Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification

Itai Yanai[1], Hila Benjamin[1,2], Michael Shmoish[1], Vered Chalifa-Caspi[3], Maxim Shklar[1], Ron Ophir[3], Arren Bar-Even[1], Shirley Horn-Saban[3], Marilyn Safran[3], Eytan Domany[2], Doron Lancet[1*] and Orit Shmueli[1]

Departments of [1] Molecular Genetics, [2] Physics of Complex Systems, and [3] Biological Services, Weizmann Institute of Science, 76100 Rehovot, Israel

*To whom correspondence should be addressed

Running head: Midrange genome-wide transcription profiles

Keywords: tissue specificity, human transcriptome, expression profiling, DNA arrays, cluster analysis, suppressed expression.

Abstract

Motivation: Genes are often characterized dichotomously as either housekeeping or single-tissue specific. We conjectured that crucial functional information resides in genes with midrange profiles of expression.

Results: In order to obtain such novel information genome-wide, we have determined the mRNA expression levels for the one of the largest hitherto analyzed set of 62,839 probesets in 12 representative normal human tissues. Indeed, when using a newly-defined graded tissue specificity index τ, valued between 0 for housekeeping genes and 1 for tissue specific genes, genes with midrange profiles having $0.15 < \tau < 0.85$ were found to constitute >50% of all expression patterns. We developed a binary classification, indicating for every gene the I_B tissues in which it is overly expressed, and the 12 - I_B tissues in which it shows low expression. The 85 dominant midrange patterns with $I_B = 2$ to 11 were found to be bimodally distributed, and to contribute most significantly to the definition of tissue specification dendrograms. Our analyses provide a novel route to infer expression profiles for presumed ancestral nodes in the tissue dendrogram. Such definition has uncovered an unsuspected correlation, whereby *de novo* enhancement and diminution of gene expression go hand in hand. These findings highlight the importance of gene suppression events, with implications to the course of tissue specification in ontogeny and phylogeny.

Availability: All data and analyses are publically available at the GeneNote website, http://genecards.weizmann.ac.il/genenote/ and, GEO accession GSE803.

Contact: Doron Lancet, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, telephone +972 8-934-4121, fax +972 8 9344487, email: doron.lancet@weizmann.ac.il

Supplementary Information: Four tables

Introduction

The ontogeny of complex multicellular organisms is enabled by the differential expression of genes across various cell types. Expression profiling with DNA arrays offers the opportunity to systematically identify such patterns (Halfon and Michelson 2002; Slonim 2002). Housekeeping genes are expressed in all cell types, whereas other genes are expressed in a more restricted selection of tissues. In previous research on the tissue specificity of genes, emphasis has mainly been on the extremes of one-tissues specific (Hsiao et al. 2001; Su et al. 2002) and housekeeping genes (Eisenberg and Levanon 2003; Lercher et al. 2002; Warrington et al. 2000). However, many genes may show midrange patterns of expression, i.e. are expressed at a high level in a subset of the tissues, and at a much lower level or not at all in other tissues. This term is related to the cross-tissue "breadth" of gene expression, rather than high or low overall expression intensities. Here, we investigate the occurrence and potential significance of midrange patterns of expression, noting that important information about a given tissue may be harbored not only in tissue-specific enhancement of expression, but also in tissue-specific suppression.

Some recent high-throughput DNA arrays studies of gene expression have been aimed at characterizing healthy tissue transcription patterns. One of these examined the transcription profiles in 28 normal human tissues and 45 mouse tissues, utilizing 12,000 oligonucleotide probesets (Su et al. 2002). cDNA arrays have also been used to examine expression of over genes across normal human tissues (Saito-Hisaminato et al. 2002). These, as well as other surveys on normal tissues (Haverty et al. 2002; Hsiao et al. 2001) were limited to only the more well-characterized genes, and did not afford a total genome-wide view. Studies on a more complete gene set focused on a

comparison between diseased and non-diseased states (Bakay et al. 2002; Iacobuzio-Donahue et al. 2002; Mariani et al. 2002). In a recent report (Shmueli et al. 2003), as well as in the current work, we queried 12 normal human tissues with a complete gamut of 62,839 probesets, representing 23,271 identifiable human genes. This is one of the largest sets employed to date, and includes nearly 12,000 genes whose tissue expression has not been examined by the earlier studies. Most recently, Su et al have extended their expression atlas to encompass 79 human and 61 mouse tissues (Su et al. 2004).

The resulting genome-wide view of gene expression patterns is used here to reveal relationships among healthy human tissues, as well as to generate new genome annotation tools. Specifically, our data shed new light on genes with midrange profiles of expression, with implications to the fine balance of gene expression and suppression that underlie tissue specification.

Systems and Methods

Expression data preprocessing. The expression intensity of mRNA was assayed across 5 microarrays (Affymetrix GeneChips U95A-E), containing a total of 62,839 probesets, each in duplicate. PolyA+ RNA samples from the human tissues were purchased from Clontech (Palo Alto, CA, details in Table 5 in the Supplementary Materials). This collection of major human tissues includes: bone marrow, brain, heart, kidney, liver, lung, pancreas, prostate, skeletal muscle, spinal cord, spleen and thymus. These RNA samples have relatively coarsely-defined tissue delineations but are compatible in this respect to those used in other studies of transcription patterns in a group of normal human tissues (Su et al., 2002, Su et al., 2004, Saito-Hisaminato et

al. 2002). Each RNA sample was typically composed of a pool of 10-25 individuals. While such commercial pooled samples from anonymous donors are demographically ill-defined, they are advantageous in enabling others to reproduce the experiments.

Replicate experiments were done independently, mostly from RNA of identical lot numbers. Exceptions are kidney, pancreas, and prostate. Aliquots of each sample (12 μg cRNA in 200 μl hybridization mix) were hybridized to a GeneChip Human Genome U95A-E array set (Affymetrix, Santa Clara, CA, USA). Preparation and hybridization of cRNA were done according to the manufacture's instructions (Affymetrix 2001).

The expression value for each gene was determined using the MicrroArray Suite version 5.0 (MAS 5.0) software (Hubbell et al. 2002; Liu et al. 2003) with default parameters, without using the MAS 5.0 scaling and normalizing procedures. The quantilization procedure used here (see below) encapsulates some features of a preprocessing method, RMA normalization (Irizarry 2003) not used here. Affymetrix MAS 5.0 signal values were normalized by taking the $\log_{10}$ of all values (substituting -1 for zero intensities) and then subtracting the mean for the particular array and adding the total experimental mean (Shmueli et al. 2003). Finally, intensities less than $\log_{10}30$ were set to $\log_{10}30$ to eliminate the perturbation by the noise present in the low intensities. Variations in this threshold resulted in no significant changes. The MAS 5.0 intensities, ranging on a decimal logarithmic scale from $\log_{10}30$ to roughly 4, were converted into a quantile scale. The expression data, averaged over the two replicates, were divided into 11 bins, whereby 10 equal density quantiles spanned the

values above $\log_{10}30$, and an 11^{th} " zero bin" included the remaining low intensity values. Henceforth, the quantiled profiles were used in the analysis.

Statistical analysis of differential expression. Single-classification ANOVA with equal sample sizes was employed on the preprocessed 24 element expression vector composed of 12 tissues in two replicates. For each tissue profile, the sum of the squares of the differences between the replicates was compared with the sum of the squares of the differences between the averages of the tissue expressions. To account for the multiple comparison problem inherent in calculating the P-values for all 62,839 probesets, we calculated the false discovery rate of the P-values (Benjamini and Hochberg 1995). We chose a 1% error rate, which gave a P-value cutoff of 0.0036. This resulted in 22,936 profiles that were defined as "differentially expressed". The remaining profiles were further divided into non-expressed profiles, defined as having all 12 values in the zero quantile, and housekeeping profiles, whose expression is non-zero quantile in all tissues and all intensities are of a similar value (standard deviation smaller than 1 quantile unit). The remaining profiles were defined as non-differentially expressed. The algorithms described below were deployed on the 22,936 differentially expressed profiles.

