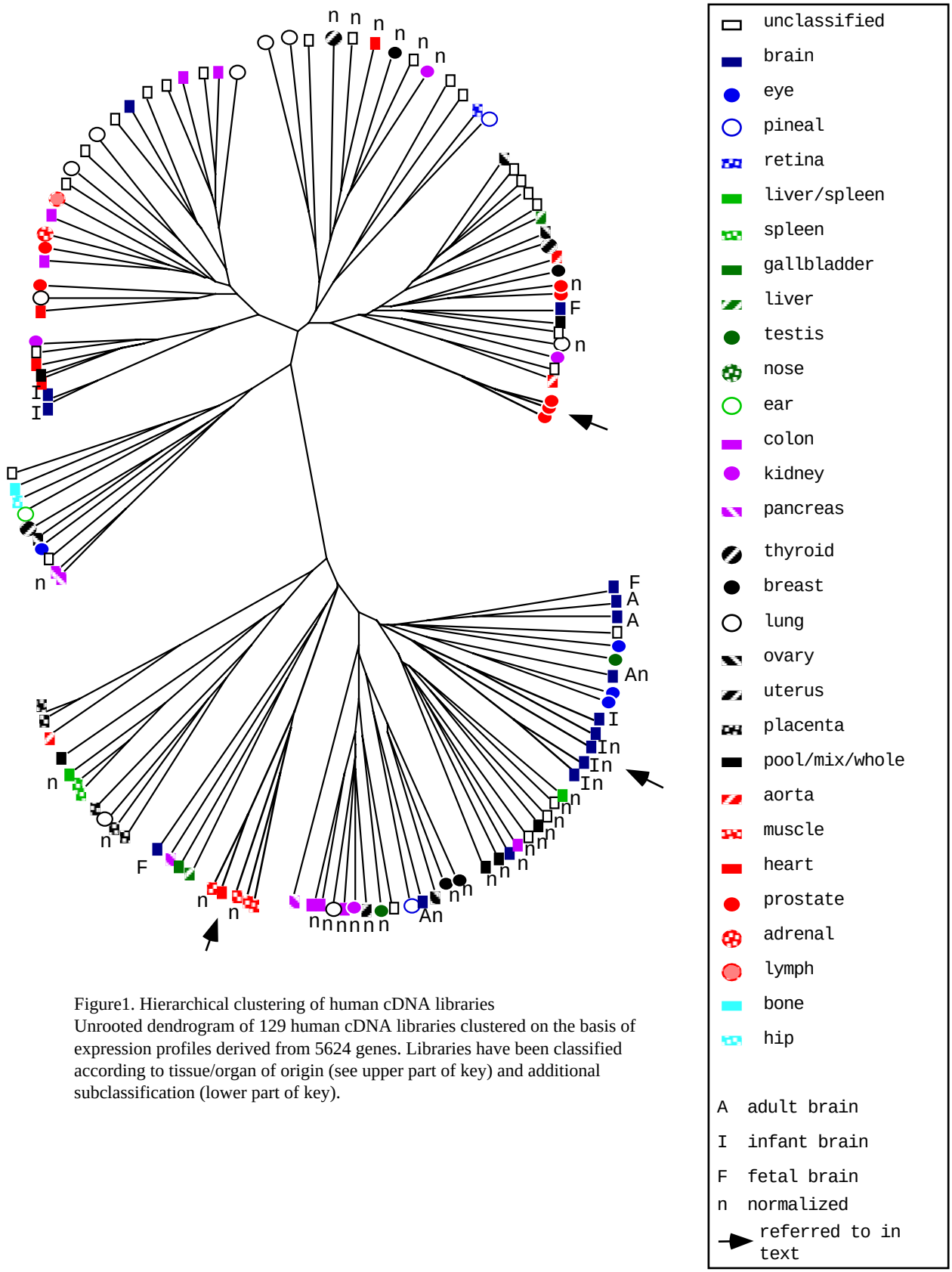

Figure1. Hierarchical clustering of human cDNA libraries
Unrooted dendrogram of 129 human cDNA libraries clustered on the basis of expression profiles derived from 5624 genes. Libraries have been classified according to tissue/organ of origin (see upper part of key) and additional subclassification (lower part of key).

