Original Research

Medusa structure of the gene regulatory network: dominance of transcription factors in cancer subtype classification

Yuchun Guo[1,2], Ying Feng[1], Niraj S Trivedi[1,3] and Sui Huang[1,4]

[1]Vascular Biology Program, Children's Hospital, Harvard Medical School, Boston; [2]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge; [3]Bioinformatics Program, Boston University, Boston, MA, USA; [4]Institute for Biocomplexity and Informatics, University of Calgary, Calgary, Alberta, Canada

Corresponding author: Sui Huang, Institute for Biocomplexity and Informatics, Biological Sciences Bldg, Rm 539D, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4. Email: sui.huang@ucalgary.ca

Abstract

Gene expression profiles consisting of ten thousands of transcripts are used for clustering of tissue, such as tumors, into subtypes, often without considering the underlying reason that the distinct patterns of expression arise because of constraints in the realization of gene expression profiles imposed by the gene regulatory network. The topology of this network has been suggested to consist of a regulatory core of genes represented most prominently by transcription factors (TFs) and microRNAs, that influence the expression of other genes, and of a periphery of 'enslaved' effector genes that are regulated but not regulating. This 'medusa' architecture implies that the core genes are much stronger determinants of the realized gene expression profiles. To test this hypothesis, we examined the clustering of gene expression profiles into known tumor types to quantitatively demonstrate that TFs, and even more pronounced, microRNAs, are much stronger discriminators of tumor type specific gene expression patterns than a same number of randomly selected or metabolic genes. These findings lend support to the hypothesis of a medusa architecture and of the canalizing nature of regulation by microRNAs. They also reveal the degree of freedom for the expression of peripheral genes that are less stringently associated with a tissue type specific global gene expression profile.

Keywords: gene regulatory network, medusa network, transcription, gene expression pattern, core network

Experimental Biology and Medicine 2011; **236**: 628–636. DOI: 10.1258/ebm.2011.010324

Introduction

Genome-scale gene expression profiles are informative combinatorial configurations of the expression status of individual genes across the genome. Such profiles are efficiently assessed as 'transcriptomes' at the level of mRNA expression using DNA microarrays. Transcriptomes have widely been used to discriminate between tumor subtypes and phenotypic properties, such as diagnostic or prognostic features of tumor samples.[1–3] Herein, m samples (expression profiles) are clustered based on their n attributes which represent the expression level of the same set of n genes. However, the very existence of specific, informative gene expression profiles that underlie the distinct characteristics of tissues is often taken for granted. It is clear that the set of biologically stable and observable transcriptomes constitute only a tiny fraction of the universe of combinatorically possible configurations of gene expression in a genome.[4] This collapse of the vast space of theoretical gene expression configurations is a manifestation of the fact that expression of individual genes is not independent from each other. Instead, genes form a genome-wide network of regulatory interactions which massively constrains the vast gene expression space to the subset of biologically realizable expression profiles.[5] For instance, if gene A is an unconditional inhibitor of expression of gene B, then all profiles that have A and B both highly expressed would be 'forbidden' in the normal tissue.

The elementary process by which genes directly interact with each other is transcriptional regulation in which a 'regulating' gene encodes a transcription factor (TF) which binds to the promoter and/or enhancer region of a regulated ('target') gene and influences expression of the latter. Every gene is a 'target' gene but not every gene is regulating other genes. In a simple dichotomy, thus a gene can be either a TF or a non-regulating protein, such as a structural protein or a metabolic enzyme. Let us call the latter group of non-regulating genes as 'effector genes'. Since each gene, be it a TF or an effector gene, is a target gene of some other TFs, in a first approximation, the set of TF genes can be regarded as controlling the entire gene expression profile

throughout the genome. (Here we neglect indirect, global effects on transcription exerted by non-TF genes, such as metabolic enzymes, etc. that may, for instance, affect cellular pH which in turn could influence gene transcription.)

The entirety of transcriptional interactions can be conceptualized as a network, or directed graph, with genes representing the nodes and the regulation relationship between genes being the edges.[6,7] We are only at the beginning of determining the genome-wide architecture of that network for mammalian genomes using techniques such as ChIP-chip or ChIP-Seq.[8–10] Available data are both incomplete and crude, notably lacking the information on the 'promoter logics'[11] through which the inputs configuration of a node maps to its output (induction, repression of target). Nevertheless, a vast body of theoretical work exists aimed at the fundamental nature of the constraints imposed by the regulatory interactions in generic, complex networks. One goal of such studies, which use simulations of *in silico* networks whose general architecture is inspired by those observed in biology, is to relate the architecture of the regulatory networks to the dynamics and stability of observable transcriptomes.[7]

Considering the regulatory network as a complex dynamical system, a *state* of the network is defined by the configuration of the expression status of all the genes in the network and practically corresponds to a measurable gene expression profile which in turn maps into a phenotypic state of a cell or tissue. A central concept that arose from the theoretical treatment of complex networks is that because of the interaction constraints (given some basic conditions in the wiring diagram of the gene regulatory network[7]), the network produces stable 'attractor' states – spontaneously orchestrated collective behavior in the potential cacophony of expression of thousands of genes. A basic notion is that such naturally self-stabilizing, distinct network states correspond to the gene expression profile of the discrete cell types[7,12]. Recent experimental evidence support the notion that the gene expression profile associated with stable phenotypic states of cells are attractors of a genome-scale network.[5] Because of the complex web of constraints that prohibit most random gene expression configurations, the occupation of the theoretical gene expression space is not a continuum, but is sparse, limited to small regions, corresponding to attractor states and connecting trajectories.[5]

