
© 2012 The Author(s).
Reprinted with permission.
Data-Driven Information Retrieval in Heterogeneous Collections of Transcriptomics Data Links SIM2s to Malignant Pleural Mesothelioma

José Caldas 1,∗ Nils Gehlenborg 2,3,4,∗, Eeva Kettunen 5, Ali Faisal 1, Mikko Rönty 6, Andrew G. Nicholson 7, Sakari Knuutila 8, Alvis Brazma 2, Samuel Kaski 1,9,†

1 Aalto University School of Science, Department of Information and Computer Science, Helsinki Institute for Information Technology HIIT, Helsinki, Finland
2 Functional Genomics Group, European Bioinformatics Institute, Cambridge, United Kingdom
3 Graduate School of Life Sciences, University of Cambridge, Cambridge, United Kingdom
4 Current address: Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA
5 Health and Work Ability, Biological Mechanisms and Prevention of Work-Related Diseases, Finnish Institute of Occupational Health, Helsinki, Finland
6 HUSLAB, Department of Pathology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
7 Department of Histopathology, Royal Brompton Hospital, London, United Kingdom
8 Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
9 Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

Associate Editor: Dr. Jonathan Wren

ABSTRACT
Motivation: Genome-wide measurement of transcript levels is an ubiquitous tool in biomedical research. As experimental data continues to be deposited in public databases, it is becoming important to develop search engines that enable the retrieval of relevant studies given a query study. While retrieval systems based on meta-data already exist, data-driven approaches that retrieve studies based on similarities in the expression data itself have a greater potential of uncovering novel biological insights.

Results: We propose an information retrieval method based on differential expression. Our method deals with arbitrary experimental designs and performs competitively with alternative approaches, while making the search results interpretable in terms of differential expression patterns. We show that our model yields meaningful connections between biological conditions from different studies. Finally, we validate a previously unknown connection between malignant pleural mesothelioma and SIM2s suggested by our method, via RT-PCR in an independent set of mesothelioma samples.

Availability: Supplementary data and source code are available from http://www.ebi.ac.uk/fg/research/rex.
Contact: samuel.kaski@aalto.fi

1 INTRODUCTION
DNA microarrays are a frequently used high-throughput tool for measuring gene expression, which is reflected in the continuously increasing amount of data available in public repositories such as the Gene Expression Omnibus (GEO) (Barrett et al., 2009) or ArrayExpress (Parkinson et al., 2009). The thousands of gene expression studies in these repositories make it increasingly challenging to retrieve data sets that are relevant to the user. At the same time, the availability of these collections gives us the opportunity to develop retrieval methods that take into account the gene expression data from these studies to deliver biologically meaningful results and provide insights into the molecular mechanisms at work in the deposited studies.

There are two possible types of solutions for the task of retrieving relevant studies from databases. Knowledge-driven approaches are based on the metadata used to describe the deposited gene expression studies. Various forms of string matching algorithms have been applied to retrieve studies based on a textual query (Zhu et al., 2008). Advanced solutions incorporate controlled vocabularies or ontologies for semantic query expansion (Malone et al., 2010). Given high-quality annotations, the likelihood of biological relevance of the results is high, but methods using this paradigm are limited to retrieving studies annotated with a known label. Moreover, these approaches are fundamentally limited by the

∗These authors contributed equally to this work.
†To whom correspondence should be addressed.

© The Author(s) 2011. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
fact that the text-based description of a study and its results contains only a fraction of the information in the actual gene expression data.

Data-driven or content-based approaches to information retrieval or meta-analysis (Hunter et al., 2001; Segal et al., 2004; Lamb et al., 2006; Fujibuchi et al., 2007; Kapushesky et al., 2009; Caldas et al., 2009; Hu and Agarwal, 2009; Huang et al., 2010; Kupershmidt et al., 2010; Engreitz et al., 2011) have a high potential for discovering novel and biologically meaningful relationships between the studied tissues, organisms, and biological conditions, since similarities between studies are derived from shared expression patterns. Differential expression is a natural encoding for a study, as it describes the biological variation between the studied conditions. It is also a very useful basis for data-driven retrieval in heterogeneous collections of gene expression studies, and meta-analysis in general, as it addresses issues such as inter-platform incommensurability.

