Published online 27 March 2007 *Nucleic Acids Research, 2007, Vol. 35, No. 7* ***2343–2355***
doi:10.1093/nar/gkm119

Graph-based identification of cancer signaling pathways from published gene expression signatures using PubLiME

Giacomo Finocchiaro[1,2], Francesco Mattia Mancuso[1,2], Davide Cittaro[1] and Heiko Muller[1,2,*]

[1]The FIRC Institute of Molecular Oncology Foundation, Via Adamello 16, 20139 Milan, Italy and [2]Department of Experimental Oncology, European Institute of Oncology, Via Ripamonti 435, 20141 Milan, Italy

Received December 6, 2006; Revised January 22, 2007; Accepted February 12, 2007

ABSTRACT

Gene expression technology has become a routine application in many laboratories and has provided large amounts of gene expression signatures that have been identified in a variety of cancer types. Interpretation of gene expression signatures would profit from the availability of a procedure capable of assigning differentially regulated genes or entire gene signatures to defined cancer signaling pathways. Here we describe a graph-based approach that identifies cancer signaling pathways from published gene expression signatures. Published gene expression signatures are collected in a database (PubLiME: Published Lists of Microarray Experiments) enabled for cross-platform gene annotation. Significant co-occurrence modules composed of up to 10 genes in different gene expression signatures are identified. Significantly co-occurring genes are linked by an edge in an undirected graph. Edge-betweenness and k-clique clustering combined with graph modularity as a quality measure are used to identify communities in the resulting graph. The identified communities consist of cell cycle, apoptosis, phosphorylation cascade, extra cellular matrix, interferon and immune response regulators as well as communities of unknown function. The genes constituting different communities are characterized by common genomic features and strongly enriched *cis*-regulatory modules in their upstream regulatory regions that are consistent with pathway assignment of those genes.

INTRODUCTION

Gene expression microarrays have been applied to studying a wide variety of biological conditions, including cancer. Analysis of microarray data is generally performed by applying a number of filtering steps, the application of statistical tests, cluster analysis, definition of classifier gene sets, annotation of genes identified in clusters and the validation of differential expression using alternative techniques such as quantitative PCR. With the accumulation of publicly accessible datasets stored in repositories like ArrayExpress (1), GEO (2), CIBEX (3) and Oncomine (4), it is becoming increasingly feasible to cross-compare in-house generated data to published datasets and to perform meta-analysis, which can be very informative.

However, since gene expression studies are being performed using a variety of commercially available as well as custom microarray platforms, meta-analysis is hampered by the need for cross-platform annotation. As a consequence, researchers are forced to compare their data to published data that have been generated on compatible microarray platforms or to undergo a painstaking cross-platform annotation effort which limits the number of datasets available for meta-analysis. Furthermore, the difficulties in rendering datasets compatible for meta-analysis often conditions the choice of published datasets to those which are most ‘interesting’ from the point of view of biological intuition, precluding the discovery of unexpected connections between diverse datasets.

We (5) and others (6) have proposed possible solutions to this problem by developing repositories that are helpful in performing meta-analysis tasks and in identifying similar datasets in an unbiased manner. Nevertheless, despite of the use of sophisticated gene annotation tools

*To whom correspondence should be addressed. Tel: +39 02 574303263; Fax: +39 02 574303244; Email: heiko.muller@ifom-ieo-campus.it
The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

such as DAVID (7), Onto-tools (8) and GoMiner (9), the interpretation of gene expression signatures remains essentially restricted to the interpretation of gene lists identified in isolated experiments and subject to personal judgment, as different hypotheses can have similar statistical validity. Furthermore, using approaches based on pre-assembled gene lists, discovery of unknown pathways is impossible.

Here we discuss a graph-based integrative approach that identifies biologically meaningful pathways from the analysis of co-occurrence patterns observed in published gene expression signatures. Genes whose expression is governed by similar signals are expected to co-occur significantly in distinct signatures and to form a strongly interconnected community with different communities being targeted by different signaling pathways. Using this approach, individual genes as well as entire signatures can be assigned to pathways whose composition is entirely determined by the signatures in the repository as well as the signature being analyzed rather than by pre-configured gene sets that have been grouped according to various measures of similarity.

The analysis consists in the following steps:

(1) Identification of significantly co-occurring gene modules composed of up to 10 genes in the PubLiME repository.
(2) Generation of a graph representation of co-occurring gene modules where genes are represented by nodes that are linked by an edge when the genes are part of the same significant co-occurrence module.
(3) Identification of strongly interconnected communities (i.e. pathways) in the resulting graph using two different approaches: edge-betweenness clustering combined with graph modularity as a quality measure (10), which identifies separate communities and k-clique clustering (11) identifying partially overlapping communities, both yielding highly consistent results.

Following this procedure, we identified communities that correspond to defined biological pathways composed of regulators of the cell cycle, phosphorylation cascades, apoptosis, extracellular matrix, immune and interferon responses, as well as communities of unknown function. We show that the genes that constitute different communities are characterized by common genomic features and display strongly enriched *cis*-regulatory modules in their putative promoter regions that are consistent with pathway assignment.

MATERIALS AND METHODS

Generation of a repository of published cancer gene signatures

We searched the Affymetrix database of publications using Affymetrix technology and PubMed to identify cancer-related gene expression microarray studies appearing in the years 1999–2005. Four hundred ninety nine published cancer-related gene expression microarray studies were scrutinized for scope of the study, microarray platforms employed, organism studied and feasibility of cross-platform annotation of published gene expression signatures. Two hundred seventy three studies (233 human and 40 mouse) were selected for manual extraction of gene expression signatures. The selected publications report two basic types of signatures: signatures resulting form unbiased screening of cancer specimens as well as studies identifying differentially regulated genes in cell-line-based model systems. Cross-platform annotation of gene expression signatures was performed according to a procedure described in (12). Medline annotations of these publications were downloaded by calling NCBI Entrez Utilities (http://utils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html from a Perl script via the LWP module. Parsing of downloaded XML files was performed by a Perl script using the DOM module. Data regarding publications and gene expression signatures were imported into a relational MySQL database that is accessible via a web-interface (http://bio.ifom-ieo-campus.it/publime/).

Significance of co-occurrence of genes in gene expression signatures

Estimates of significance of co-occurrence of genes in a gene expression signature are based on randomization of signatures. Gene expression signatures are composed of non-redundant sets of genes. Thus, randomized signatures must be composed of non-redundant gene sets as well, leading to constraints on the composition of randomized signatures which precludes calculating the probability of a gene for being part of a given signature based on the number of signatures that a gene is part of and the number and sizes of analyzed signatures. Therefore, a gene-swap procedure is used where prior to swapping two genes between signatures, a test is performed ensuring that the genes to be swapped between signatures are not already present in the respective target signatures. To ensure complete randomization of signatures, the number of swaps performed in a single simulation is chosen to be 10 times the sum of all signature sizes. Ten thousand simulations are run and at each simulation, the presence/absence of each gene in each signature is determined. The occurrence probability of a gene in a given signature is then calculated as the number of times the gene was found being part of that signature divided by 10 000.

Given the occurrence probabilities of genes per signature, co-occurrence probabilities are calculated by multiplying the occurrence probabilities of the genes under study (this set of genes is called a module) in each signature. Co-occurrences of two up to ten genes were analyzed. Co-occurrence probabilities are signature-specific.