Probesets to genes analysis. The association of probesets to genes was performed using the GeneAnnot algorithm (Chalifa-Caspi et al. 2003a; Chalifa-Caspi et al. 2003b). GeneAnnot comprehensively identifies relationships between oligonucleotide array probesets and annotated genes in GeneCards (Safran et al. 2002) by performing pairwise alignments between the probe sequences and gene transcripts, and assigning sensitivity and specificity scores to them. A further step of probeset annotation,

conducted by GeneTide (Shklar et al. 2004), was to assign annotation based upon the transcript from which the probeset was derived. This was carried out by an integration of transcript annotation data from several resources such as UniGene (Wheeler et al. 2003) and AceView (http://www.humangenes.org). Furthermore, these target sequences were aligned against the human genome using BLAT (Kent 2002), and assigned a gene according to their genomic location using GeneLoc (Rosen et al. 2003).

Algorithms

Tissue specificity index. The index τ is defined as:

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

where N is the number of tissues and x_i is the expression profile component normalized by the maximal component value. For example, expression profile '0 8 0 0 0 2 0 2 0 0 0 0' is said to have $\tau = 0.95$. Other definitions, e.g. based on entropy or geometric considerations, were pursued but found less robust in terms of sensitivity to extreme profile component values.

Binary patterns. We first defined the 'gap' index for each expression profile as the maximum difference between the two neighboring values in the sorted quantile vector. When the same 'gap' was found more than once in a profile, the first gap, between the smaller neighboring values with that gap was taken. The 'gap' was used to convert expression profiles into binary form. For those 8,224 differentially expressed profiles with a 'gap' of at least 3, expression above the 'gap' was interpreted as over-expressed (1) and the rest as under-expressed (0). This set of 8,224

probeset profiles form our ‘mingap’ set. The remaining 14,712 differentially expressed profiles were classified to the best matching binary patterns detected by ‘gap’ as follows. The Euclidean distance was calculated between each of the 14,712 profiles and the mean expression profile of each of the binary patterns. The pattern to which this distance was smallest was selected as the matching binary pattern for the profile. The binary index, I_B, corresponding to each binary pattern is defined as the number of 1’s in the pattern.

Unsupervised clustering. The Superparamagnetic Clustering (SPC) algorithm (Blatt et al. 1996) was applied to the same set of profiles used in the binary pattern analysis. Before clustering, each profile was centered and normalized such that its mean was centered to zero, and its norm became one (as described by (Kannan et al. 2001)). The SPC parameters used are detailed in Table 5 (Supplementary materials).

Ancestral tissue reconstructions. Given two binary tissue expression profiles, an ancestor profile was inferred by first assuming that instances of agreement (both 1’s or both 0’s) are unaltered in the ancestor. In the disagreement cases (1 and 0, or 0 and 1), maximum parsimony is applied, with a majority call of expression in the remaining tissues. Our method for inferring the ancestors of each node in a dendrogram including the deep internal nodes, involved following the linkages of the hierarchically clustered tree and successively inferring each node.

Availability. All analyses were implemented in Matlab (www.mathworks.com). Scripts and intermediate data are all available upon request.

Implementation

Expression profile categorization. Expression profiles were generated for a set of 12 representative normal human tissues (Fig. 1). This was done with a total of 62,839 oligonucleotide probesets, of which nearly 75% corresponded to annotated human genes, encompassing 23,271 GeneCards entries (Safran et al. 2002), and the rest could not be associated with currently known gene-related sequences (Table 1). The 50,214 probesets included in the four less commonly used arrays U95B-E provided novel expression information on 11,418 GeneCards genes. This genome-wide view of human tissue expression patterns is available in the GeneNote database (Shmueli et al. 2003) http://genecards.weizmann.ac.il/genenote/. The expression profiles were classified into four categories: differentially expressed, housekeeping, unexpressed and uncategorized (Fig. 1 and Table 1). It is seen that a majority (~90%) of the probesets in the first two categories are related to known genes, while most of the unannotated probesets are included in the last two categories, as expected.

Distribution of tissue specificities. To examine the complete expression pattern diversity, we developed a Tissue Specificity Index, τ, a quantitative, graded scalar measure of the specificity of an expression profile. τ values interpolate the entire range between 0 for housekeeping genes and 1 for strictly one-tissue specific genes. It is seen (Fig. 2A) that τ values near 0 and 1 tend to be more probable than the intermediate values, generating a U-shaped distribution. However, as many as 57% of all profiles have intermediate specificities: $0.15 \leq \tau \leq 0.85$, constituting the largest group, greater than the housekeeping and one-tissue specific sets combined.

To evaluate the robustness of the shape of the τ distribution to additional tissues and another organism, we calculated the same distributions for a previously published set (Su et al. 2002) where 27 human and 45 mouse normal tissues with replicates were analyzed for one fifth of the gene representations examined here. We found that the shape of the τ distributions was largely similar in all three data sets. Indeed, nearly identical percentages of profiles with intermediate specificity ($0.15 \leq \tau \leq 0.85$) are detected: 56% for mouse and 57% for human.

Do our tissue specificity (τ) estimates from 12 tissues scale up when a more comprehensive number of tissues is examined? A recently published study (Su et al. 2004) provides human gene expression profiles across 74 non-cancerous human tissues. We found a high correlation (R=0.85) between the τ indices of genes across the two datasets for differentially expressed genes (Fig. 3). Two clusters of τ values differ markedly: low τ in GeneNote, high τ in the new study, and vice versa. The former correspond to genes specific to tissues not present in GeneNote, and the latter to spleen, not present in the more recent study. Congruence between the tissue specificity values based upon 12 and 74 tissues demonstrates the power of our newly defined tissue specificity index, and shows that our choice of tissues is fairly representative of the complete tissue-set transcriptome. An analysis of the distribution of τ values for the new data set (Fig. 2A, green line) shows a relatively high preponderance of intermediate τ values, likely stemming from the use of sub-tissues such as different brain regions.

Binary expression patterns. The one-dimensional tissue specificity index is limited in its capacity to identify and categorize specific classes of expression patterns. To

overcome this, we developed a procedure that converts an arbitrary expression profile into a binary pattern. The quantiled expression profiles are mapped from a very large set of the cardinality 11^{12} (more than 3.1 billion) to a reduced set of $2^{12} = 4{,}096$ possible patterns. This analysis was initially performed on a subset of 8,224 probeset expression profiles that fulfilled a specific intensity gap criterion (the mingap set, see Algorithms).

Of the possible 4,094 binary patterns (excluding the all-0 and all-1 patterns), 859 were actually observed in this set. The probesets of the first microarray (U95A) detected only 498 of these patterns, while the remaining 42% of the patterns were found only on the four additional arrays (U95B-E). Further, the four additional arrays strengthen 127 patterns that were populated by only one profile in the first array. Subsequently, the differentially expressed profiles not included in the mingap set were binarized by matching each one to its closest binary counterpart.