Data-driven information retrieval or meta-analysis methods typically consist of the following four components: (1) a decomposition of the experimental design of studies into differential expression (pairwise comparison) of genes or gene sets; (2) a method to measure the significance of differential expression (e.g., fold-change, t-test, or Gene Set Enrichment Analysis (GSEA); Subramanian et al., 2005), which serves as a basis for encoding the studies; (3) a method to extract biological patterns of interest from the encoded studies; and (4) a relevance measure between studies, conditions, or microarrays. Supplementary Table S1 describes the various existing approaches for each of the components. Depending on their scope, most existing methods include only a subset of the components. For instance, the well-known meta-analysis module map method (Segal et al., 2004) does not include an approach for computing the relevance between studies. Conversely, several information retrieval methods do not make use of any method to extract shared expression patterns (Hunter et al., 2001; Lamb et al., 2006; Fujibuchi et al., 2007; Hu and Agarwal, 2009; Kupershmidt et al., 2010).

Three challenges that are particularly significant in the context of large and highly heterogeneous gene expression repositories but that so far have not been addressed are the decomposition of studies with an arbitrary experimental design, facilitation of the biological interpretation of the retrieval results, and the systematic evaluation of the retrieval performance. For instance, most methods are designed to deal only with studies comparing case vs. control. Other methods are able to handle studies with arbitrary designs, but decompose studies into comparisons in ways that induce study-specific bias and hinder the interpretation of the retrieval results. As an example, comparing two phenotypes (e.g. normal vs. disease) in a multi-factorial study while ignoring additional experimental variables may introduce confounding factors.

In this paper, we propose RXeasy (data-driven Retrieval of Experiments), which extends our earlier data-driven information retrieval method (Caldas et al., 2009). An overview of the key steps of the method is provided in Figure 1. First, for the decomposition of studies into pairwise comparisons, we introduce an approach that takes into account the fact that a comparison depends not only on the phenotypes being compared, but also on the phenotypes which are held constant in the comparison, i.e. the context. In each comparison, the other experimental factors need to have the same values in order to avoid confounding factors. Unlike in our previous work, this approach is applicable to any type of experimental factor. The underlying data-driven modeling has also been extended.

Our proposed unsupervised learning model enables the detection of associations between studies and the interpretation of these associations in terms of recurrent patterns of differential expression. The new model additionally takes into account correlations in the activity of gene expression patterns; moreover, while the earlier method worked purely on the level of gene sets, we now additionally model the activity of the specific genes in the sets to increase accuracy and enable more specific interpretations. Finally, we propose a novel ontology-based approach for evaluating the retrieval results, to deal with the wide range of biological and medical subject areas spanned by the studies in the repository.

We apply RXeasy to a collection of 1092 studies taken from the ArrayExpress repository, involving three species (human, mouse, and rat) and corresponding to a total of 6925 phenotype comparisons (in our previous feasibility study, we applied our method to less than 800 comparisons derived from human studies). We show that the inferred differential expression patterns correspond to functionally coherent core intersections of gene sets. We also demonstrate that the numerical retrieval performance of our method is competitive with existing approaches. In a series of case studies, we point out that connections between conditions found by our method have been confirmed in independent studies. These case studies illustrate how conditions can be connected on a molecular level, and provide evidence for the validity of our approach.

In an experimental validation study, we explored a connection found by our method that hints at a potential role of the basic helix-loop-helix transcription factor Single-minded homolog 2 (SIM2s) in malignant pleural mesothelioma (MPM), which has not been previously described in the literature. Using real-time polymerase chain reaction (RT-PCR), we were able to detect significant SIM2s under-expression in an independent set of MPM tumors, indicating that SIM2s may effectively have a role in MPM. This shows that our data-driven information retrieval approach can indeed be used to obtain novel biological insights from large and heterogeneous collections of transcriptomics data.