Given the co-occurrence probabilities of a module, the significance of the number of observed co-occurrences must be evaluated. If co-occurrence probabilities were equal for all signatures (which would be the case if all signatures were of equal size), the probability distribution of co-occurrences of a module would be given by the Binomial Distribution in which the number of trials is equal to the number of signatures, the number of successes is the number of co-occurrences and the probability of success is the co-occurrence probability of a module. However, the signature-specific nature of co-occurrence probabilities caused by differences in signature sizes implies that the co-occurrence probability distribution of a module is given by a Binomial Distribution with trial-specific probabilities. Calculating this distribution for large numbers of signatures is not feasible because it implies summation of a number of terms that is given by the binomial coefficient $\binom{S}{k}$ where k is the number of co-occurrences and S is the number of signatures which assumes large values even for relatively small numbers of signatures. Therefore, we apply a Z-score transformation to the number of co-occurrences k of a module given by

$$\frac{k-\mu}{\sigma}.$$

Here, μ and σ designate the mean and standard deviation of the Binomial Distribution with trial-specific probabilities which are calculated as:

$$\mu = \sum_{i=1}^{S} p_i \quad \text{and} \quad \sigma = \sqrt[2]{\sum_{i=1}^{S} p_i - \sum_{i=1}^{S} p_i^2}$$

where p_i designates the signature-specific co-occurrence probability and S is the number of signatures. Please note that for all p_i being equal $\mu = S^*p$ and $\sigma = \sqrt[2]{S^*p - S^*p^2}$ which are the mean and standard deviation of the Binomial Distribution, respectively.

Edge-betweenness clustering and graph modularity

Edge-betweenness is defined as the fraction of shortest paths in a graph that pass through an edge. Edges that are connecting two strongly interconnected communities (hence are in between those communities) will be part of shortest paths more often than edges within the communities when shortest paths are calculated for all nodes of the two communities. Having identified the edge with the largest edge-betweenness, the edge is removed from the graph. Then, the algorithm restarts calculating shortest paths in the remaining graph, identifies the next edge with highest edge-betweenness, removes it from the graph and so on. The procedure is supplemented with a quality measure (graph modularity) which tells the algorithm the optimal number of edges to be removed from the graph. Graph modularity is calculated as the difference of the fraction of observed edges within a community minus the expected fraction of edges within a community if the edges were linking the nodes of the graph at random. This difference is calculated for all communities and the sum thereof represents graph modularity whose value for highly modular graphs is found to be larger than 0.3 and rarely exceeds 0.7 (10).

Promoter analysis

The identification of co-occurrence modules of genes is performed using a list representation of signatures as PubmedID-gene pairs. The identification of *cis*-regulatory modules is performed using the same software and statistics with the only difference that lists of promoter-motif pairs are used. To obtain these lists (supplementary file 'Symbol_TF.txt'), we assembled a collection of consensus transcription regulatory motifs derived from Transfac database (13) (supplementary table 'consensus_motifs.xls'). Next, we identified all matches of consensus motifs within 500 bp upstream of the annotated transcription start site for all Refseq genes in the human genome (UCSC hg18), excluding duplicated promoter sequences and Refseqs with ambiguous genome mapping. Multiple occurrences of a motif in the same promoter were ignored. The swapping procedure described earlier was used to obtain genome-wide occurrence probabilities for each motif in each promoter region. Co-occurrence of combinations of motifs in the promoters of community forming genes is then analyzed and gives the number of promoters containing the module. Modules are required to be present in at least one-third of all promoters of community forming genes. The promoter-specific co-occurrence probability of a module is then calculated by multiplying the occurrence probabilities of the composing motifs. A Z-score using the Binomial Distribution with trial-specific probabilities is calculated as described earlier. A Z-score cutoff of 5 is chosen to define significant *cis*-regulatory modules. The same procedure is carried out in randomized versions of promoter-motif lists for each community. The number of significant *cis*-regulatory modules for each community and module size is divided by the corresponding number of modules identified in randomized promoter-motif lists so as to obtain a signal-to-noise ratio (SNR).

In order to visualize the modules identified at the module size giving the best signal-to-noise ratio, we generated a graph representation of motif modules where motifs are nodes linked by an edge if they are part of the same significant *cis*-regulatory module. In this representation, the motif that is most frequently co-occurring with other motifs will be characterized by a high node degree that can be visualized by varying node size. A similar representation is obtained using the PageRank algorithm (JUNG software package) that, in addition to node degree, also takes into consideration the structure of the graph in order to identify the node that will be visited most frequently upon random walks along the edges of a graph.

Software

Custom Java-based software was used for determining occurrence probabilities, co-occurrence probabilities and the identification of significant co-occurrence modules from PubLiME data. JUNG (http://jung.sourceforge.net/index.html) and Netsight (http://jung.sourceforge.net/netsight/) software were used for edge-betweenness clustering and graph visualization. For calculating graph modularity, a customized Java class implementing graph modularity calculation as described by (10) was written and run within the JUNG framework. CFinder software was used for k-clique clustering (11). Boxplots and Q–Q plots were prepared using R.

RESULTS

PubLiME content and meta-analysis

Gene expression microarray data generated from numerous studies are generally published as supplementary data and/or they are deposited in public gene expression data repositories like ArrayExpress (1) and GEO (2). The fact that there are a considerable number of different microarray platforms available turns meta-analysis of gene expression data into a nontrivial task because individual genes are often represented in a many-to-many relationship on different array platforms. While individual experiments can be scrutinized in great depth, meta-analysis of microarray data is concerned with cross-experiment and often cross-platform comparison of gene expression data (Figure 1A). Complications in this process arise from the use of different identifiers and formats in different publications as well as from differences in gene content and probes employed by different microarray platforms.

In order to facilitate meta-analysis, we chose to generate a repository of gene expression signatures that hosts the differentially regulated gene sets identified by individual studies as gene lists in a relational database (Figure 1B). The database is composed of three logical units regarding individual genes, gene lists (i.e. signatures) and publications. Currently, the database hosts 1041 human gene lists collected from 233 publications and 241 mouse gene lists from 40 publications. The types of identifiers used in different gene lists are shown in Figure 1C. Genbank accession number is the most widely used, followed by Affymetrix probeset, gene symbol, Unigene title and Unigene cluster identifier. All identifiers used in publications were mapped to Entrez Gene identifiers. 31243 reported identifiers of human genes were mapped to 7476 Gene identifiers and 8323 reported mouse identifiers were mapped to 4277 mouse Gene identifiers meaning that roughly one-third of all human and mouse genes have been reported as differentially regulated at least once.

The publications cover a wide range of different conditions. Figure 1D shows a classification of publications by experimental approach. There are roughly equal numbers of studies employing model systems and patient samples. Some studies use both approaches. The model systems comprise screening of cell lines with inducible gene expression, drug/hormone/cytokine/radiation/serum/knockdown treatment of cell lines, co-culture/infection of cell lines with pathogens, somatic gene knockout in cell lines and comparison of cell lines with different tumorigenic or drug-resistance properties. Patient samples and model systems are derived from cancers of a wide range of organs (Figure 1E). Publications studying cancer of blood cells are most abundant but very heterogeneous. There are patient sample studies on leukemia (T-ALL, B-ALL, AML, CLL and CML), lymphoma, thymoma and many model system studies. A complete account of conditions is found in the supplementary file 'conditions_overlaps.xls'.

We analyzed the occurrence frequency of genes in published signatures and found it to be strongly uneven, with few genes being present in more than 20 signatures while most genes are found in less than 5 signatures (Figure 2A). The 20 most frequently occurring genes are listed in Table 1. Possible explanations for this distribution are biased selection of conditions, higher expression levels of most represented genes leading to more reliable detection and use of alternative promoters in different conditions, causing changing expression in a larger number of conditions. Biased selection of conditions appears unlikely considering the data shown in Figure 1E, although it cannot be excluded entirely. To test the remaining two hypotheses, we analyzed the Unigene Hs.198 EST expression profiles. Expression levels are expressed in transcripts per million in 49 different tissues. We tested whether the 500 most frequently occurring genes have a significantly higher expression level in different tissues than the average of all genes with an Entrez GeneID. Wilcoxon signed rank sum tests were performed for all tissues. The negative decadic logarithm of the obtained *P*-values shown in Figure 2B illustrates that the 500 most frequently occurring genes are indeed expressed at higher levels in nearly all tissues. Similarly, we analyzed whether those genes are expressed in more tissues. A gene was considered expressed when at least one EST corresponding to that gene has been identified in that tissue. The last column of Figure 2B shows that the 500 most frequently occurring PubLiME genes are expressed in a larger number of different tissues with a highly significant *P*-value. We sought to identify common genomic features of the 500 most frequent genes. CpG islands covering transcription start sites are known to be associated with promoter activity and are characterized by frequent alternative start sites (14,15). Therefore, we analyzed CpG island lengths of CpG islands covering annotated transcription start sites for the 500 most cited PubLiME genes and compared it to the genomic average. We found that the 500 most frequently occurring genes have significantly longer CpG island promoters as well as more annotated alternative 5′-ends (Figure 2C), suggesting that these genes might be expressed from a number of alternative promoters responding to different stimuli. Apart from offering an explanation for the nonuniform distribution of gene occurrences in PubLiME, these results suggest that PubLiME hosts biologically relevant information.