The results of the binarization are shown in Fig 4. The different panels 4.i (i=1 to 12) have profiles (parsed from among the 8,224 gene representations of the differentially expressed and 4,216 housekeeping genes) with high expression in i tissues and under-expression in 12 minus i tissues. Panel 4.12 contains the strictly-defined housekeeping genes. In panel 4.1 (single-tissue specificity), brain, bone marrow, pancreas, skeletal muscle and liver are more highly represented, while spinal cord, kidney, heart, and spleen have relatively few profiles. In panel 4.2, prevalent two-tissue specific patterns are brain and spinal cord, heart and skeletal muscle, bone marrow and thymus, and kidney and liver. Bone marrow, spleen, and thymus tissues define a major three-tissue pattern in panel 4.3. Panels 4.9 to 4.11 depict profiles with expression in all but 3, 2 or

1 tissue(s), respectively. Notably, the same five tissues with the most single tissue specific profiles (brain, bone marrow, pancreas, muscle, and liver) also have the greatest number of single tissue suppressed profiles.

Of the individual profiles appearing in Fig. 4, 5,220 are not well annotated to any known gene and should therefore be considered interesting, as their function can be preliminarily ascertained based upon their expression profile. Table 2 (Supplementary Material) shows the expression profile along with the identifier of the sequence from which the probeset was derived.

We subsequently defined the 99 most populated binary patterns among the 22,936 differentially expressed genes as such that represent at least 25 probeset profiles, including the housekeeping (all 1's) and null (all 0's) patterns (Fig. 2B). The number of populated binary patterns in each binary index, I_B, shows a clear bimodal distribution (Fig. 2C), with peaks at binary index values of 2 and 10. This behavior reflects the same bimodal trend seen for τ values in Fig. 2A. Whereas all 12 one-tissue specific patterns are included, only about one third of the two-tissue expressed patterns ($I_B = 2$) and about a quarter of the two-tissue repressed patterns ($I_B = 10$) are included in this set, suggesting biases towards specific oligo-tissue combinations. The peak at high I_B values in Fig.2C corresponds to profiles with low expression in 1 to 3 tissues and high expression in the others. We use the term suppression to describe instances where genes are expressed at lower levels in a few tissues. This does not necessarily imply an active process where the expression of a gene is specifically turned off. It could equally be due to a loss of activation in expression or a dilution of mRNA levels in one tissue relative to another, due to a different cellular

composition".

To test the validity of the supervised binary clustering, we also applied unsupervised Superparamagnetic clustering (SPC) (Blatt et al. 1996; Getz et al. 2000) to our data (Fig. 5A). SPC is suitable for the clustering of gene expression profiles due to its stability against noise and the inherent measure of cluster stability (Getz et al. 2000). The identified seventy SPC clusters showed a strong correlation with the 97 binary clusters (Fig. 5B). Some binary patterns are represented by multiple SPC clusters, thus serving to further refine the relevant binary patterns (Fig. 5C). The high level of overall agreement between the two clustering methods lends additional credence to the binary classification proposed here.

Tissue relationships based upon the expression repertoire. Inter-tissue distances were calculated between pairs of tissue vectors, each containing the 22,936 expression values of the set of differentially expressed profiles. The resulting tissue dendrogram (Fig. 6A) shows a specific set of groupings relating to different degrees of inter-tissue similarities. The dendrogram reveals a set of tissue relationships that is consistent with previous knowledge (Hsiao et al. 2001). The immunological tissues, bone marrow, thymus and spleen, along with the lung, cluster together. Pairs of related tissues are also coupled in the dendrogram: heart and skeletal muscle, kidney and liver, brain and spinal cord, and prostate and pancreas.

To isolate those profiles that specify the underlying relationships among the tissues, we split the differentially expressed profiles into two groups: those with I_B=1 and those with I_B= 2 to 11. We found that the tree based upon the second group, with

midrange profiles (Fig. 6B) recovers the most important features of the dendrogram based on the entire set: a united nervous system, muscle tissues juxtaposed, and immune system mutually coherent. In contrast, the dendrogram based upon the I_B=1 group (Fig. 6C) is very different and appears much more removed from known tissue relationships. For example, the spinal cord is closest to heart and very distant from brain. One could argue that the visible non-zero off-diagonal values in the I_B=1 patterns (Fig. 4.1) would contribute sufficient information, so as to generate a more biologically-realistic tissue dendrogram.

Inferring ancestral tissue profiles. The availability of genome-wide expression profiles for each of the tissues provides a unique opportunity to obtain additional information regarding mutual relationships among different tissues. Specifically, it is possible to derive from the tissue dendrogram an inferred gene expression profile for each of the "ancestral" tissues represented by the internal nodes (Fig. 7A). As an example, it is seen (Fig. 7B) that in most cases of expression in brain but not in spinal cord, under-expression is inferred for the ancestral tissue, reflecting *de novo* specificities for brain. On the other hand, most of the high expressions found in spinal cord but not in brain are also positive in the inferred ancestral tissue, suggesting that the difference corresponds to brain-specific suppressions. More generally, we found that for most ancestral tissues, including all but one of the most closely-related doublets, the tissue with more genes showing novel expression also exhibits more genes with novel suppressions (Fig. 7C). This phenomenon is also gleaned by visual inspection of panels 1 and 11 of Fig. 4, as described above.

Discussion

This paper proposes a set of novel genome-wide specific annotation tools. First, each of the 23,271 genes targeted by 46,185 probesets (Table 1) has one or more tissue expression profiles, documented in GeneNote. A set of tools has been developed to allow one to generate a consensus expression pattern for each of these genes, with the exclusion of outliers. Second, every gene is marked with a specific value of τ, identifying it as belonging to a particular range on a graded tissue specificity measure between extreme tissue specificity and a complete absence of such specificity. Third, a gene with a differentially expressed profile is related to a binary pattern, indicating the combination of tissues in which it is more highly expressed and suppressed. We believe that these binary patterns are more amenable to intuitive scientific interpretation than classification based on standard clustering algorithms. It is reassuring, though, that a high degree of correlation is demonstrated between the two systems. All the above information provides tools for assigning potential function to novel and hitherto un-annotated genes. As the annotation tools presented here are easily generalized, we believe they can be fruitfully applied to a wide spectrum of datasets, for example to sets with tumor and non-tumor samples.

The binary pattern analysis is particularly useful in revealing expression profiles that constitute unusual tissue combinations. For example, the pattern number 36 in Fig. 2B denotes high expression in bone marrow, pancreas and liver; pattern number 47 denotes high expression in heart, prostate and spinal cord. In general, among the 98 binary patterns of Fig. 2B that show expression in at least one tissue, one pattern corresponds to the housekeeping expression profile, and another 12 denote single-

tissue specific profiles. The remaining 85 patterns are defined here as denoting midrange profiles of expression. Of these, a maximum of 33 patterns may be considered as consistent with tissue clustering as defined by the dendrogram of Fig. 6A, as they correspond to the groups of tissues defined by the terminal and internal nodes of the dendrogram. Thus, a majority of the dominant binary patterns corresponding to midrange profiles may be viewed as unexpected. Such patterns are difficult to explain in terms of tissue similarities, including the sharing of common cell types among disparate tissues. Alternatively, there may be yet undiscovered underlying transcription control mechanisms that could be discerned by future research. Some such unexpected expression patterns may be a neutral mode of expression (Khaitovich and et al. 2004; Yanai et al. 2004).

The approach explored here focuses on midrange profiles of transcription, with elevated expression/suppression in specific tissue combinations, and intermediate values of the tissue specificity index τ. Our analysis has revealed that midrange profiles constitute a majority of the tissue specificity expression patterns. Despite its ubiquity, this category has received remarkably little attention relative to its housekeeping and tissue-specific counterparts. Of the nearly 100 most populated binary patterns, more than 80% are midrange patterns. A recent expression study in maize has also shown that a relatively small portion of genes tend to be organ specific while the remaining show diverse expression (Cho et al. 2002).