The studies of structure and dynamics of gene regulatory networks also led to the concept that the genomic transcriptional network contains a *core* set of genes consisting of TFs that regulate each other, hence forming what in graph theory is called a 'connected graph'. This core enslaves the rest of the network, the non-core or '*periphery*' genes consisting of either non-regulating but regulated *effector* genes, such as metabolic enzymes, etc. (and of those peripheral TFs that only control other effector genes but whose regulatory activity does not feed back to the core TFs) (Figure 1). Thus, the overall structure of the genomic regulatory network has been described as a 'medusa network'[13] – with the 'head', representing the 'core' TFs, that computes and executes the network constraints and thereby controls the 'medusa arms' representing the peripheral effector genes.[13] Thus the expression pattern of the core genes is expected to contain almost all information necessary to establish the genome-wide gene expression profile. A practically relevant hypothesis then is that in the analysis of genome-wide gene expression profiles associated with profound phenotypic differences, such as cell types or tumor types, the expression status of genes of the core alone may be most discriminatory in defining the expression profiles characteristic of a cell or tissue phenotype.

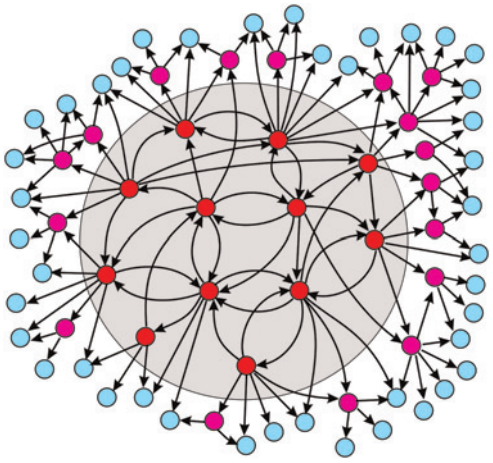

Figure 1 Schematic representation of a 'medusa' network architecture for the gene regulatory network. Circles represent individual genes (red = regulators = genes with transactivational (regulatory) activity; blue = peripheral 'effector' genes). Arrows denote regulatory activity. The grayed central area outlines the core which is technically a strongly connected graph ('medusa head'). Genes outside the gray circle represent the peripheral, non-core genes ('medusa arms'). Note that some transcription factors (TFs) (purple) are not part of the core since their regulatory action does not feed back to other TFs in the core (A color version of this figure is available in the online journal)

Clearly, a core gene must be a TF in order to send out network connections but not all TFs are necessarily part of the core. The genome of higher multicellular organisms consists of approximately 5–10% TFs.[14] The fraction of TFs scales in a characteristic manner with the complexity of the organisms and the genome,[14] in line with the fact that complex, multicellular organisms have to generate a higher diversity of discretely distinct, stable gene expression profile (attractors) corresponding to cell types. Note that not all TFs will be members of the core since, as mentioned above, some TFs may control only effector genes – thus not contributing to global network dynamics (='peripheral TFs', purple in Figure 1). The distinction between 'core TFs' and 'peripheral TFs' can at the moment not be made due to the incompleteness of transcriptional network data. Nevertheless, it is for practical purpose reasonable to assume that the small subset of genes in a genome that are nominally annotated as TFs approximates the regulating core of the genome-wide network that determines the transcriptome.

Specifically, since the genomic network architecture is still sketchy, here, we evaluate the functional consequence of the medusa network by asking whether the expression profile consisting only of the subset of genes that are considered TFs have a better discriminatory power than the much larger, genome-wide set of genes in classifying samples of lung tumors into the tumor types. Thus, we are testing a biological hypothesis and are not seeking to identify the

strongest discriminator genes in the tradition of data mining endeavors.[15] We use a published data-set of transcriptomes of samples of various types of lung cancer (normal lung, small cell carcinoma, squamous cell carcinoma and carcinoids)[16] and show that despite the small number of genes used as attributes to cluster the expression profiles of tumor specimen, the set of 538 TFs performed better than when all the 9072 genes or the same number of metabolic genes are used. Thus, the genome-wide transcriptome to some extent is a manifestation of the expression patterns of the TFs.

Methods

The data-sets and data preprocessing

Gene expression profile data from normal lung and pulmonary tumors previously published by Bhattacharjee *et al.*[16] (http://www.broad.mit.edu/mpr/lung) were used for the analysis shown in Figures 2–4. The data were obtained as Affymetrix array raw image (DAT) files and globally normalized by scaling to a target intensity of 1500 using the microarray suite (MAS) 5.0 program (Affymetrix). A total of $m = 64$ samples were used in this analysis, comprising four different tissues: squamous cell lung carcinoma ($n = 21$), pulmonary carcinoid ($n = 20$), small-cell lung carcinoma ($n = 6$) and normal lung ($n = 17$). Expression data were filtered by removing the genes that are not considered expressed above background noise in all samples (as defined by the 'absent' call based on internal statistics in MAS 5.0, using standard P values of 0.05), resulting in a input data matrix for the analyses of ($n = 9072$ genes) × ($m = 64$ samples). The data matrix was $\log_2$-transformed to obtain a normal distribution of the originally log-normally distributed 'signal' values to prevent bias by outlier genes in the clustering. Each sample was standardized to the z-score to further minimize global sample-to-sample variability due to external factors. The resulting value x_i^j for gene i in sample j was used for further calculations.