PubLiME can be queried via a web-interface. Possible searches include single gene searches returning all publications where the gene has been reported as differentially regulated, gene list searches returning publications reporting similar gene sets where similarity is evaluated based on the hypergeometric distribution, as well as searches regarding publications where fields such as Authors, Abstract, Title, Mesh terms and PubMed identifiers can be interrogated (Supplementary Figures S1 and S2).

Identification of significant co-occurrence modules of genes

We interrogated PubLiME with the aim of identifying significantly co-occurring gene sets composed of up to 10 genes. We call a set of significantly co-occurring genes a co-occurrence module and the number of co-occurring genes the module size (Figure 3A). This nomenclature was

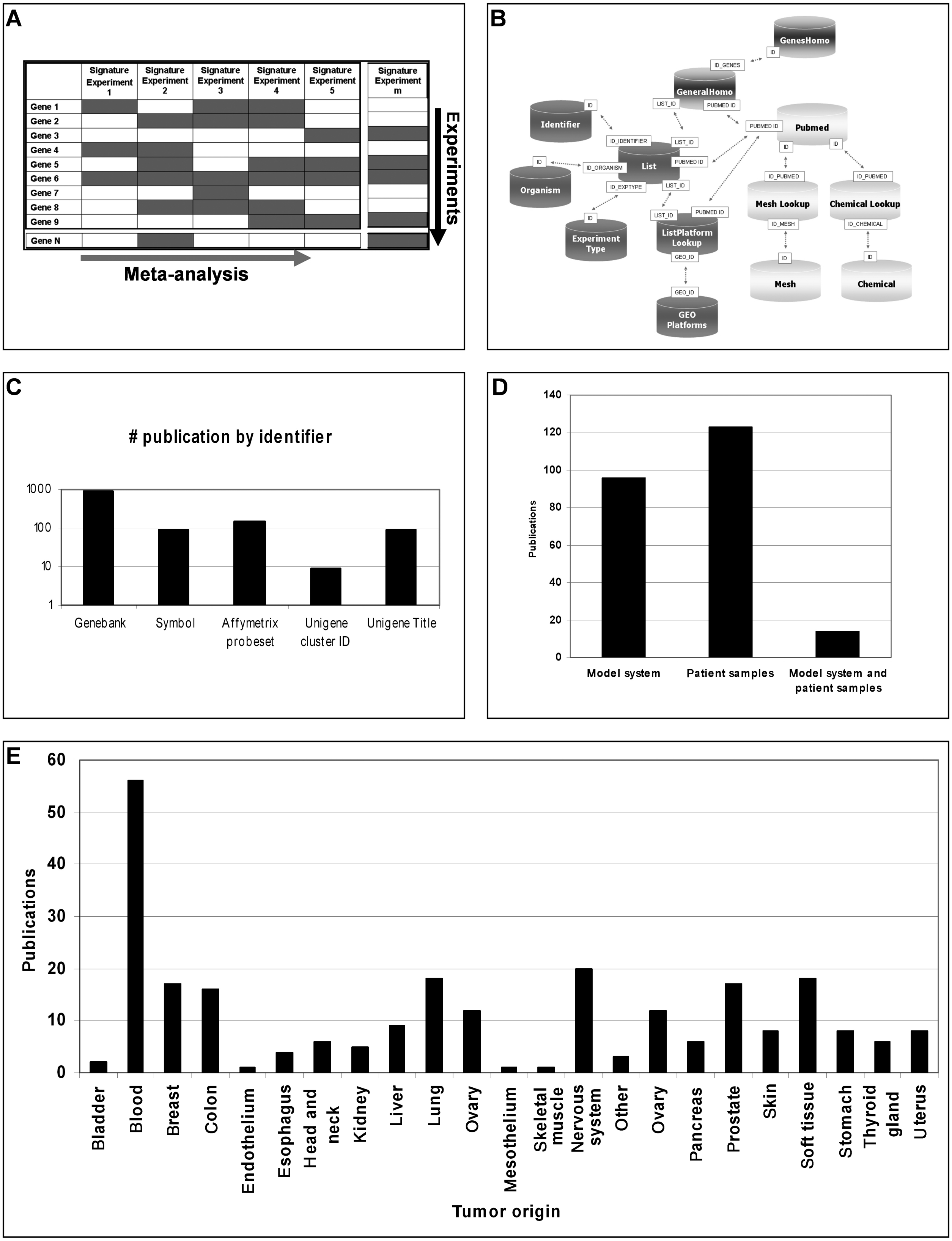

Figure 1. PubLiME, a repository of published gene expression signatures. (**A**) PubLiME stores gene expression signatures as lists of gene identifiers reported by different publications without reference to numerical detail. The resulting simplicity of database design enables efficient cross-experiment and cross-platform gene annotation needed for meta-analysis. (**B**) Gene expression signatures are deposited in a relational MySQL database with three logical areas of database schema: tables are related to genes, lists and publications. (**C**) Signatures classified by the type of identifier used in the original publication. (**D**) Number of publications studying model systems, patient samples or both. (**E**) Number of publications classified by origin of tumors studied.