Most focused arrays with specific subsets of genes used by various authors contain mostly tissue-specific genes whose level is elavated in a single tissue. Such arrays may be considered "too focused". Our results and analyses, suggesting the importance

of genes with midrange expression profiles, could have serious impact in terms of array design planning as well as experimental design planning.

A dominant property of midrange profiles is the surprising preponderance of patterns with tissue specific gene suppression (I_B=9 to 11), which are almost as populated as oligo-expression patterns (I_B=2 to 4). The most underrepresented set of profiles are the midrange profiles with I_B =5 to 7. Our results also indicate that in the evolution of a tissue, *de novo* expression and *de novo* suppression go hand in hand.
It thus appears that gene suppression plays a major role in tissue evolution and is tightly coupled with novel expression in the origin of distinct tissues. Such tissue-specific gene suppression may be mediated by specific pathways of transcription control (Hsia and McGinnis 2003), as well as by other cellular mechanisms, including those mediated by RNA interference (Cerutti 2003). One practical conclusion related to tissue-specific arrays is that these should preferably contain, in addition to single-tissue specific genes, also genes that manifest more complex patterns of expression-suppression.

Conclusion

Understanding the signaling and control pathways that govern organ development during ontogeny constitutes a fundamental problem of developmental biology (Burgess et al. 2002). Studies in model organisms such as *Drosophila* have demonstrated the complex interplay of signaling molecules that underly developmental events associated with the embryonic maturation of tissues and cell types (St Johnston 2002). The exact spatial and temporal expression of genes, and the interaction of their protein products elicit a developmental code of organ commitment

and early patterning. This code likely manifests itself in the pattern of gene expression in each of the tissues. Furthermore, when new tissues are formed in ontogeny or phylogeny, their ancestral precursors should have their own expression patterns, complexly related to those of the more highly differentiated derived tissues. To validate this concept in the future, direct experimental testing of expression patterns at early stages in embryogenesis will be required. Our analysis of ancestral tissue expression, which points to a correlation between novel tissue expression and suppression, and the availability of a tissue dendrogram relating to the full gamut of genes of the human genome, can serve as a valuable tool for such studies.

Acknowledgements

This work was made by possible by the generosity of the Abraham and Judith Goldwasser Foundation. It was further supported by the Crown Human Genome Center. IY is a Koshland Scholar, DL is the incumbent of the Ralph and Lois Silver Chair in Human Genomics, and ED is the incumbent of the Henry J. Leir Professorial Chair.

References

Affymetrix. 2001. Microarray Suite User Guide, Version 5. *Affymetrix, http://.com/support/technical/manuals.affx*.

Bakay, M., Zhao, P., Chen, J., and Hoffman, E.P. 2002. A web-accessible complete transcriptome of normal human and DMD muscle. *Neuromuscul Disord* **12 Suppl 1:** S125-141.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B* **57:** 289-300.

Blatt, M., Wiseman, S., and Domany, E. 1996. Superparamagnetic clustering of data. *Physical Review Letters* **76:** 3251-3254.

Burgess, R., Lunyak, V., and Rosenfeld, M. 2002. Signaling and transcriptional control of pituitary development. *Curr Opin Genet Dev* **12:** 534-539.

Cerutti, H. 2003. RNA interference: traveling in the cell and gaining functions? *Trends Genet* **19:** 39-46.

Chalifa-Caspi, V., Shmueli, O., Benjamin-Rodrig, H., Rosen, N., Shmoish, M., Yanai, I., Ophir, R., Kats, P., Safran, M., and Lancet, D. 2003a. GeneAnnot: Interfacing GeneCards with high throughput gene expression compendia. *Brief. Bioinformatics* **in press**.

Chalifa-Caspi, V., Yanai, I., Ophir, R., Rosen, R., Shmoish, M., Benjamin-Rodrig, H., Iny-Stein, T., Shmueli, O., Safran, M., and Lancet, D. 2003b. GeneAnnot: Comprehensive two-way linking between oligonucleotide array probesets and GeneCards genes. **in press**.

Cho, Y., Fernandes, J., Kim, S.H., and Walbot, V. 2002. Gene-expression profile comparisons distinguish seven organs of maize. *Genome Biol* **3:** research0045.

Eisenberg, E. and Levanon, E.Y. 2003. Human housekeeping genes are compact. *Trends Genet* **19:** 362-365.

Getz, G., Levine, E., and Domany, E. 2000. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci U S A* **97:** 12079-12084.

Halfon, M.S. and Michelson, A.M. 2002. Exploring genetic regulatory networks in metazoan development: methods and models. *Physiol Genomics* **10:** 131-143.

Haverty, P.M., Weng, Z., Best, N.L., Auerbach, K.R., Hsiao, L.L., Jensen, R.V., and Gullans, S.R. 2002. HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res* **30:** 214-217.

Hsia, C.C. and McGinnis, W. 2003. Evolution of transcription factor function. *Curr Opin Genet Dev* **13:** 199-206.

Hsiao, L.L., Dangond, F., Yoshida, T., Hong, R., Jensen, R.V., Misra, J., Dillon, W., Lee, K.F., Clark, K.E., Haverty, P. et al. 2001. A compendium of gene expression in normal human tissues. *Physiol Genomics* **7:** 97-104.

http://www.humangenes.org.

Hubbell, E., Liu, W.M., and Mei, R. 2002. Robust estimators for expression analysis. *Bioinformatics* **18:** 1585-1592.

Iacobuzio-Donahue, C.A., Maitra, A., Shen-Ong, G.L., van Heek, T., Ashfaq, R., Meyer, R., Walter, K., Berg, K., Hollingsworth, M.A., Cameron, J.L. et al. 2002. Discovery of novel tumor markers of pancreatic cancer using global gene expression technology. *Am J Pathol* **160:** 1239-1249.

Irizarry, R., Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP. 2003. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* **4:** 249-264.

Kannan, K., Kaminski, N., Rechavi, G., Jakob-Hirsch, J., Amariglio, N., and Givol, D. 2001. DNA microarray analysis of genes involved in p53 mediated apoptosis: activation of Apaf-1. *Oncogene* **20:** 3449-3455.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Khaitovich, P. and et al. 2004. A Neutral Model of Transcriptome Evolution. *PLoS Biology* **2:** (in press).

Lercher, M.J., Urrutia, A.O., and Hurst, L.D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31:** 180-183.

Liu, G., Loraine, A.E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D., and Siani-Rose, M.A. 2003. NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res* **31:** 82-86.

Mariani, T.J., Budhraja, V., Mecham, B.H., Gu, C.C., Watson, M.A., and Sadovsky, Y. 2002. A variable fold-change threshold determines significance for expression microarrays. *Faseb J.*

Rosen, N., Chalifa-Caspi, V., Shmueli, O., Adato, A., Lapidot, M., Stampnitzky, J., Safran, M., and Lancet, D. 2003. GeneLoc: exon-based integration of human genome maps. *Bioinformatics* **19 Suppl 1:** I222-I224.

Safran, M., Solomon, I., Shmueli, O., Lapidot, M., Shen-Orr, S., Adato, A., Ben-Dor, U., Esterman, N., Rosen, N., Peter, I. et al. 2002. GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18:** 1542-1543.

Saito-Hisaminato, A., Katagiri, T., Kakiuchi, S., Nakamura, T., Tsunoda, T., and Nakamura, Y. 2002. Genome-wide profiling of gene expression in 29 normal human tissues with a cDNA microarray. *DNA Res* **9:** 35-45.

Shklar, M., Shmueli, O., Strichman-Almashanu, L., Shmoish, M., Iny-Stein, T., Safran, M., and Lancet, D. 2004. Terra Incognita Discovery Endeavor. Comprehensive EST assignment to GeneCards genes. *Presented at ISMB*.