2 METHODS

2.1 Data

2.1.1 Gene Expression Studies Data sets from transcriptomics studies in human (Homo sapiens), mouse (Mus musculus) and rat (Rattus norvegicus) were obtained from the ArrayExpress Archive on 26 October 2009 by selecting all data sets that include a preprocessed expression matrix and sufficiently curated annotation. The data sets fulfilling these criteria are also included in the ArrayExpress Atlas database (Kapushesky et al., 2009) and the same underlying data were used to construct our collection. A total of 1092 microarray data sets were retrieved. Out of these, 479 were from human, 445 were from mouse, and 168 were from rat studies.

2.1.2 Gene Sets For our analysis we used the canonical pathway gene set collection (C2.CP) provided by the Molecular Signature Database (Version 2.5) (Subramanian et al., 2005). This collection contains 639 gene sets that represent pathways from a range of public databases.

2.1.3 Tumor Specimens and RT-PCR Tumor tissue specimens were obtained from ten malignant pleural mesothelioma (MPM) patients that were diagnosed with mesothelioma tumor at Royal Brompton and Harefield NHS Trust, United Kingdom. Of those, six were epithelial and four were biphasic
MPMs. As a control we used a microscopically normal scraped pleural tissue lining of the lung of a 39 year old, previously healthy male patient operated at the Helsinki University Central Hospital for a non-neoplastic intrabronchial inflammatory polyp. We then measured the expression levels of MMP2, MMP3, MMP14, SNAI1, SNAI2, MYOM2, SIM2l, and SIM2s via RT-PCR. We provide the full details of our experimental procedure in Supplementary Text S1.

2.2 Information Retrieval Framework

2.2.1 Study Decomposition

The collected and preprocessed data sets were decomposed into binary comparisons between two conditions, denoted by A and B, to be able to determine differentially expressed genes and gene sets. We applied the following criteria:

1. All samples for the conditions A and B are annotated with exactly one of two different factor values that belong to the same experimental factor.
2. If there are additional experimental factors used in the study, the factor values of each of those must be the same for all samples associated with conditions A and B. These factor values form the context of the comparison.
3. For each condition there must be at least three samples.
4. Neutral factors are removed before studies are decomposed into comparisons. Neutral factors are factors that would not result in meaningful comparisons and have a very large number of associated factor values within a study. The factors “age” (without stratification) or “individual” are examples for such cases. The full list of neutral factors is shown in Supplementary Text S2.

We extracted all possible comparisons according to these rules, which resulted in a total of 6925 comparisons. Of those, 1976 are from human studies, 2137 are from mouse studies and 2812 are from rat studies. The extracted comparisons were further classified into whether they are interpretable or not. We define a comparison to be interpretable if either A or B can be considered as a “control” or “normal” state in the experiment. Such conditions are, for example, wild type strains when different genotypes are being compared, a mock treatment when the effects of drugs are analyzed, or healthy tissues when cancers are studied. The assumption is that the effects observed in an interpretable comparison can be attributed to the non-control condition.

In order to identify interpretable comparisons, we assembled a list of control factor values by manually classifying all factor values used in the collection of data sets. The full list of control factor values is shown in Supplementary Text S3. We were able to classify a total of 908 comparisons as interpretable, with 325 coming from human, 429 coming from mouse and 154 coming from rat studies. The number of interpretable comparisons is almost nine times higher than in our earlier study, where only 105 interpretable comparisons were used. Furthermore, in our earlier work we only considered comparisons of disease against some control as interpretable, whereas here we considered interpretable comparisons derived from a wide range of different experimental factors.