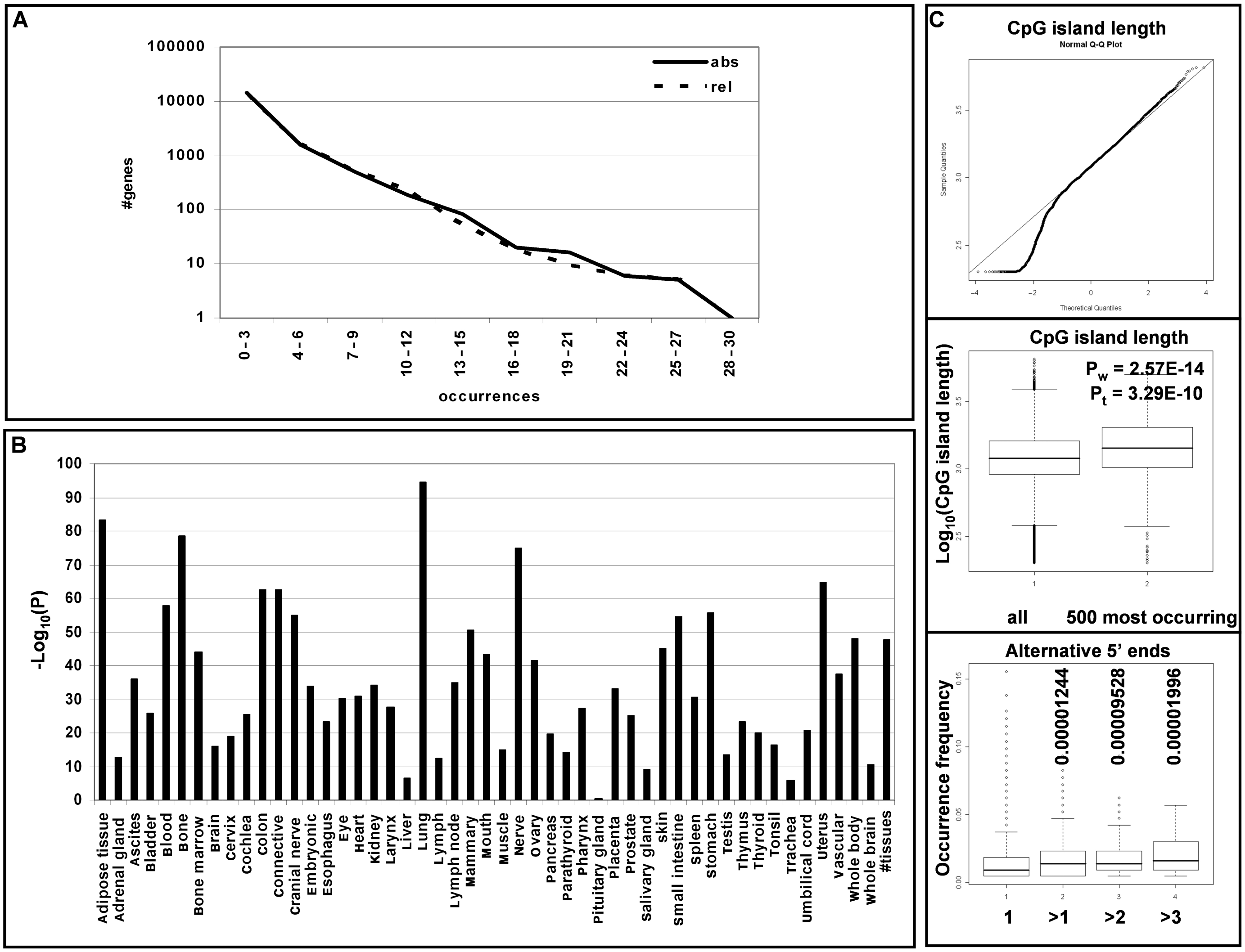

Figure 2. Nonuniform distribution of gene occurrences. (**A**) Graph showing absolute and relative occurrence frequency of PubLiME genes. Relative occurrence frequency is calculated by dividing the absolute occurrence number by the number of studies where the gene was represented on the microarray platform employed. (**B**) Five hundred most frequently occurring genes are expressed at higher levels and in more tissues (last column) as compared to genomic average. Expression is measured in transcripts per million as reported in Unigene Hs.198 expression profiles. Log-transformed Wilcoxon signed rank sum test *P*-values are shown. (**C**) Most occurring genes have longer CpG islands overlapping annotated transcription start sites and more annotated alternative transcripts with different 5′-ends. *Upper panel*: Q–Q plot of genome-wide distribution of CpG island lengths which are overlapping annotated transcription start sites. The plot shows that CpG island lengths follow a nearly normal distribution with some overrepresentation of shorter CpG islands. *Middle panel*: comparison of genome-wide distribution of CpG island lengths ('all') with the distribution of CpG island lengths that are overlapping with start sites of 500 most occurring PubLiME genes. P_t: *T*-test *P*-value, P_w: Wilcoxon signed rank test *P*-value (performed because CpG island lengths distribution is not perfectly normal). *Lower panel*: the distribution of relative occurrence frequencies was calculated for PubLiME genes having exactly one annotated 5′-end or more than the indicated number of different alternative 5′-ends. *P*-value is calculated using Wilcoxon signed rank sum test.

adopted from studies of significantly co-occurring transcription factor binding motifs in promoter sequences which are referred to as *cis*-regulatory modules. The number of signatures a module is required to be part of is called support (Figure 3A). Significance of co-occurrence is evaluated using a generalized form of the Binomial Distribution with trial-specific probabilities (see 'Materials and Methods' section for details). Thus, there are three parameters that influence the number of significant co-occurrence modules: module size, support and *Z*-score. Each significant co-occurrence module is represented as a fully connected undirected graph where genes are nodes and edges are drawn between all pair-wise combinations of genes (Figure 3B). For a set of significant co-occurrence modules, this procedure leads to a network whose numbers of nodes and edges are determined by the stringency of the analysis. Performing the analysis on randomized signatures allows evaluating the effectiveness of a given set of parameters.

The impact of varying the three analysis parameters is illustrated in Figure 3C–E. Signal-to-noise ratios (SNR) are calculated by dividing the number of nodes/edges/modules in the real network by the corresponding number of nodes/edges/modules in the randomized network. Figure 3C shows the effect of raising the *Z*-score cutoff on the SNR for module size 3 and support 5. Figure 3D

Table 1. Genes that are occurring most frequently in PubLiME

Symbol	Occurrences	Detectable	rel_occ_freq
CCND1	30	193	0.155
MYC	27	196	0.138
IL8	25	198	0.126
VEGF	25	198	0.126
TNFAIP3	25	198	0.126
FN1	25	198	0.126
CLU	24	199	0.121
FOS	24	199	0.121
IGFBP4	23	200	0.115
CDKN1A	23	200	0.115
TGFBI	22	201	0.109
TOP2A	22	201	0.109
JUNB	21	202	0.104
PCNA	21	202	0.104
SPARC	21	202	0.104
STAT1	21	202	0.104
SERPINE1	20	203	0.099
IGFBP3	20	203	0.099
GADD45A	20	203	0.099
LGALS1	20	203	0.099

Note: Relative occurrence frequency (rel_occ_freq) is calculated by dividing column 2 (absolute occurrences) by column 3 (number of times the gene was present on the array).

illustrates the impact of the support parameter for module size 3 and Z-score cutoff 7. Support 5 was found to give the best SNR and was used for all further analyses. In other words, we required a module to be observed in at least five publications. This requirement eliminates the need to include genes in the analysis which are present in less than five signatures. Thus, 1642 out of 7476 human genes annotated in PubLiME are included in the analysis. Figure 3E depicts the number of modules in the real and randomized networks as a function of module size for support 5 and Z-score cutoff 7. Most modules are being identified with module sizes 5 and 6 while the best SNR (>1000) is obtained with module sizes 7 and 8. While at module size 3 a SNR of 38 is obtained, the SNR at module size 2 was found less than five for varying Z-score cutoffs, indicating that co-occurrence analysis at module size 2 is not very informative. We conclude that highly significant co-occurrence modules can be identified in PubLiME. Please note that modules are defined from purely qualitative data.

Community formation of genes in co-occurrence module graph

To address the question whether the co-occurrence modules are forming distinct gene sets, we sought to identify strongly connected communities in co-occurrence module graphs. Figure 4A shows the graph representation of co-occurrence modules with module size 3, support 5 and Z-score cutoff 5. Community identification in this graph was carried out following an approach proposed by Newman (10), which is based on recursively removing edges with the largest edge-betweenness. In short, edge-betweenness is a measure that identifies edges in 'between' highly connected communities. Recursive removal of edges leads to a partitioning of the graph into separate communities. The fraction of edges connecting nodes within communities as opposed to edges connecting nodes between communities can be used to define graph modularity which assumes values above 0.3, rarely exceeding 0.7, upon removal of relatively few edges in modular graphs (10) (see 'Materials and Methods' section for details).

We applied this algorithm to co-occurrence module graphs for module sizes 10 to 3 (see supplementary file 'community_composition.xls'). Figure 4B shows the development of graph modularity upon removing edges from the graph shown in Figure 4A and its randomized counterpart. Graph modularity reaches a value of 0.52 upon removal of 231 edges indicating the presence of communities in this graph while in the randomized graph the modularity is mainly negative. We noticed that removing the number of edges corresponding to maximal graph modularity often leads to a partitioning of the graph into communities with highly related functions (as measured by Gene-Ontology-term enrichments). Therefore, we adopted a different strategy for community definition. Starting from high module sizes, we first identified the most significant community forming genes. With decreasing module size, more genes will be inserted into the different communities. The number of edges to be removed from the graph at a given module size was required to minimize the separation of genes that had been assigned to a community at higher module sizes. A detailed report of community composition at different module sizes is shown in the supplementary file 'community_composition.xls'. Genes with changing community assignment at different module sizes were excluded from the analysis (10 in total). Figure 4C depicts the communities that are identified upon removal of 91 edges from the graph of module size 3, support 5 and Z-score cutoff 5, designated C1 to C7. Communities composed of four or fewer genes were not considered further. Gene-Ontology-term enrichment analysis was used to define the putative biological role of these communities: C1—cell cycle ($P = 3.8$ E-23), C2—phosphorylation ($P = 2.9$ E-11), C3—interferon induction ($P = 1.8$ E-11), C4—extracellular matrix ($P = 6.6$ E-13), C5—immune response ($P = 2.6$ E-4), C6—unknown, C7—cell cycle ($P = 3.6$ E-5) and apoptosis ($P = 7.6$ E-5).