Extraction of TF and metabolic genes

Using the gene ontology functional annotation online tools DAVID[17] (http://david.abcc.ncifcrf.gov/), the list of human TF genes and metabolic genes were extracted using the gene ontology terms 'GO:0003700: transcription factor activity' and 'GO: 0008152: metabolic process', respectively.

Comparison of differentially expressed genes and MicroRNA data-set

For the mRNA and microRNA expression profile in Figure 5, data from various types of human cancer and normal tissues from the previous work by Lu and co-workers[18] (www.broad.mit.edu/cancer/pub/miGCM) were used. The expression data were obtained as GCT (gene cluster text file)-format normalized data matrix. The tissues with at least four samples are used in this analysis. Here the starting input data matrix was ($N = 14507$) × ($M = 75$).

Davies–Bouldin index

We used the Davies–Bouldin index (DBI) to evaluate the classification of the different gene sets. DBI is a metric used for measuring the quality of clustering results.[19,20]

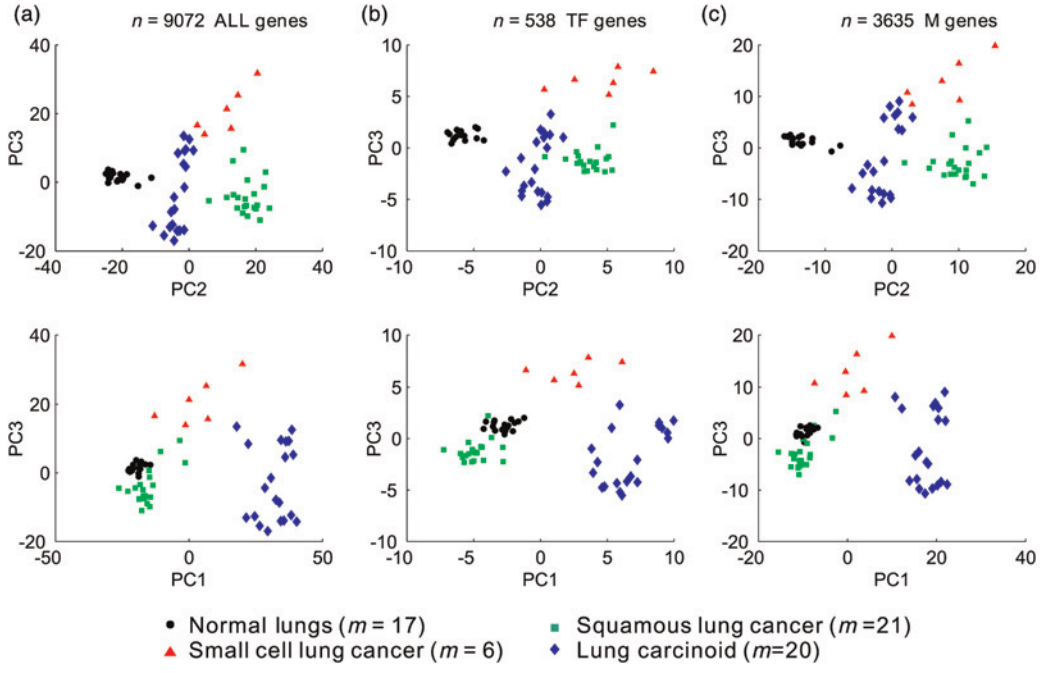

Figure 2 Natural distinction of pulmonal tumor samples using principal component analysis for three different sets of attributes (genes). Each sample is mapped as a point (symbol) onto the two-dimensional planes spanned by two of the first three principal components (PC1, PC2 and PC3) for each of the three gene sets (a, *ALL* set; b, *TF* set, c, *M* set). A total of 64 samples were included in the analysis and represented by the symbol according to the diagnosis shown in the legend (A color version of this figure is available in the online journal)