Shmueli, O., Horn-Saban, S., Chalifa-Caspi, V., Shmoish, M., Ophir, R., Benjamin-Rodrig, R., Safran, M., Domany, E., and Lancet, D. 2003. GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biologies* **326:** 1067-1072.

Slonim, D.K. 2002. From patterns to pathways: gene expression data analysis comes of age. *Nat Genet* **32 Suppl:** 502-508.

St Johnston, D. 2002. The art and design of genetic screens: Drosophila melanogaster. *Nat Rev Genet* **3:** 176-188.

Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A. et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A* **99:** 4465-4470.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101:** 6062-6067.

Warrington, J.A., Nair, A., Mahadevappa, M., and Tsyganskaya, M. 2000. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol Genomics* **2:** 143-147.

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. et al. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res* **31:** 28-33.

Yanai, I., Graur, D., and Ophir, R. 2004. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics* **8:** 15-24.

Legends to Figures

Figure 1: Gene expression profiles across 12 normal human tissues.

Classification of the 62,839 expression profiles (horizontal lines) into four groups indicated by the right bar: HK, housekeeping; DE, differentially expressed; UC, uncategorized; NE, not expressed (Table 1). The left bar indicates the origin from Array A (peach) and Arrays B-E (brown). The profiles within each category were sorted in ascending order according to τ, the tissue specificity index. Expression intensities are color coded by quantile values on the bottom bar. Tissue abbreviations: BRN, Brain; SPC, Spinal cord; BMR, Bone marrow; SPL, Spleen; TMS, Thymus; LNG, Lung; PNC, Pancreas; PST, Prostate; HRT, Heart; MSL, Skeletal muscle; KDN, Kidney; LVR, Liver.

Figure 2: Tissue specificity index and expression pattern repertoire.

A. Distribution of τ values for 27,152 profiles which include the 22,936 differentially expressed and 4,216 housekeeping profiles (bars). The τ distributions are also shown for the 12,626 profiles across 74 human tissues (green curve) from a recent study (Su et al. 2004), as well as 27 human tissues (red curve), and 12,654 across 45 mouse tissues (yellow curve) from another set (Su et al. 2002). **B.** A summary representation of the most populated binary patterns (columns), where blue circles indicate high expression. The patterns, enumerated on the abscissa, are sorted according to binary value. Tissue abbreviations as in Fig. 1. **C.** The frequency distribution of I_B values of the binary patterns shown in (B). The green curve indicates the expected distribution following a random binomial model.

Figure 3: Comparison of tissue specificity indices (τ) in different data sets. The τ indices of the current work (GeneNote) were compared with those of a recently published set with 74 human tissues (Su et al. 2004). The two sets were compared based upon the probeset mappings (U95 and U133) released by Affymetrix, where if more than one U133 probeset could be matched for a given U95 probeset, only one was taken. This restriction resulted in 13,124 probeset pairs. The expression intensities of the two sets were normalized by quantile normalization, and subsequently the mean of each expression profile for each set was scaled to the total mean of the profile. Replicates were averaged and signal quantilization was carried out as described in Methods. Shown are the τ pairs for the 9,450 differentially expressed genes of our set.

Figure 4: Profile clustering in 12 binary pattern sets. Each panel contains all mingap set expression profiles with binary patters having the I_B value indicated on top, sorted by the pattern's binary value. Housekeeping profiles are taken as $I_B = 12$. The profiles of each pattern were further sorted according to the mean of expression values corresponding to the non-zero elements in the corresponding binary pattern. The color code for expression intensity is as in Fig. 1. Table 3 (Supplementary Material) specifies the accession identifiers for each expression profile shown.

Figure 5: Superparamagnetic clustering (SPC). A. A summary representation of the results of superparamagnetic clustering (SPC). Each

column (enumerated on the abscissa and sorted by gap-computed binary values) represents an average pattern of all profiles in a cluster. The predefined SPC minimal cluster size was set to 10 profiles. **B.** Each element in the matrix is a comparison between the profiles of an SPC cluster and those of a binary cluster. Binary clusters were obtained for the 8,224 profiles of the mingap set, and only the patters with at least 10 profiles are shown. The color-coded score (right bar) is calculated as the number of shared profiles between the two cluster patterns, divided by the size of the smaller of the two clusters. **C.** Augmented view of 11 individual SPC clusters. Centered and normalized quantiled expression profiles of the cluster's members are shown. The clusters 2, 23 and 65 (corresponding to their rows in (B)) manifest expression in both brain and spinal cord; 2 is equally expressed in both tissues, 23 is higher in the brain and 65 is higher in spinal cord. Similar relationships are seen in the cluster triplets 12, 42 and 59, expressed in both heart and skeletal muscle, and 15, 36 and 46, expressed in both liver and kidney (not shown). Table 4 (Supplementary Material) specifies the accession identifiers for all profiles clustered by SPC.

Figure 6: Midrange specificity profiles and the tissue dendrograms. Tissue dendrograms, based upon a hierarchical clustering using average linkage of the differentially expressed genes. **A,** based on all differentially expressed (DE) profiles; **B,** based upon the 4,921 mingap profiles with I_B = 2-11; **C,** based upon 3,303 profiles with I_B = 1. A systematic analysis of this trend (data not shown) suggests that the contribution to the dendrogram from profiles with high binary values (I_B = 8-11) is

greater than that from the symmetrically disposed patterns with low binary index (I_B = 2-5).

Figure 7: Ancestral tissue patterns. A. The tissue dendrogram, based upon the differentially expressed genes, is shown with superimposed tissue vectors and inferred ancestral tissue vectors based upon the binary profiles of the mingap set. **B.** Two instances of ancestral tissue (ANC) inferences for brain and spinal cord (left) and for kidney and liver (right). **C.** Correlation of novel tissue expression and novel tissue suppression of profiles with reference to their inferred ancestor. For each internal ancestral node, a ratio was calculated between the novel expressions in the two derived tissue vectors with respect to the ancestral tissue vector (abscissa). Similar ratios were also calculated for the novel suppressions (ordinate). Points represent all internal nodes, with tissue pair order selected to have an abscissa value higher than 1. Only two of the internal nodes (labeled red and orange, similarly marked in panel A) had an ordinate value lower than 1, indicating lack of correlation. The node corresponding to the brain spinal-cord ancestor is highlighted in green.

Table 1

	Differentially Expressed	Housekeeping	Not Expressed	Uncategorized	Total
GeneCards	20,589	3,181	9,859	12,726	46,185
Not Annotated	2,347	1,035	6,502	6,600	16,654
Total	22,936	4,216	16,361	19,326	62,839

Partition of U95A-E probesets into four expression categories (Fig. 1). The probesets to GeneCards associations were done using the GeneAnnot algorithm (Chalifa-Caspi et al. 2003b), and annotation from the original transcripts from which the probesets were derived (see Systems and Methods).