2.2.2 Differential Expression

We use the signal-to-noise ratio as a measure of differential expression of each gene in each comparison. We then apply GSEA version 2.04 (Subramanian et al., 2005) to test for the overrepresentation of pre-defined gene sets among the most up or down-regulated genes, and collect the 50 gene sets with the highest normalized score, ignoring the direction of differential expression. Unlike in previous work (Caldas et al., 2009), we also consider the most differentially expressed genes in each gene set, a subset known as the leading edge subset (Subramanian et al., 2005). We provide additional details in Supplementary Text S4.
2.2.3 Unsupervised Learning Method

We propose a latent variable mixture model for analyzing the GSEA results. Patterns of gene set and gene differential expression are represented as mixture components and GSEA comparisons are encoded as soft combinations of those components. The model structure is shown in Figure 2. We assume there are \( T \) mixture components, or submodules, with the \( t \)-th submodule consisting of two vectors of Bernoulli distributions, \( \phi_t \) and \( \psi_t \). The vector \( \phi_t \) has length equal to the number of gene sets and models the binary activation status of each gene set; the vector \( \psi_t \) has length equal to the total number of genes in the data set and models the leading edge subset of each gene set. The activation status of a gene set \( j \) in a given GSEA comparison and the composition of its leading edge subset are assumed to be generated by first picking a submodule \( t \); then, the binary activation status of gene set \( j \) is a sample from a Bernoulli distribution parameterized by \( \phi_{t,j} \), while for each gene \( g \) in that gene set we generate its leading edge subset membership by sampling from a Bernoulli distribution with parameter \( \psi_{t,g} \). In order to model correlations between submodules, we incorporate a two-level submodule selection procedure (Li and McCallum, 2006); we assume that each GSEA comparison has a discrete distribution over so-called modules, parameterized by a vector \( \theta \); each module \( m \) has a discrete distribution over submodules, parameterized by a vector \( \eta_m \). The selection of a submodule \( t \) is made by first choosing a module \( m \) using \( \theta \), and then choosing submodule \( t \) using \( \eta_m \). The variables \( u \) and \( v \) in Figure 2 indicate the chosen module and submodule, respectively. Finally, we endow each \( \theta \) and \( \eta_m \) with conjugate symmetric Dirichlet prior distributions, and each \( \phi_t \) and \( \psi_t \) with conjugate symmetric Beta prior distributions, parameterized by \( \alpha_\phi \), \( \alpha_\psi \), \( \alpha_u \), and \( \alpha_v \), respectively. The conjugate prior distributions are primarily chosen for the purpose of analytical tractability, as it allows us to derive a collapsed Gibbs sampler for inference and estimation, which has been shown to work well in latent variable mixture models (Griffiths and Steyvers, 2004).

We use a collapsed Gibbs sampler (Griffiths and Steyvers, 2004) to compute approximate posterior distributions for \( u \) and \( v \) as well as estimates for \( \theta \), \( \eta_m \), \( \phi_t \), and \( \psi_t \) given the observed GSEA results and a pre-defined number of modules and submodules.

The relevance of a GSEA comparison \( r \) to a query \( q \) is computed as the expected probability that the parameters of comparison \( r \) generated the data in comparison \( q \). Using a general probabilistic formulation, this amounts to computing

\[
rel(q,r) \stackrel{\text{def}}{=} \int P(x|\Psi_r)P(\Psi|X)d\Psi,
\]

where \( X \) is the input data and \( \Psi \) is the collection of random variables upon which inference is performed (Buitine et al., 2004).

Finally, our model allows computing for each comparison the marginal probability that each gene set is active. Using the inferred estimates for the model variables, the marginal probability of a gene set being active in a given comparison is given by the following expression:

\[
P(\text{gene set } s \text{ is active | comparison } i) = \frac{\sum_{m=1}^{M} \sum_{t=1}^{T} \theta_{im} \eta_{mt} \phi_{ts}}{\sum_{m=1}^{M} \sum_{t=1}^{T} \theta_{im} \eta_{mt} \phi_{ts}} \quad (1)
\]

The full details of our model are described in Supplementary Text S5.

2.3 Performance Evaluation

In our previous work, the evaluation of retrieval results relied on a manual classification of comparisons into “cancer-related” and “not cancer-related” (Caldas et al., 2009). This was possible because the number of comparisons was fairly small. For the REs method described here, we developed a scalable approach that employs an ontology-based relevance score to evaluate the performance of the method.