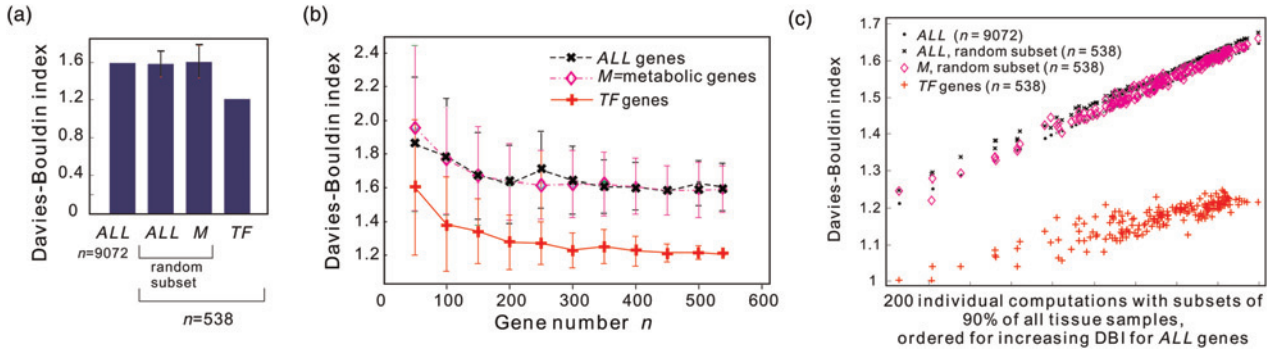

Figure 3 Davies–Bouldin index (DBI) showing that transcription factor genes can distinguish tissues better than other genes. (a) DBI for clustering the lung tumor samples computed for genes from the *TF* set is significantly lower than from the *ALL* set ($n = 9072$) and from randomly selected gene subsets ($n = 538$) from the *ALL* set and the metabolic *M* set. For the random subsets the mean of 100 random sampling of 538 genes is shown; error bars indicate the standard deviation from 100 samplings. Lower values of DBI indicate a stronger clustering performance. (b) Comparison of DBI for subsets of genes constructed by random selection of a decreasing number of genes ($n = 538$ down to 50, horizontal axis) from the set *ALL* genes, *TF* genes and *M* genes. For each gene number *n* the random sampling was repeated 100 times. Error bars indicate the standard deviation of 100 samplings. (c) Robustness to sample variability of the clustering capacity of transcription factors (TFs). Two hundred different configurations of randomly selected samples that comprise 90% of the original 64 set of samples were used to compute the DBI with the four gene sets. Each computation is displayed side-by-side as a column along the horizontal axis, ordered for increasing DBI value of the case when *ALL* genes were used. Thus, each column represents a test with a unique sample configuration, consisting of four data points, corresponding to the DBI of the data-sets with all genes, all 538 *TF* genes, 538 random genes (100 repeat sampling) and 538 random metabolic genes (100 repeat sampling), respectively. Error bars from repeated gene sampling are omitted for clarity (A color version of this figure is available in the online journal)

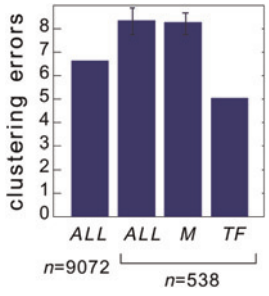

Figure 4 Contribution of *TF* gene set to k-means clustering quality. *k*-means clustering of the same subsets of genes shows that the *TF* subset had the lowest number of errors in the classification. Error bars indicate standard deviation from bootstrapping where more genes than the used subset were available (see Methods) (A color version of this figure is available in the online journal)

Here we use DBI to compare the classification quality of the data-set. The *a priori* known diagnoses are used to separate the data into the nominal tissue *groups* (e.g. normal lung versus small cell lung cancer, etc.), hence allowing the calculation of the DBI based on the within-group distances of samples in the same group and the between-group distances of the different groups. The within-group distances reflect the 'compactness' of the groups, and the between-groups distances correspond to the separation of the groups. A data-set with a lower DBI has better-separated and more compact groups, and therefore suggest a better classification.

Analysis by hierarchical clustering, *k*-means clustering and principal component analysis

For the principal component analysis (PCA) (Figure 2), hierarchical clustering, k-means clustering (Figure 4), we used the software Matlab (Mathworks, Natick, MA, USA). Clustering was performed in the 'sample dimension,' i.e. classifying the M tissue samples with respect to their N attributes, i.e. expression values of N genes, using Euclidean distance as a (dis)similarity measure between globally normalized samples and the 'average linkage' method to build the dendrogram.

For the error counting in the *k*-means clustering, the following algorithm was used:

(1) For each data-set, the *k*-means clustering method is applied, with $k = 4$ as the number of clusters desired (=the number of distinct tissue classes [normal and tumor types in the data]);
(2) The clustering result is compared with the *a priori* known four tissue classes (gold standard). The majority of samples of the same class in each cluster define the class label of a cluster and the number of samples grouped into wrong clusters is counted;
(3) A DiffMatrix and a MatchMatrix are calculated as follows: for each cluster i of the k-means clustering output and each class j of the *a priori* classes, the number of different samples is used as element (i, j) of the DiffMatrix, and the number of matched samples is used as element (i, j) of the MatchMatrix;
(4) Calculate matching pairs: find which cluster of the k-means clustering output corresponds to which cluster of the pre-known clusters by using the DiffMatrix and MatchMatrix;
(5) For each corresponding pair, calculate number of mismatched samples as error count;
(6) Repeat the above steps for 10,000 times; the average error count will be calculated.

Significance analysis of microarrays analysis

Significance analysis of microarrays (SAMs) (Figure 5) is a widely used technique that rank orders significant genes