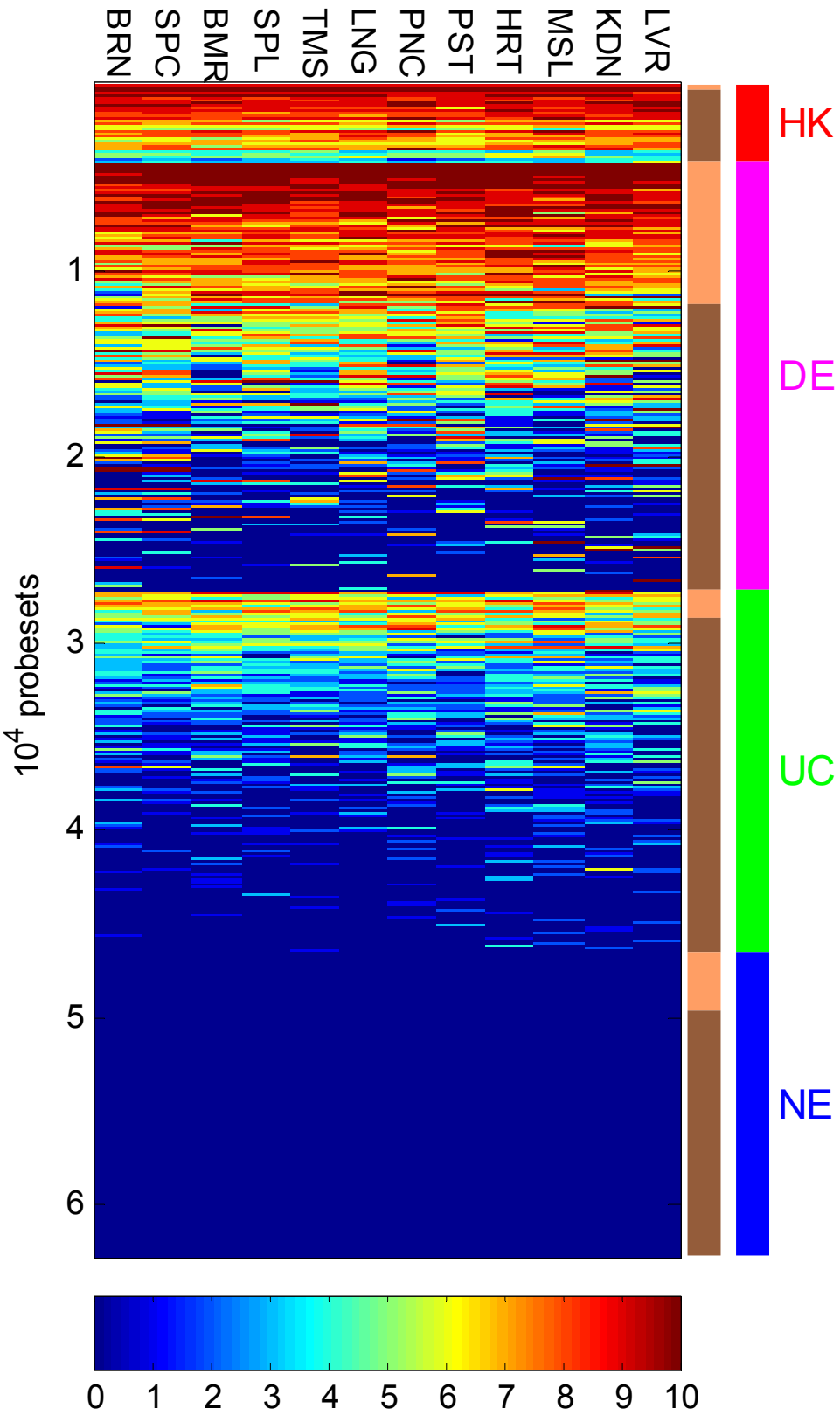
BRN
SPC
BMR
SPL
TMS
LNG
PNC
PST
HRT
MSL
KDN
LVR
HK
DE
UC
NE
10^4 probesets
1
2
3
4
5
6
0 1 2 3 4 5 6 7 8 9 10

Figure 1

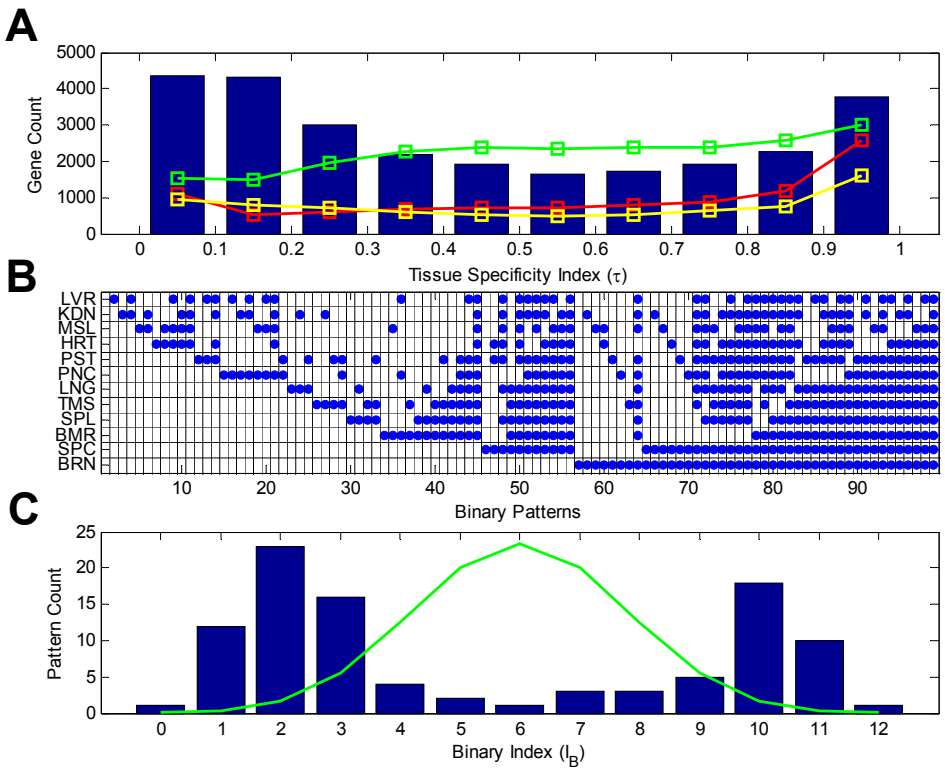
A
Gene Count
5000
4000
3000
2000
1000
0
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
Tissue Specificity Index (τ)
B
LVR
KDN
MSL
HRT
PST
PNC
LNG
TMS
SPL
BMR
SPC
BRN
10 20 30 40 50 60 70 80 90
Binary Patterns
C
Pattern Count
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9 10 11 12
Binary Index (I_B)