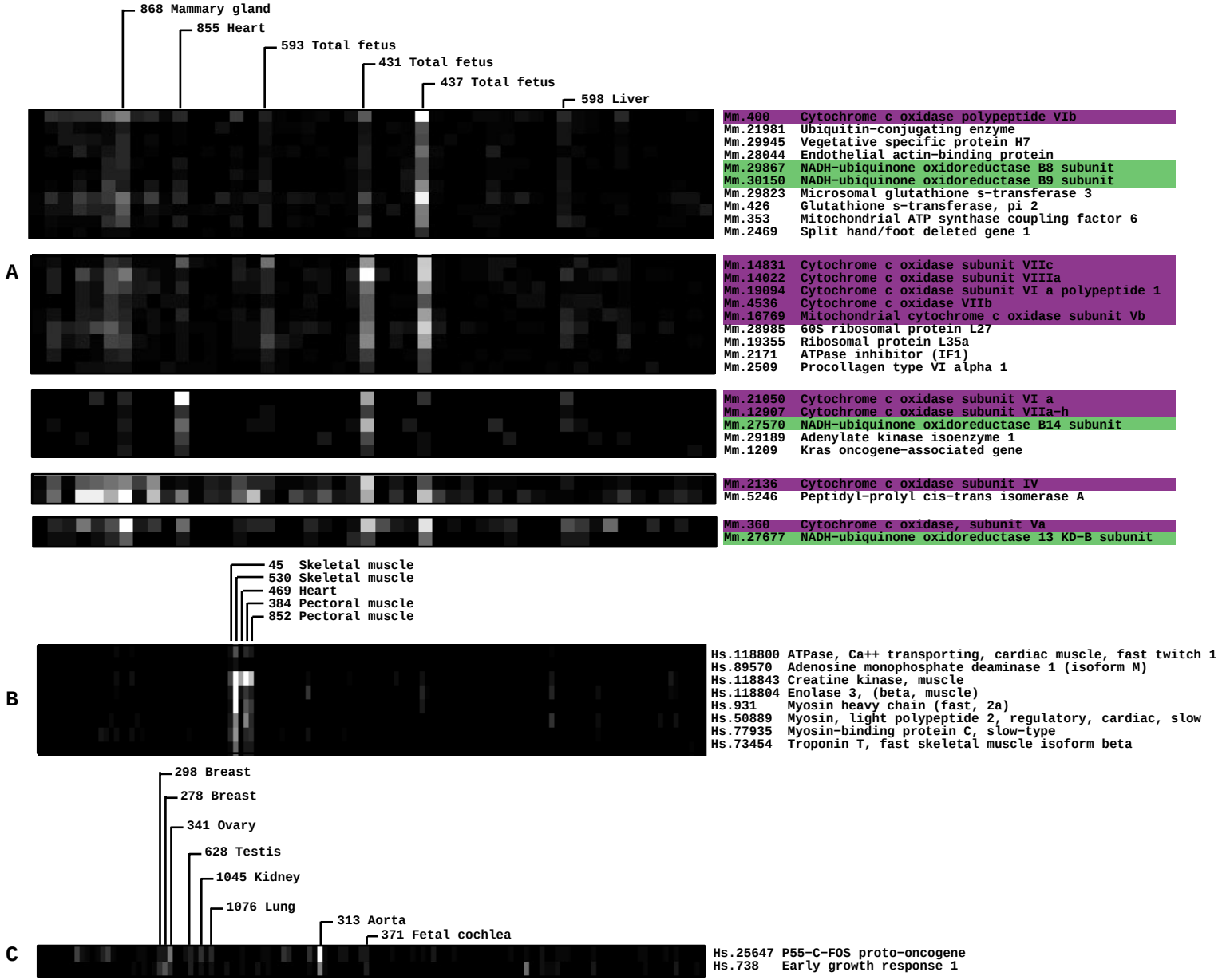

Figure 2. Clustered gene expression profiles.
(A) Five discrete, complete−link clusters (correlation coefficient, r >= 0.75) from the mouse UniGene dataset (3889 unigenes x 48 cDNAlibraries). A short description of the gene product is given with each unigene. The colormap has been scaled in order to represent the expression data as clearly as possible; white represents 50 or more ESTs, black represents 0 ESTs and shades of grey, intermediate values. Highlighted unigenes encode cytochrome c oxidase components (dark grey) or NADH−ubiquinone oxidase components (light grey). Libraries contributing significantly to the expression profiles are identified.
(B) Discrete, complete−link cluster (r >= 0.75) of 8 unigenes from the human dataset (5624 unigenes x 129 cDNA libraries). Colormap scale: white >= 65 ESTs, black = 0 ESTs, 0 < greys < 65.