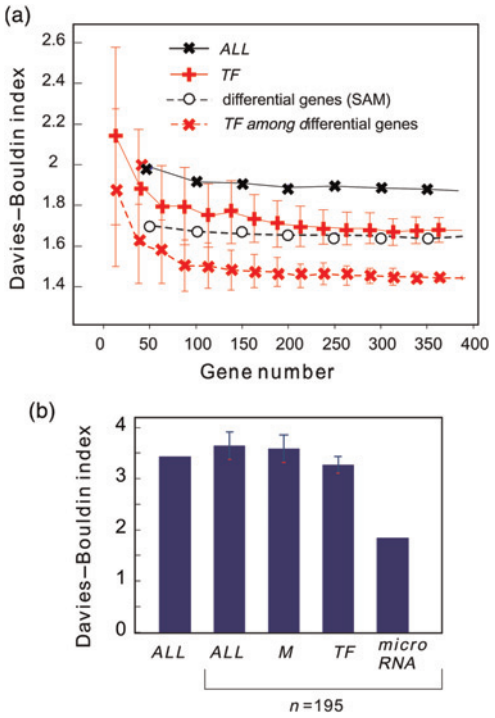

Figure 5 Differentially expressed genes and microRNAs have better discriminatory capacity than the transcription factors (TFs). (a) Davies–Bouldin index (DBI) analysis analogous to Figure 3b, but comparing the set 195 differentially expressed genes with an equal number of TFs from a distinct set of expression profile data (with *ALL* containing ~14,000 genes after filtering) of a variety of tumors.[21] A set of differentially expressed genes was determined based on significance analysis of microarray analysis (see Methods). Note the improvement of clustering (lower DBI) for the significance analysis of microarray (SAM) genes compared with TF genes only at low gene numbers (horizontal axis). (b) Similar to the lung tumor data (Figure 3), using the data-set of various types of human cancer and normal tissues[18] that contains $n = 195$ microRNAs, the DBI was calculated for clustering of pulmonal tumors based on these microRNAs and the indicated subset of n genes of equal set size. The DBI of microRNA expression profiles ($n = 195$) was approximately. half that of the other gene sets of equal number of genes. The DBI of randomly selected TF genes ($n = 195$) was still significantly lower than that of randomly selected genes ($n = 195$) or metabolic genes ($n = 195$). For the random $n = 195$ subsets of genes, the mean of 100 random sampling is shown; error bars indicate the standard deviation (A color version of this figure is available in the online journal)

based on a gene-specific *t*-test that uses permutations to compute significance.[21] SAM was performed with a delta of 0.06 on lung data containing four separate tissue types, each type containing five chips. Software that performed SAM was acquired from: http://www-stat.stanford.edu/~tibs/SAM/.

Results

TF expression profiles as attributes for distinguishing tumor samples

To compare the discriminatory power of TF expression patterns with that of other classes of genes or all genes in classifying tumor samples, we first defined a subset of genes as 'TF' genes among the genes measured by the microarrays and a subset of metabolic genes as control, representing non-regulating, 'effector gene' controls. These subsets were extracted as defined by the gene ontology (GO) categories for TFs and metabolic genes as described in the Methods section.[17,22] This extraction resulted in three sets of genes, named as follows: (1) the set *ALL* comprises the entire data-set with all the genes on the microarrays (after filtering of non-expressed genes, $n = 9072$ for Figures 2–4) reported by Bhattacharjee *et al.*;[16] (2) a set *M*, the subset of all metabolic genes ($n = 3635$); and (3) the set *TF*, the subset of TF genes ($n = 538$). Overall, the *M* set of metabolic genes has an average log-transformed expression value more similar to that of the *ALL* data-set with all genes; while the genes of the *TF* set had a lower average expression (Table 1), consistent with the common finding that TF genes are expressed at a low level.

We first compared how these three subsets of genes differed from each other when serving as attributes for comparing all the $m = 64$ tissue samples to each other. Thus we calculated the ($m \times m$) correlation matrix for all pairwise comparisons between the m lung tissue samples with respect to $n = 538$ genes which, to assure equal number of genes, were either randomly selected genes from the *ALL* gene set, randomly selected from the *M* set or comprised all the 538 genes in the *TF* set. Then the three ($m \times m$) matrices of correlation coefficients were compared with that obtained when all genes were used and a score was calculated as the correlation coefficient between the values in the respective ($m \times m$) matrix obtained from the subsets of genes and those from the *ALL* gene set. The correlation score of the 538 randomly selected genes against the score from the original set of *ALL* genes can be considered as a baseline for demonstrating the deviation incurred by the use of a smaller subset of genes. The result showed that a randomly selected subset of 538 metabolic genes had a correlation score similar to that of a same size set of genes randomly selected from the entire ('*ALL*') set genes. In contrast, for the subset of genes consisting of TFs, the correlation score was significantly lower (last row in Table 1). This indicates that the *TF* subset had different properties in relating the 64 samples to each other that was not due to the smaller number of the sample attributes (genes) alone.

To visually compare these three different data-sets with regard to clustering the four tissue/tumor types, we used PCA to reduce the dimension of the attribute space (expression values of the genes). Figure 2 shows the overall relationship among the 64 samples when projected to the plane spanned by the first three principal component (PC) axes. The overall distribution of the samples were very

Table 1 Three sets of genes to be compared: 'ALL', 'M' = metabolic and 'TF' = transcription factors

Set Genes	ALL All genes	M Metabolic genes	TF Transcription factors
Total number of genes in set	9072	3635	538
Average expression value	6.8716	6.9283	6.7227
Correlation score of 538 gene-subset with entire set	0.9720	0.9514	0.9025

In the last row, correlation scores of between the three 538 gene-subsets of genes and the whole set is shown. 'Correlation' refers to correlation of the matrix of the ($m \times m$) correlation coefficients between all m pairs of tissues calculated in each case

similar between the *ALL* set and *M* set, whereas visual inspection reveals that the *TF* set appears different. This finding is consistent with the correlation scores above (Table 1). Clearly, the drastic reduction of the attributes from more than 9000 gene expression values to less than 600 does not appear to affect the ability of the profiles to cluster these lung tumor samples according to their subtypes. Interestingly, when the *TF* gene set was used, the six small cell lung cancer samples form a distinct group that is even more clearly separable from the other groups with respect to both PC1 and PC3, whereas this separation was less obvious when *ALL* or *M* genes were used. However, although using TFs as features seems to provide a better separation of different tissue classes in this simple PCA analysis, this improvement is at this point rather subjective.