A similar definition of communities was obtained by using CFinder, a software that identifies partially overlapping communities by searching k-cliques sharing at least one edge (11) (supplementary file 'community_composition.xls'). Only communities containing more than four genes were considered. In general, CFinder assigns fewer genes to communities and tends to break–up related gene sets as shown by Gene-Ontology-term enrichment. However, assignment of single genes to different communities is surprisingly rare and limited to the two largest communities, suggesting that the communities do reflect distinct rather than strongly overlapping biological functions. Interestingly, the gene assigned to most communities is MYC (four communities). We conclude that genes making up published gene expression signatures can be partitioned into separate communities using two different approaches of community identification.

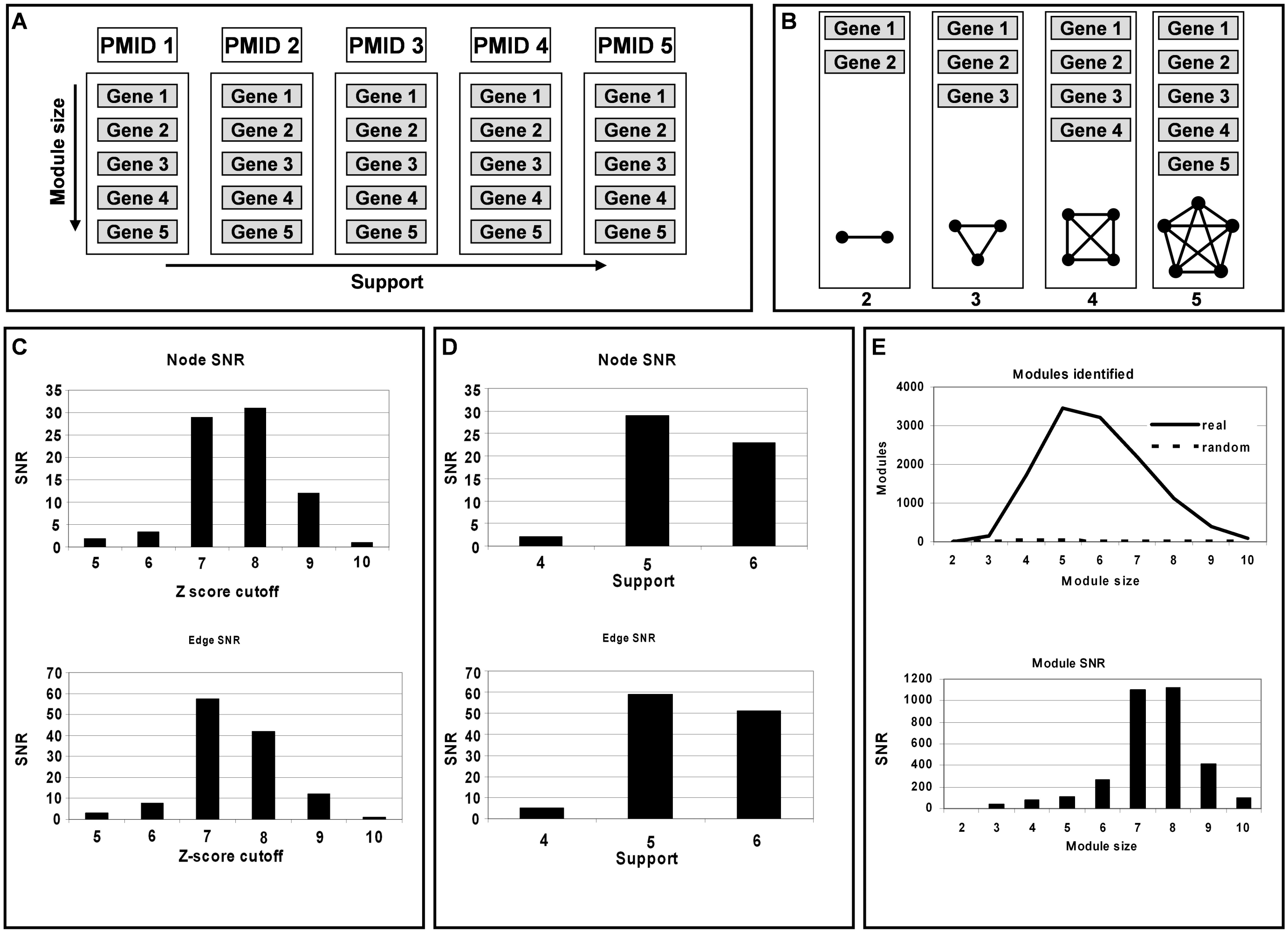

Figure 3. Co-occurrence analysis of genes in PubLiME signatures. (**A**) Parameters of co-occurrence analysis. A module is a set of genes. Module size indicates the number of genes required to co-occur in different signatures. Support is the number of signatures a module is required to be part of. PMID is the PubMed identifier. (**B**) Graph representation of a module. Genes are represented as nodes. Edges are drawn between all pair-wise combinations of genes creating a fully connected graph. Numbers under each graph indicate module size. (**C**) Signal-to-noise ratio (SNR) as a function of *Z*-score cutoff. SNR is calculated as the number of nodes (*upper panel*) or edges (*lower panel*) in the network resulting from the combination of graph representations of all significant modules divided by the corresponding number of nodes (edges) in a network resulting from the analysis of randomized signatures. Module size = 3, support = 5. (**D**) SNR as a function of support. Module size = 3, *Z*-score cutoff = 7. (**E**) Number of nodes and edges as a function of module size in networks resulting from the analysis of real and randomized signatures. Support = 5, *Z*-score cutoff = 7.

Promoter analysis of community genes

Since significant co-occurrence of genes is suggestive of co-regulation, we sought to identify enriched *cis*-regulatory modules of transcription factor binding motifs in the promoters of genes forming different communities. Identification of significant *cis*-regulatory modules was carried out following the same procedure used for detection of co-occurrence modules of genes in different signatures, i.e. co-occurrence of combinations of binding motifs in promoters of community genes is tested. The resulting graph of *cis*-regulatory modules is not tested for the presence of communities, however, but visualized with the aim of identifying the most common motif which will be characterized by the highest node degree (# of edges). In addition, the PageRank algorithm (JUNG software) is used that identifies the node visited most frequently upon random walks along the graph (Figure 5, node size represents PageRank).