Figure 2

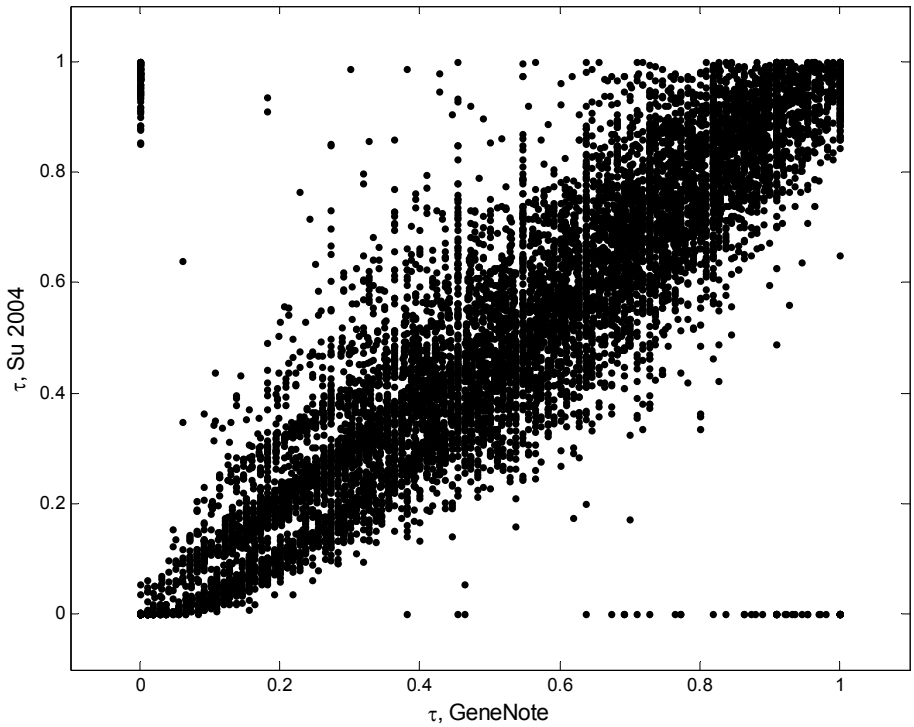

Figure 3

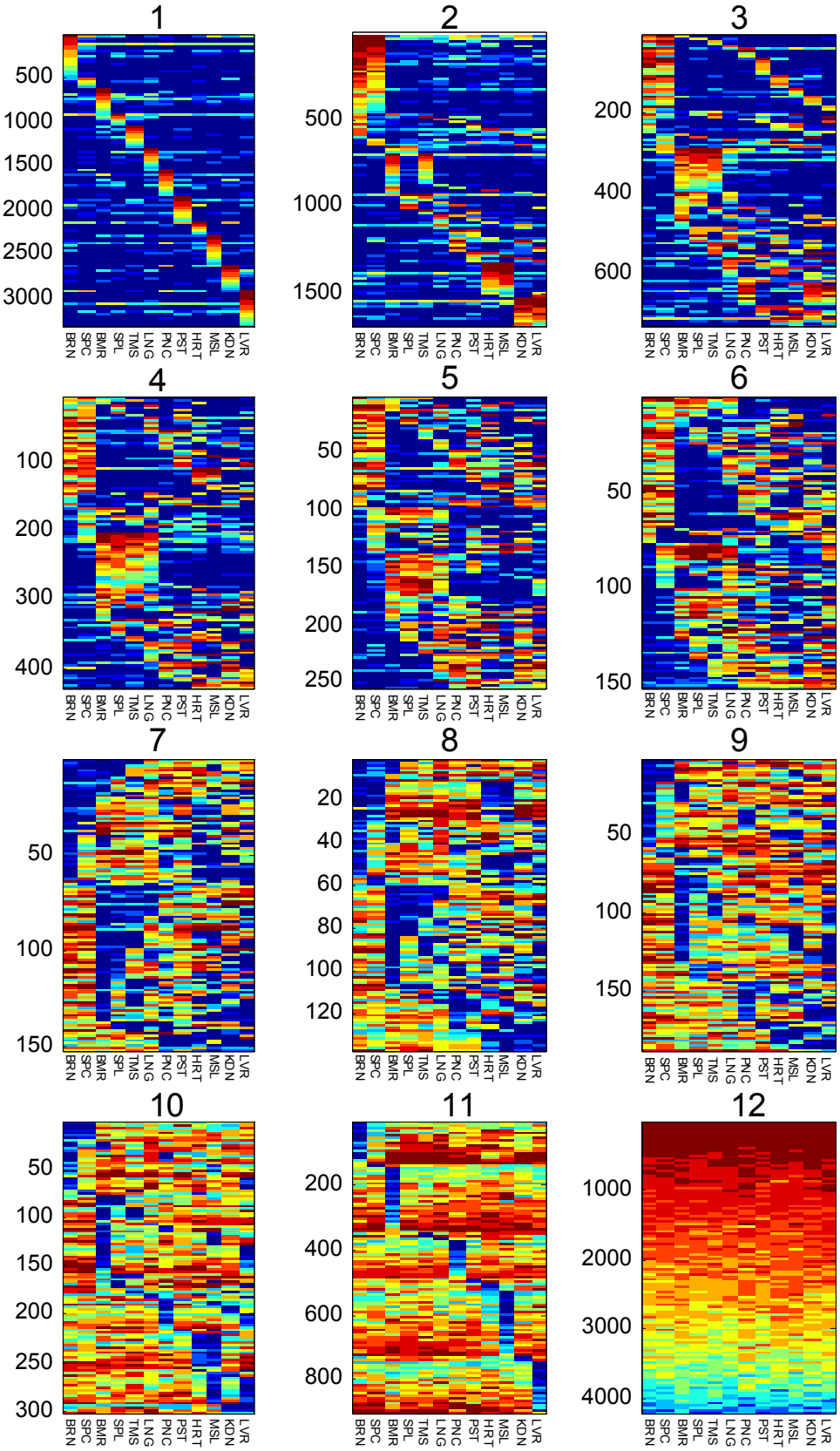
1
2
3
4
5
6
7
8
9
10
11
12
BRN SPC SPC BMR SPL TMS LNG PNC PST HRT MSL KDN LVR

Figure 4

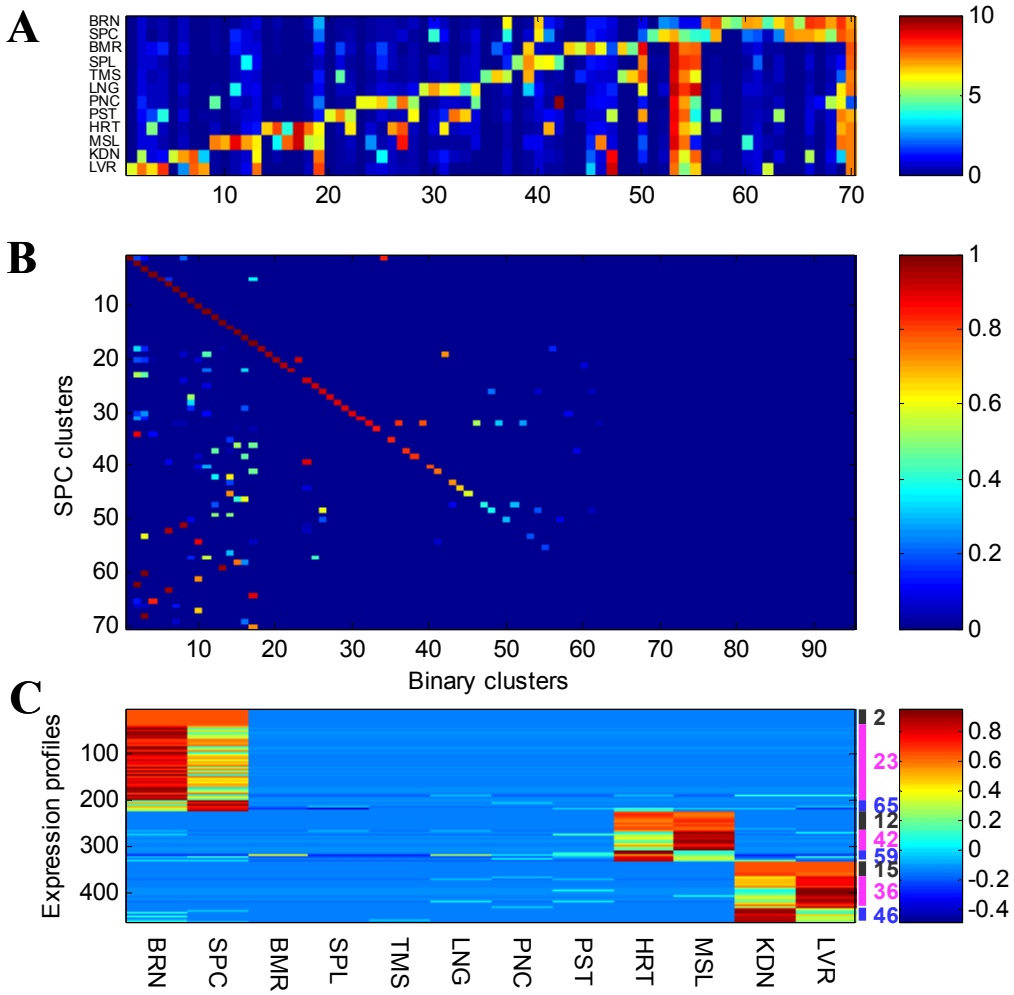
A
BRN
SPC
BMR
SPL
TMS
LNG
PNC
PST
HRT
MSL
KDN
LVR
10
20
30
40
50
60
70
10
5
0
B
SPC clusters
Binary clusters
10
20
30
40
50
60
70
80
90
1
0.8
0.6
0.4
0.2
0
C
Expression profiles
100
200
300
400
2
23
65
12
42
59
15
36
46
0.8
0.6
0.4
0.2
0
-0.2
-0.4
BRN
SPC
BMR
SPL
TMS
LNG
PNC
PST
HRT
MSL
KDN
LVR