TF genes have better capacity to distinguish tumor tissues than other genes

In order to objectively and quantitatively assess how well the subsets of genes discriminate the sample groups, we determined to what extent the gene expression profiles of the different subsets of genes can distinguish the (*a priori* known) types of tumors by comparing an index of clustering quality, the DBI.[19] DBI is a quantity that assesses the compactness of clusters and the separateness between clusters; essentially, it takes the ratio of all intracluster distances over the intercluster distances in the *n*-dimensional gene expression space where each point represents a sample profile (see Methods for details). Thus, a lower DBI in a data-set indicates a better clustering quality. Traditionally, DBI is used to find the optimal classification scheme for a given data-set where no *a priori*, nominal groups exist. Here we face the reversed problem since we already know the 'optimal' classification, namely, the nominal diagnosis of the tumor types provided by histology, pathology and clinical features. Thus, we have the 'gold standard' for the classification of the samples. Our question is which set of attributes (genes) would produce the lowest DBI given the known clusters. Profiles are assigned to the clusters based on *a priori* known diagnosis, not based on a cluster algorithm, and then, given this classification, the DBI is computed using the expression values of the three gene subsets as classification attributes.

Again, to neutralize the confounding effect of number of genes (attributes) in the various gene sets, with the *TF* set having least genes ($n = 538$ gene), we used random subsets of $n = 538$ genes from the set of *ALL* genes and random subsets of the same number of genes from the set *M* of metabolic genes. Subsets were drawn randomly 100 times (bootstrapping) and DBI computed and averaged. We found that the TF subset of $n = 538$ genes had a significantly lower DBI in forming the four tissue clusters than the gene sets with all the genes ($n = 9072$), or the random selection of an equal size subset form all genes ($n = 538$), or the metabolic genes (Figure 3a). The difference in DBI between the *TF* gene set and the random selection of a same number of genes are statistically significant based on the bootstrap analysis.[23]

Although the average expression values across all the tissues for TF genes were lower than the average of all the genes (Table 1) and hence TF gene expression is more susceptible to experimental noise, the TF genes still outperformed the other sets of genes in terms of better separating the four different types of tissues than the other gene sets.

To further challenge the discriminatory capacity of the gene sets, we next reduced the number of genes from each set and calculated the DBI with respect to the four known clusters. Subsets of decreasing number of n' genes were randomly selected from each of the three gene sets and the DBI was calculated for each subset of size n'. Each test (random selection of subsets of size n' for each of the gene sets) was repeated 100 times and the resulting DBI results were averaged. This analysis showed that the DBI for each of the three set of the genes remained essentially unchanged as the gene number was reduced from 538 to around 200 when the DBI began to increase, i.e., the clustering worsened (Figure 3b). Thus, about 200–250 randomly selected genes appear to be diverse enough to capture the distinction in gene expression patterns between the different types of lung tumors and normal lung tissue. In the case of the *TF* gene set, the DBI remained lower than that of the other control gene sets by 0.3–0.4 units of DBI for the same number of genes. Remarkably, even with less than 10% of all the *TF*s in this data-set, the DBI of 1.6 was still lower than when the entire set *ALL* gene set was used.

The discriminatory capacity of TF genes is robust to sample selection

In current statistics or machine learning, classification is often performed using a set of most discriminatory genes (e.g. a diagnostic signature) that was first determined by an independent algorithm. Such diagnostic gene sets for tumors are often affected by the composition of the set of the samples used, and often do not overlap between different studies.[24] This dependency on the sample configuration usually suggests over-fitting. In contrast to such analysis here we used *a priori* biological function (*TF* versus *metabolic* genes) rather than statistical analysis as criteria for the subset of attributes – independent of their individual discriminatory capacity. Thus we expect that the result will be less dependent on the sample configuration. To test whether different sample configurations of the data-set will affect the results, we randomly selected 90% of the $m = 64$ samples to calculate the DBI and repeated this 200 times. As shown in Figure 3c, the 200 sets of DBI values among the gene sets used varied significantly across the different sample configurations that contained as much as 90% of the original sample set. Nevertheless, for each configuration (ordered in Figure 3c along the *x*-axis according to the DBI obtained for *ALL* genes), the difference in DBI between *TF* subset and the other gene sets was clearly maintained. All DBI values for the *TF* gene set form a band at the bottom of the graph. DBI values of the other three data-sets form another band on the top. These two groups are well separated by a DBI difference of 0.2–0.5 for all the repeated runs (Figure 3c). Thus, the TF genes are better at clustering the

lung tumor samples independent of the particular sample configuration.

TF genes cluster lung samples more accurately than all genes

Hierarchical clustering is a simple, practical, standard method for analysis of sets of gene expression profiles that is very robust to reduction of the number of features (genes).[25] In fact, when we subjected the data-sets to hierarchical clustering using our four gene sets, *ALL*, a subset of *ALL*, *TF*, and *M* (the latter three sets with $n = 538$ genes), the tumor types clustered nearly as well as the entire set of 9072 genes (ALL) in that the randomized subset of 538 genes from the *ALL* gene set produced only one 'miss-classification' among the 64 samples (data not shown). This reflects the appreciated inherently robust separation of root classes by hierarchical clustering of high-dimensional gene expression data when naturally distinct groups are present. Since dendrograms are ambiguous in defining clusters ('cutting of trees' at subjective level of branching)[26] and difficult to validate at the level of statistical ensembles, we used non-hierarchical *k*-means clustering to quantify the quality of clustering into the four *a priori* groups. Subjecting the same data-sets to *k*-means clustering ($k = 4$) allowed us to systematically count the number of misclassified tissue samples in a bootstrap analysis (see Methods). The results, summarized in Figure 4, now exposed the superior quality of clusters based on TFs (~5 errors) as compared with using the same number of randomly selecting genes from the entire set or from metabolic genes (~8 errors). In the latter case, reduction of gene number to ($n = 538$) significantly increased classification errors. By contrast the *TF* subset ($n = 538$ genes) even performed better than when using *ALL*, $n = 9072$ genes (~6.7 errors), consistent with the results based on DBI (Figure 2).