For the cell cycle community, E2F was identified as the motif having the largest node degree and page rank. The genes making up this community are strongly enriched ($P = 2.71$ e-17) for genes which we have previously shown to be under control of E2F transcription factors (RFC4, CDC25A, RFC3, MAC30, RRM1, RRM2, BARD1, MCM7, CCNE1, CHAF1A, EZH2, MCM4, PCNA, TFDP1, HMGB2 and FEN1) (16,17). Motif searches in the promoter regions of these genes indicated E2F, Sp1, GC-boxes and NF-Y (CCAAT boxes) as strongly enriched transcription factor binding motifs (17). In general, a number of motifs were identified whose role in the regulation of genes making up the respective communities is either known or compatible with biological

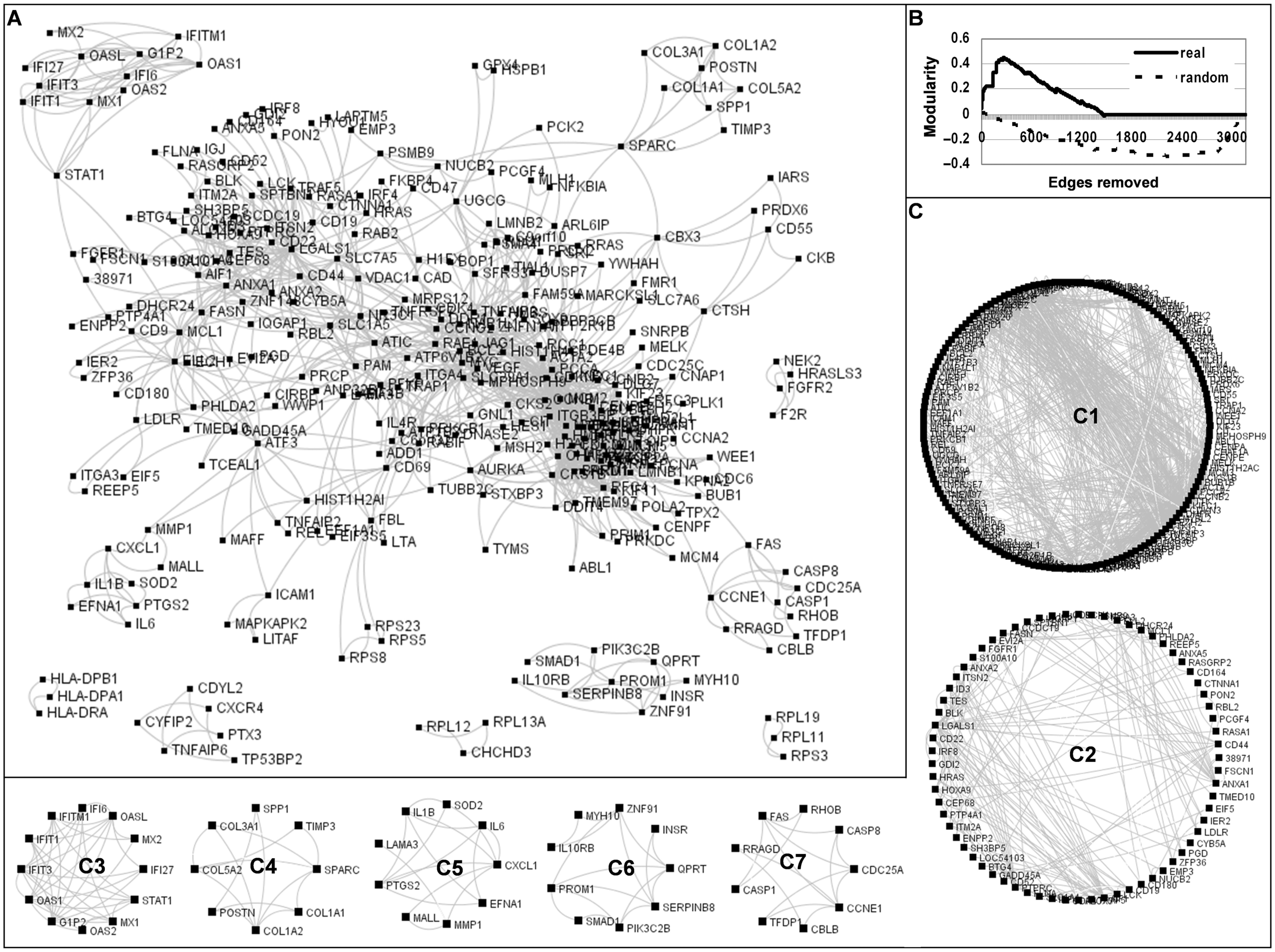

Figure 4. Identification of communities in the graph formed by significant co-occurrence modules. (A) Visualization of co-occurrence module graph with module size 3, support = 5, Z-score cutoff = 5. Some highly interconnected parts of the graph are connected to the body of the graph by very few edges which will be removed from the graph early in the process of edge-betweenness clustering leading to the formation of separate communities and to a corresponding steep increase in graph modularity. (B) Graph modularity as a function of the number of edges removed from the graph in the real and randomized co-occurrence module graph of module size = 3, support = 5, Z-score cutoff = 5. The maximum of graph modularity determines the number of edges to be removed from the graph for the definition of communities. (C) Communities defined by removal of 91 edges from the co-occurrence module graph of module size = 3, support = 5, Z-score cutoff = 5.

intuition. For example, promoters in the extracellular matrix community are found enriched for SOX5 and SOX9 (Figure 5) which were reported to cooperatively activate expression of the COL1A1 promoter (18). Promoters of the immune response community are rich in NFKB and OCT factor binding motifs, known regulators of inflammation (19) and tissue-specific expression of immune system genes (20). In the interferon community we find strong enrichment of interferon regulatory factors 1 and 7 (IRF1, IRF7), known mediators of the interferon response (21,22).

Publications reporting signatures enriched in pathway targets

We queried PubLiME for the identification of publications reporting gene lists that are significantly overlapping with pathways targets. The results of these queries are documented in the supplementary file 'conditions_overlaps.xls'. The pathway targets have been identified in publications reporting human expression profiles. Thus, for human expression profiles, this approach is somewhat circular. It nevertheless provides a detailed account of conditions that lead to deregulation of pathway targets. For murine expression profiles, instead, significant and meaningful overlap with pathway targets can be taken as an independent proof that the pathways identified are biologically relevant and to illustrate that cross-platform and cross-organism annotation are working correctly.

We identified six publications reporting significant overlap with cell cycle targets (C1), two publications with enrichment of immune response targets (C5), one publication with enrichment for interferon response genes (C3), two publications enriched for ECM targets (C4) and one publication with enrichment for cell cycle/apoptosis