(C) Expression profiles of two human unigenes, taken from hierarchical clustering (nearest neighbors on a dendrogram of 5624 human unigenes). Colormap scale: white >= 65 ESTs, black = 0 ESTs, 0 < greys < 65.

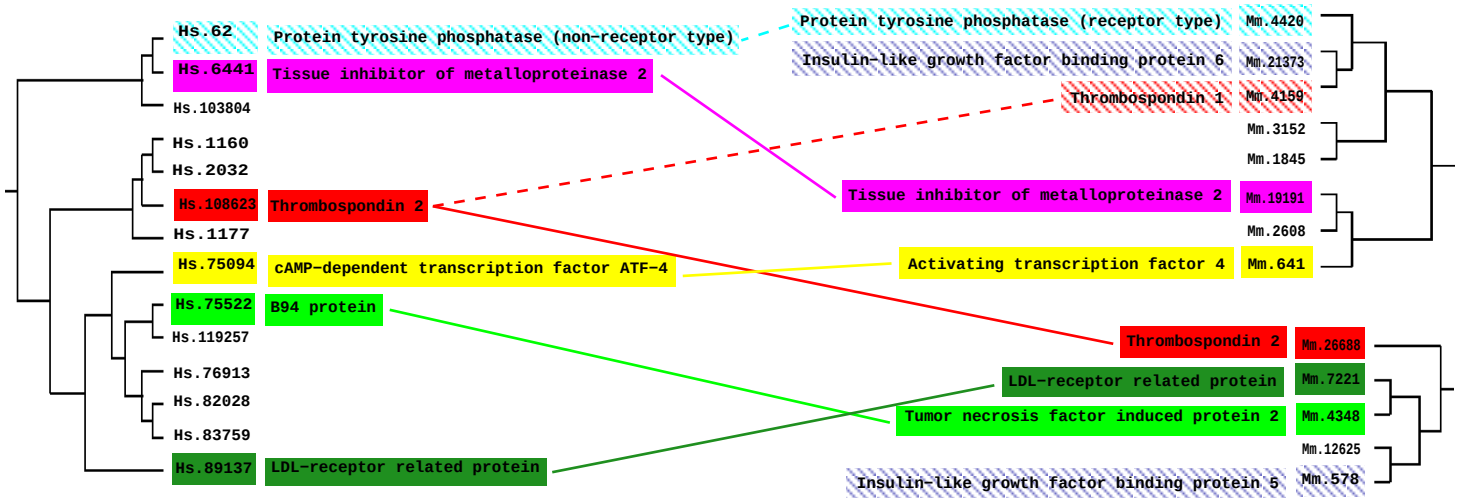

Figure 3. Conservation of correlations in hierarchical clustering of human and mouse unigenes.
Fragment from human dendrogram (5624 genes) (A)and two fragments from mouse dendrogram (3889 genes) (B) showing conservation of correlations. For each unigene, the UniGene number is shown along with a brief description of the predicted gene product. Orthologous relationships between human and mouse genes are shown in solid colours with solid lines. Hashed boxes and dotted lines represent significant alignment scores (gapped−TBLASTx score > 380) between the human and mouse genes.