Figure 5

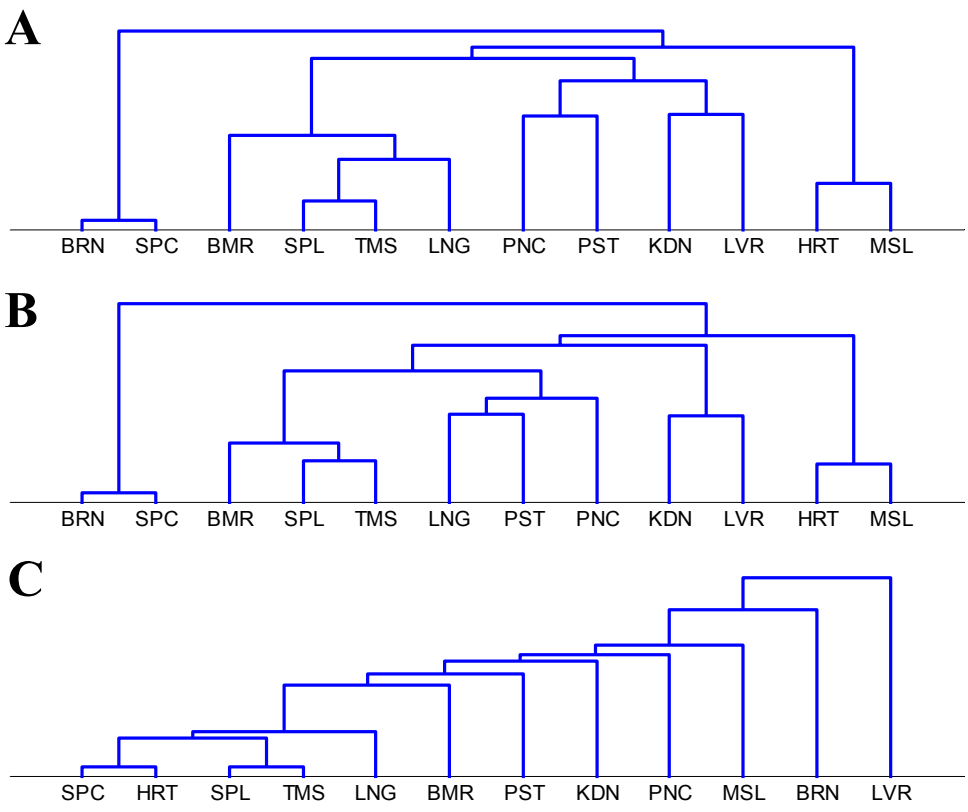
A
BRN SPC BMR SPL TMS LNG PNC PST KDN LVR HRT MSL
B
BRN SPC BMR SPL TMS LNG PST PNC KDN LVR HRT MSL
C
SPC HRT SPL TMS LNG BMR PST KDN PNC MSL BRN LVR

Figure 6

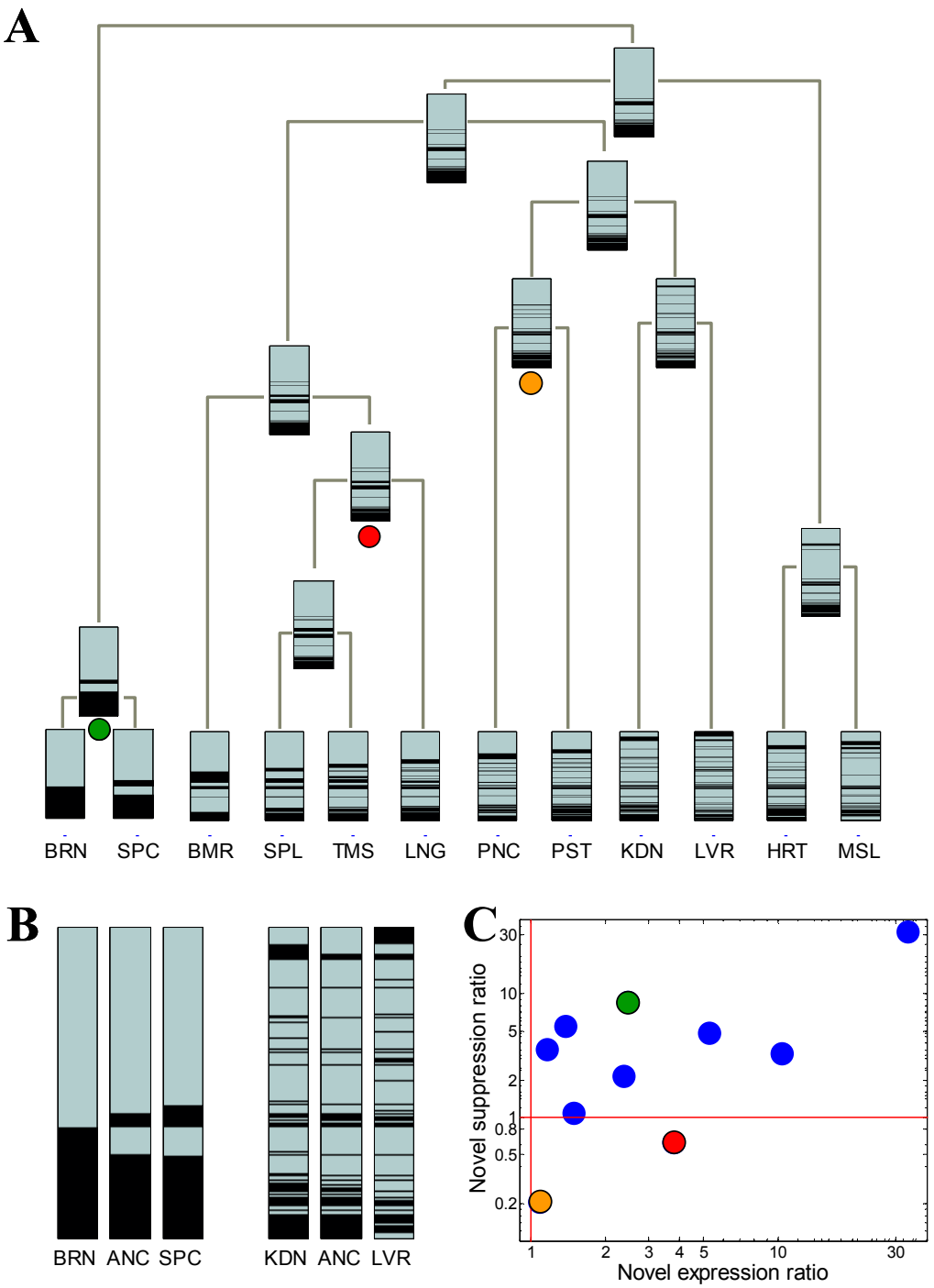
A
BRN
SPC
BMR
SPL
TMS
LNG
PNC
PST
KDN
LVR
HRT
MSL
B
BRN ANC SPC
KDN ANC LVR
C
Novel suppression ratio
30
10
5
2
1
0.8
0.5
0.2
1
2
3
4
5
10
30
Novel expression ratio

Figure 7