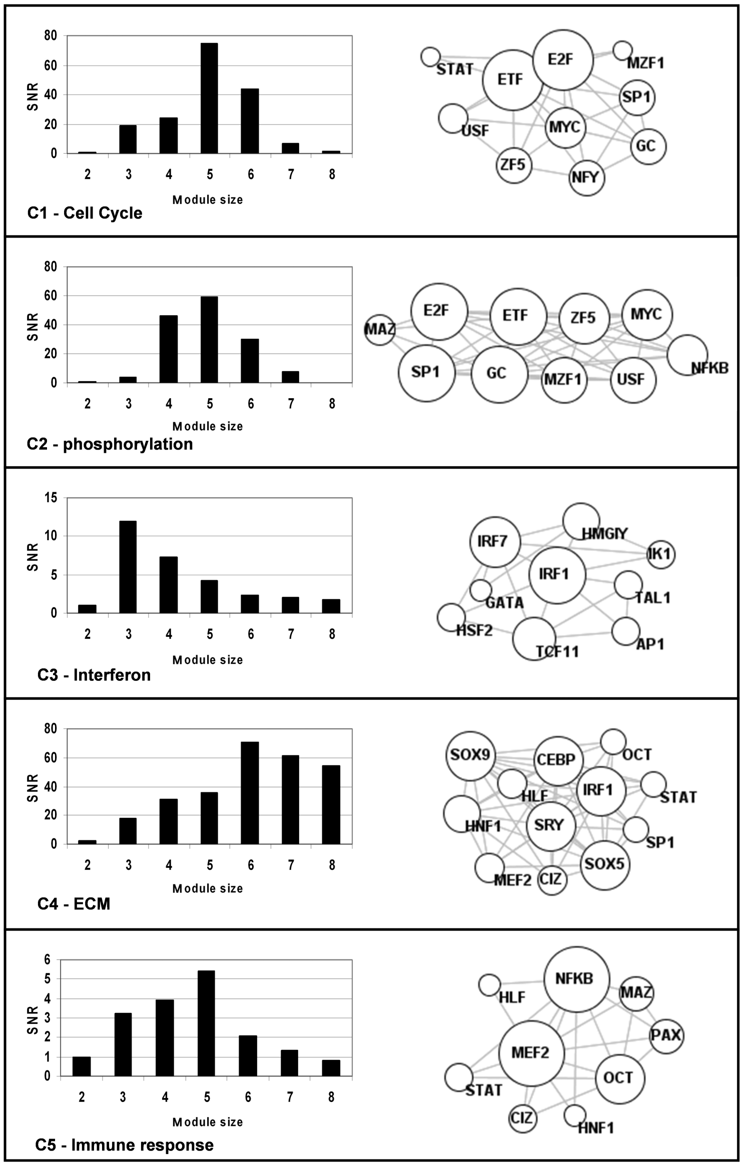

Figure 5. *Cis*-regulatory modules enriched in promoters of community forming genes. Co-occurrence of two up to ten transcription factor motifs in promoters of community forming genes were identified in real promoters and in promoters with randomized assignment of transcription factor binding motifs. A signal-to-noise ratio is calculated by dividing the number of co-occurrence modules in real promoters by the corresponding number in randomized promoters as a function of module size. *Cis*-regulatory modules obtained at module sizes giving the best signal-to-noise ratio are visualized using Netsight software. Node size indicates PageRank, a measure indicating how often a node is visited in random walks along the edges of the graph. Transcription factor binding motifs that are part of many *cis*-regulatory modules will have both a high node degree as well as a high PageRank. All modules were required to be present in at least one-third of the promoters analyzed.

targets (C7). The publications reporting the most significant overlap with cell cycle targets are studying MEFs with knockout of pocket protein family members pRB, p107 and p130 (23), MEFs stimulated with serum and comparison of growth-regulated genes with E2F target genes (24) and expression of SV40 Large T Antigen in neuroendocrine cells (25). The studies with overlapping immune response targets are reporting expression changes in dendritic cells pulsed with tumor antigens (26), and IL-12 treatment of mammary carcinoma cells *in vivo* (27). Deregulation of ECM target genes was reported in mouse cell line models with differential tumorigenicity and metastatic potential (28,29). The human signatures being enriched for pathway target genes show that there is a slight prevalence for blood cell neoplasm studies showing overlap with phosphorylation cascade genes (C2), that the interferon pathway targets (C3) are overlapping with studies employing interferon treatment of cells, and that studies with overlap to ECM genes (C4) are concerned with tumor progression and metastasis. Even though human signatures enriched in pathway targets cannot provide an independent proof, taken together with the overlapping murine signatures, it appears that the cancer signaling pathways identified in this study do reflect biologically relevant phenomena and that the approach presented here, in conjunction with more extended datasets, can help identifying critical regulators of the oncogenic process.

DISCUSSION

Here we present and illustrate the utility of a repository of cancer-related gene expression signatures, PubLiME. As opposed to other repositories of microarray data such as ArrayExpress, GEO, CIBEX and Oncomine, PubLiME stores gene identifiers of genes found differentially regulated in microarray experiments but no numerical data. This approach facilitates cross-platform annotation of gene expression signatures needed for efficient meta-analysis by enormously simplifying database design.

The meta-analysis of PubLiME content presented here is based on using purely qualitative data. No reference is made to any numeric detail and no distinction is made between up and down-regulation of genes. There are three main reasons for proceeding this way: First, the concept of up or down-regulation requires the definition of a base line condition relative to which changes are measured. This is feasible when few conditions are analyzed. During meta-analysis of hundreds of conditions, there will be many base-line conditions defined by different studies. Since current gene expression technology does not provide copy numbers of RNAs, the relative expression levels of genes in different base-line conditions cannot be obtained. Therefore, we just consider whether a gene displays changing expression levels in a given set of conditions. Secondly, the interpretation of the direction of change can be very misleading in the absence of a detailed numerical model of the underlying gene network, which is currently unavailable. For example, we observed counter-intuitive up-regulation of BCL2 by E2F1 even though E2F1 induces apoptosis (16). Thirdly, genes are often represented in a many-to-many relationship on different array platforms. During meta-analysis, it is necessary to define summary measures, which is nontrivial because one has to reconcile often contradictory readouts of different probes measuring the same gene. If one gene is measured by two probes in, say, 200 different conditions, there are $2^{200} = 1.60694E + 60$ different readouts to be reconciled for a single gene!

Considering these potential complications, we explored the possibility of detecting biologically meaningful associations of genes from co-occurrence patterns of gene combinations in different gene expression signatures. The hypothesis tested here is that downstream target genes of cancer signaling pathways should significantly co-occur in gene expression signatures identified in diverse conditions impacting the activity of cancer signaling pathways. We have adopted a combination of co-occurrence analysis of genes in different signatures with a graph-based approach aimed at identifying strongly interconnected communities of co-occurring genes. We found that such communities do exist and that the genes constituting those communities share considerable similarities in biological function as determined by analyzing gene annotations, *cis*-regulatory modules enriched in their promoter regions, and overlapping signatures in independent mouse experiments.

The analysis of occurrence frequencies of genes in PubLiME signatures revealed a highly nonuniform distribution which cannot be attributed to different numbers of times a gene was present on a microarray platform. Among the most frequently occurring genes we found many oncogenes raising concerns that the PubLiME signatures are biased because researchers may tend to focus on known cancer genes. However, PubLiME signatures represent the outcome of statistical analyses of microarray studies listing signatures of median length 52. Researcher bias would require the favorite gene to be explicitly added to those signatures if it was not already present. Second, among the most frequently occurring genes there are many which are not among the most widely studied cancer genes such as Clusterin (24 signatures), TNFAIP3 (25 signatures), CD24 (20 signatures) or genes with unknown function such as C5orf13 (12 signatures). Third, the most frequently occurring genes are characterized by higher expression levels, expression in a larger number of tissues, and larger CpG islands covering their annotated transcription start site. CpG islands are associated with alternative start sites of transcription (15). Indeed, genes with more annotated alternative transcripts have a higher occurrence rate in PubLiME signatures. It is tempting to speculate that the occurrence rate of genes in signatures is partly determined by the number of alternative promoters that respond to different afferent signals. In any case, these evidences make it seem unlikely that the nonuniform distribution of occurrence probabilities is due to researcher bias for favorite cancer genes.

Microarray studies are inherently error prone due to uncertainties in probe specificities and sensitivities, varying methods of analysis and sample impurities. Gene annotation is another potential source of error. Thus, robustness of meta-analysis with respect to noise is a valid concern. We found our approach to be robust because even in the presence of 20% of mis-assigned genes per signature, the modularity of the graph and the communities identified hardly changed (Supplementary Figure S4). We believe that the robustness is a result of module sizes bigger than two applied in this analysis. Since an edge is drawn between two genes when they are part of the same significant co-occurrence module, for module size two there is only one module (the one composed of the two genes under analysis) that determines the presence or absence of an edge. For module sizes larger than two, every pair of genes is part of many modules and only one of them needs to be significant for an edge to be drawn. Thus, even though moderate levels of noise will impact the significance of a considerable number of modules, only extreme levels of noise will eliminate all of them. Noise resistance is also illustrated by the fact that communities identified at highly stringent large module sizes (4–10) contained fewer genes (as expected) but the genes that were part of the same community at large module size did hardly ever change community at module size three (see supplementary file 'community_composition.xls'). The second reason for noise tolerance is the relatively large number of signatures stored in PubLiME (233 human, 40 mouse). For a co-occurrence module to be considered in further analysis, it is required to occur significantly in at least five different signatures. Thus, the modules analyzed for community formation have been validated by independent studies.