Gene sets that cluster better than TFs

Selecting subsets of genes based on their GO classification as attributes for clustering tissue samples is here motivated by the biological rationale that a core set of regulatory genes, well approximated by the set of TF genes, controls the genome-wide expression profile. Although the *TF* set we used in fact performed better than random subsets, it is not clear how TFs fare when compared with a set of genes selected on purely statistical basis for their discriminatory capacity. A number of statistical methods have been developed for bidirectionally unsupervised identification of sets of genes that classify samples into subclasses (e.g. see references[15,27]). However, here the subclasses are known *a priori* (cancer subtypes). Thus, we used the simplest criterion, differential expression in the various diagnosis groups, to obtain a set of cancer type specific genes for comparison with the TFs. Using the standard SAM method,[21] we selected a subset of most significantly differentially expressed in another set of tumor expression data[21] (see Methods), defining a set of 195 '*SAM*' genes that were strongly differentially expressed between tumor subtypes. We performed DBI analysis analogous to Figure 3, using either the set of $n = 195$ SAM genes or the same number randomly selected *TF* genes and successively decreased the number of genes used for calculating the DBI. As Figure 5a shows, the TF genes performed almost as well as the SAM genes and both groups also clustered the samples much better than when *ALL* genes were used, i.e. had much lower DBI. It was only at lower number of genes that the clustering performance of SAM genes became significantly better than that of *TFs*.

If TFs exhibit good clustering performance because they control the gene expression profile, we need to consider whether microRNAs also perform as well. MicroRNAs have been shown to cluster pulmonal tumors, consistent with the idea that microRNAs play a pivotal role in development and cancer.[18] Here we posit that their role in core regulatory networks may also afford them particular discriminatory capacity. MicroRNAs directly control the expression of genes, in part through destabilization of transcripts or translational inhibition.[28] But unlike TFs, their action circumvents the complex promoter logics, imposing the repression of expression independent of other regulatory inputs at the promoter. Such overriding regulatory capacity has been referred to as 'canalizing function'[29] based on the theory of network dynamics,[7] and plays a pivotal role in the generation of 'ordered dynamics' and stable attractor states that correspond to distinct gene network states, such as those representing cell types.[7] Figure 5b shows a DBI analysis analogous to Figure 3, but using a different data-set that has microRNAs[18] (see Methods), comparing the DBI in clustering tumors for a set of 195 microRNA genes and for randomly selected sets of 195 metabolic (*M*), *TF* and any other genes. Clearly, the microRNAs outperformed even the TF genes. With only 195 *TF* genes used, this set of genes still showed clear, albeit more moderate, improvement compared with the *ALL* genes set or a set of *M* genes of the same size (Figure 5b). Overall, these findings may reflect the fact that microRNAs more tightly control the gene expression patterns through their canalizing functionality, and thus are in line with the notion that genes that are strong regulators constitute good attributes for clustering.

Discussion

Since the gene expression profiles that in the first place permit the classification of normal tissue or tumor samples are produced by the dynamical constraints of a molecular network of gene interactions, the discriminatory capacity of a expression profile as classifier is not solely determined by the number of genes (the features or attributes for classification) used but also by properties of the genes in terms of their role in establishing the stable gene expression patterns. In the history of microarray data analysis, numerous statistical approaches have been proposed to identify sets of genes (gene signatures) that have maximal discriminatory power to differentiate a set of samples (gene expression patterns) either in a supervised or in a unsupervised setting.[15,27] However, such statistical data mining analyses remain in general agnostic to what

molecular and biological mechanism in the first place created these non-random expression patterns to begin with.

Here we approach gene expression analysis from a fundamentally different direction: not from the data mining framework but with a biological rationale regarding the genesis of the expression pattern. We view tissue and condition specific, recurrent gene expression patterns as the product of the integrated dynamics of the gene regulatory network. Specifically, a recurrent gene expression pattern reflects a recurrent and hence likely stable state of the network, or attractor state, a 'locked-in' state due to the nonlinear regulatory interactions between a set of genes capable of regulating each other's expression activity.[5,7] In principle, this set of core regulatory genes enslaves the entire genome and determines the transcriptome that consists of regulators and effectors in our simplifying dichotomy. With this gene network based rationale we postulated that genes that are members of the 'core' of the genomic regulatory network, as approximated by taking all the genes that are considered having transcriptional activity (*TF* set), contain at least the same amount of information as the genome-wide gene expression profile for clustering lung tumor samples into known diagnostic groups.

In other words, the non-core, peripheral portion of the transcriptome, here epitomized by the metabolic or the random genes, did not add significant information to discriminate tissue samples. This is consistent with the notion of a 'medusa' architecture of the transcriptional regulatory network with a core and a periphery that is enslaved by the core and may fluctuate due to environmental conditions.

However, although effective in grouping the tumor samples into data-sets, the *TF* gene set is not the optimal gene set for clustering. Genes selected based on an algorithm for their differential expression perform slightly better, notably, at low gene number. Not surprisingly, microRNAs, which represent canalizing functions in regulatory networks and hence permit much less 'leeway' in regulation, substantially outperformed TFs in clustering tumor profiles. These findings have practical relevance: TFs and microRNAs may represent a universal set of attributes with high discriminatory power, which, unlike the 'diagnostic gene signatures' need not be determined statistically *ad hoc* for each new data-set. Future analysis thus should evaluate whether good diagnostic gene sets obtained from unsupervised statistical analysis (e.g. reference[15]) are enriched in the class of genes that control transcriptome dynamics.

Yet, despite the medusa network concept[13] with a core set of genes that enslaves peripheral genes and our findings that are consistent with this notion, it is reasonable to assume that the mRNA profile of the core genes alone will not fully determine the entire transcriptome. Aside from the trivial notion that microarrays measure only mRNA steady-state levels, which due to post-transcriptional regulation does not map 1:1 into transcriptional activity of the respective gene product,[30] we also note that the peripheral genes in the 'medusa arms'[13] enjoy some degrees of freedom beyond gene regulatory control – this is the reason why they perform poorer in classification. Clearly these genes respond in a specific manner to external signals which may differ between the samples. This kind of modulation of gene expression alters the expression profile with respect to the periphery without causing any change to the core. Such changes may transiently affect expression of effector genes in a coherent way, such as cytokine gene expression in response to infectious agents, and are less likely to cause wide-spread transcriptome changes that may switch attractor states. Yet they still may be coordinated in a manner that they produce expression profiles that encode for particular transient cellular state. Such subphenotypes would be missed by a clustering based on TFs or microRNAs alone.

Despite limited sample size of gene expression profiles, our results show that gene expression profile data-sets are not a 'black box' to be analyzed blindly using powerful multivariate statistical tools, but that there is a biologically defined substructure that is grounded in gene network architecture through regulatory dynamics. Analyzing the clustering capacity of the TF and microRNA subset, or more precisely defined regulators in an expanded set of gene expression profile data-sets will shed more light on the biological consequences of the medusa structure of genomic networks and reveal how much autonomy the peripheral genes possess.

Author contributions: SH conceived of the question and analysis; YG and SH developed the approach. YG, YF and NST conducted the data collection and computational analysis. All authors participated in interpretation of the data and in reviewing the manuscript. SH and YG wrote the manuscript.

ACKNOWLEDGEMENTS

The authors acknowledge the Air Force Office of Scientific Research (AFOSR) for funding (SH) and thank Dr Donald Ingber for support of YG and YF and Dr Stuart Kauffman for discussions.

REFERENCES

1 Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;**286**:531–7

2 Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;**403**:503–11

3 van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002;**347**:1999–2009

4 Huang S. The practical problems of post-genomic biology. *Nat Biotechnol* 2000;**18**:471–2

5 Huang S, Eichler G, Bar-Yam Y, Ingber DE. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys Rev Lett* 2005;**94**:128701

6 Huang S. Reprogramming cell fates: reconciling rarity with robustness. *Bioessays* 2009;**31**:546–60
7 Huang S, Kauffman S. Complex gene regulatory networks – from structure to biological observables: cell fate determination. In: Meyers RA, ed. *Encyclopedia of Complexity and Systems Science*. Berlin: Springer, 2009
8 Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. *Science* 2000;**290**:2306–9
9 Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;**4**:651–7
10 Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of *in vivo* protein-DNA interactions. *Science* 2007;**316**:1497–502
11 Yuh CH, Bolouri H, Davidson EH. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* 1998;**279**:1896–902
12 Kauffman S. Homeostasis and differentiation in random genetic control networks. *Nature* 1969;**224**:177–8
13 Kauffman S. A proposal for using the ensemble approach to understand genetic regulatory networks. *J Theor Biol* 2004;**230**:581–90
14 van Nimwegen E. Scaling laws in the functional content of genomes. *Trends Genet* 2003;**19**:479–84
15 Roden JC, King BW, Trout D, Mortazavi A, Wold BJ, Hart CE. Mining gene expression data by interpreting principal components. *BMC Bioinformatics* 2006;**7**:194
16 Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* 2001;**98**:13790–5
17 Dennis G Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol* 2003;**4**:P3
18 Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR. MicroRNA expression profiles classify human cancers. *Nature* 2005;**435**:834–8
19 Davis DL, Bouldin DW. A clustering separation measure. *IEEE Trans Patt Anal Mach Intell* 1979;**1**:224–7
20 Chen G, Dai Y. A new distance measurement for clustering time-course gene expression data. *Conf Proc IEEE Eng Med Biol Soc* 2004;**4**:2929–32
21 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001;**98**:5116–21
22 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9
23 Efron B, Halloran E, Holmes S. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci USA* 1996;**93**:13429–34
24 Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N Engl J Med* 2009;**360**:790–800
25 Barnés CM, Huang S, Kaipainen A, Sanoudou D, Chen EJ, Eichler GS, Guo Y, Yu Y, Ingber DE, Mulliken JB, Beggs AH, Folkman J, Fishman SJ. Evidence by molecular profiling for a placental origin of infantile hemangioma. *Proc Natl Acad Sci USA* 2005;**102**:19097–102
26 Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008;**24**:719–20
27 Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc Natl Acad Sci USA* 2000;**97**:12079–84
28 Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**:281–97
29 Kauffman SA. *The Origins of Order*. New York: Oxford University Press, 1993
30 Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* 1999;**19**:1720–30

(Received October 23, 2010, Accepted February 28, 2011)