The identification of significant co-occurrence modules applied here is based on the abstraction of distinct list-entry pairs where a list is represented by gene expression signatures composed of genes as entries, or by promoters listing transcription factor motifs. The co-occurrence probability of combinations of entries (modules) is calculated using a generalized form of the Binomial Distribution with trial-specific probabilities. This distribution is needed because of list length heterogeneity which causes the occurrence probabilities of entries to be list-specific. It is based on the Binomial Distribution because for every list analyzed there is a binary outcome (module present or not present). Therefore, the analysis can be thought of as a binomial trial where at each throw of a die a different but distorted die is used. We have shown that this approach is proficient in identifying significant co-occurrence modules of genes in signatures and of *cis*-regulatory modules in promoters. However, the abstract nature of list-entry pairs renders it applicable to a wider range of applications. For example, it would be interesting to analyze gene expression signatures in conjunction with ChIP on chip data for oncogenic transcription factors, which could be accomplished by transforming ChIP on chip data to a list-entry format where all the gene promoters bound by a transcription factor are listed using the same gene identifiers as those used in gene expression signatures.

Although the communities identified here are characterized by considerable stability, the analysis could be improved significantly by the availability of more gene expression signatures in PubLiME. Therefore, we encourage microarray researchers to submit their signatures for deposition in PubLiME. As mentioned previously, ChIP on chip data would be equally welcome and suitable to improve the definition of cancer signaling pathways and their downstream targets. The possibility of assembling cancer signaling pathways on-the-fly, without the need for preconfigured gene lists, could enable a novel, interactive way of microarray data analysis where a researcher can build pathways using his signature in conjunction with all other signatures in the repository, discover how his signatures impact pathway communities and which community the genes regulated in his signature belong to.

In conclusion, we show that genes occur in a strongly nonrandom fashion in published gene expression signatures. Co-occurrence analysis can be used to identify co-occurrence modules of genes that are strongly over-represented. A graph-based approach aimed at the identification of interconnected communities can be applied to co-occurrence modules of genes to show that they are forming distinct communities. The genes making up separate communities are enriched for regulators of cell cycle, apoptosis, phosphorylation cascades, extracellular matrix, immune and interferon response regulators and some of them are forming communities of unknown function. For the majority of communities, promoter searches for enriched *cis*-regulatory modules support the conclusion that the communities identified here reflect biologically relevant sets of co-regulated genes whose expression is altered in human cancer. As such, the identified communities may provide marker genes useful for clinical applications as well as hitherto unknown regulators of cancer signaling pathways that may constitute novel entry points for pharmacological intervention.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by grants from AIRC (Italian Association of Cancer Research) and CARIPLO (Cassa di Risparmio della Regione Lombardia). Part of this work has been conducted on a HP rx4640 Itanium2 server granted to IFOM in the framework of the 'HP Integrity' program coordinated by Dr. Alessandro Guffanti. Funding to pay the Open Access publication charge was provided by AIRC.

Conflict of interest statement. None declared.

REFERENCES

1. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P. *et al.* (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
2. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
3. Ikeo,K., Ishi-i,J., Tamura,T., Gojobori,T. and Tateno,Y. (2003) CIBEX: center for information biology gene expression database. *C R Biol.*, **326**, 1079–1082.
4. Rhodes,D.R., Yu,J., Shanker,K., Deshpande,N., Varambally,R., Ghosh,D., Barrette,T., Pandey,A. and Chinnaiyan,A.M. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*, **6**, 1–6.
5. Finocchiaro,G., Mancuso,F. and Muller,H. (2005) Mining published lists of cancer related microarray experiments: identification of a gene expression signature having a critical role in cell-cycle control. *BMC Bioinformatics*, **6**, S14.
6. Newman,J.C. and Weiner,A.M. (2005) L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol.*, **6**, R81.
7. Dennis,G.Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.
8. Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.
9. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
10. Newman,M.E. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.*, **69**, 026113.
11. Palla,G., Derenyi,I., Farkas,I. and Vicsek,T. (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814–818.

12. Guffanti,A., Finocchiaro,G., Reid,J.F., Luzi,L., Alcalay,M., Confalonieri,S., Lassandro,L. and Muller,H. (2003) Automated DNA chip annotation tables at IFOM: the importance of synchronisation and cross-referencing of sequence databases. *Appl. Bioinformatics*, **2**, 245–249.
13. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
14. Carninci,P., Sandelin,A., Lenhard,B., Katayama,S., Shimokawa,K., Ponjavic,J., Semple,C.A., Taylor,M.S., Engstrom,P.G. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, **38**, 626–635.
15. Smale,S.T. and Kadonaga,J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.
16. Muller,H., Bracken,A.P., Vernell,R., Moroni,M.C., Christians,F., Grassilli,E., Prosperini,E., Vigo,E., Oliner,J.D. and Helin,K. (2001) E2Fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis. *Genes Dev.*, **15**, 267–285.
17. Vernell,R., Helin,K. and Muller,H. (2003) Identification of target genes of the p16INK4A-pRB-E2F pathway. *J. Biol. Chem.*, **278**, 46124–46137.
18. Lefebvre,V., Li,P. and de Crombrugghe,B. (1998) A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene. *EMBO J.*, **17**, 5718–5733.
19. Karin,M. and Greten,F.R. (2005) NF-kappaB: linking inflammation and immunity to cancer development and progression. *Nat. Rev. Immunol.*, **5**, 749–759.
20. Schubart,K., Massa,S., Schubart,D., Corcoran,L.M., Rolink,A.G. and Matthias,P. (2001) B cell development and immunoglobulin gene transcription in the absence of Oct-2 and OBF-1. *Nat. Immunol.*, **2**, 69–74.
21. Honda,K., Yanai,H., Negishi,H., Asagiri,M., Sato,M., Mizutani,T., Shimada,N., Ohba,Y., Takaoka,A. *et al.* (2005) IRF-7 is the master regulator of type-I interferon-dependent immune responses. *Nature*, **434**, 772–777.
22. Harada,H., Fujita,T., Miyamoto,M., Kimura,Y., Maruyama,M., Furia,A., Miyata,T. and Taniguchi,T. (1989) Structurally similar but functionally distinct factors, IRF-1 and IRF-2, bind to the same regulatory elements of IFN and IFN-inducible genes. *Cell*, **58**, 729–739.
23. Black,E.P., Huang,E., Dressman,H., Rempel,R., Laakso,N., Asa,S.L., Ishida,S., West,M. and Nevins,J.R. (2003) Distinct gene expression phenotypes of cells lacking Rb and Rb family members. *Cancer Res.*, **63**, 3716–3723.
24. Ishida,S., Huang,E., Zuzan,H., Spang,R., Leone,G., West,M. and Nevins,J.R. (2001) Role for E2F in control of both DNA replication and mitotic functions as revealed from DNA microarray analysis. *Mol. Cell Biol.*, **21**, 4684–4699.
25. Hu,Y., Ippolito,J.E., Garabedian,E.M., Humphrey,P.A. and Gordon,J.I. (2002) Molecular characterization of a metastatic neuroendocrine cell cancer arising in the prostates of transgenic mice. *J. Biol. Chem.*, **277**, 44462–44474.
26. Grolleau,A., Misek,D.E., Kuick,R., Hanash,S. and Mule,J.J. (2003) Inducible expression of macrophage receptor Marco by dendritic cells following phagocytic uptake of dead cells uncovered by oligonucleotide arrays. *J. Immunol.*, **171**, 2879–2888.
27. Shi,X., Cao,S., Mitsuhashi,M., Xiang,Z. and Ma,X. (2004) Genome-wide analysis of molecular changes in IL-12-induced control of mammary carcinoma via IFN-gamma-independent mechanisms. *J. Immunol.*, **172**, 4111–4122.
28. Jechlinger,M., Grunert,S., Tamir,I.H., Janda,E., Ludemann,S., Waerner,T., Seither,P., Weith,A., Beug,H. and Kraut,N. (2003) Expression profiling of epithelial plasticity in tumor progression. *Oncogene*, **22**, 7155–7169.
29. Wang,Z., Liu,Y., Mori,M. and Kulesz-Martin,M. (2002) Gene expression profiling of initiated epidermal cells with benign or malignant tumor fates. *Carcinogenesis*, **23**, 635–643.