The Experimental Factor Ontology (EFO; Malone et al., 2010) is a representation of the relationships between experimental factor values used in the studies in ArrayExpress and essentially a directed, acyclic graph with a root. Each experimental factor value corresponds to a path between the root and a downstream node, with more specific terms generally being further away from the root. For evaluation purposes, and to compare our method to other information retrieval methods, we used the EFO as an external “gold standard”, based on which the relevance of a retrieved comparison given a query is measured. This approach is a systematic solution for evaluating retrieval results from a large, heterogeneous collection of studies that contains data on a wide range of subjects, that would otherwise require a large number of experts from different fields to evaluate the results; this expert knowledge is partially encoded in the ontology.

To evaluate retrieval performance with the EFO, we used an expert-curated mapping to associate the experimental factor values that define interpretable comparisons with terms in the EFO (Release 1.7), if possible. The mapping is also used for the ArrayExpress Atlas and available as a table in the ArrayExpress database. When the non-control condition of an interpretable comparison can be mapped to the EFO, we call the comparison an "interpretable" comparison. A total of 219 interpretable comparisons were identified based on the mapping from the ArrayExpress Atlas, with 137 coming from human studies, 39 coming from mouse studies and 43 coming from rat studies.

To compute the similarity between terms in the EFO and thus between comparisons in our collection, we employed a modified version of the Jaccard coefficient (Manning et al., 2008), which yields a graded relevance score between 0 and 1. We then applied the Normalized Discounted Gain (NDCG) measure (Jarvelin and Kekäläinen, 2002) to evaluate REs based on the modified Jaccard coefficient. The approach is described in detail in Supplementary Text S6.

2.4 Module and submodule interpretation

We used a statistical significance approach to compute a collection of gene sets and genes with a high activation probability for each module and submodule. Here, we describe the procedure only for submodules; for modules, the only difference is that it is first necessary to compute module-to-gene-set and module-to-gene probabilities by standard marginalization. For submodule \( k \), we first computed the probability that the submodule activates both gene set \( s \) and gene \( g \) via the product \( \phi_{k,s} \psi_{k,g} \delta_{s,g} \), where \( \delta_{s,g} \) asserts if gene \( g \) belongs to gene set \( s \). We then assessed which genes have a significantly high probability of being activated relative to other genes. This was done by using a one-tailed Wilcoxon rank-sum test, where the samples being compared are all the (gene set, gene) joint probabilities that involve a particular gene vs. all other joint probabilities. An equivalent approach was used for gene sets. Significance was assessed at the standard \( q \)-value threshold of \( q < 0.05 \). This allows obtaining for each submodule a list of significantly probable genes and gene sets. To further bind the two lists, we pruned the list of significant gene sets by removing those which are not overrepresented in the list of significant genes, as assessed by a hypergeometric test with a cut-off of \( q < 0.05 \).

3 RESULTS AND DISCUSSION

3.1 Case Studies

We retrieved the top 25 most relevant results for each of the 908 interpretable comparisons in our collection and created HTML-based reports for each of these queries. The full list of reports is available online at http://www.ebi.ac.uk/fgg/research/rex.

Using these reports we performed a series of case studies in order to obtain a qualitative evaluation of the retrieval performance of REX. In each case study we interpreted the retrieval results for one or more query comparisons with the help of the reported most relevant gene sets and the literature. Due to space constraints the details of the case studies are described and discussed in Supplementary Text S7, S8, and S9. In summary, we were able to use REs to identify links between conditions such as malignant melanoma and cardiomyopathies, or between pancreatic cancer, insulin signaling,
diabetes mellitus, and inflammation. REx also identified a set of comparisons from different studies that were all related to the central nervous system.

3.2 RT-PCR Experimental Validation: SIM2s Expression in Malignant Pleural Mesothelioma

We queried the database with a comparison of malignant pleural mesothelioma (MPM) vs. normal in human pleura. The top 25 most relevant comparisons are presented in Supplementary Table S2. The top two retrieved comparisons come from the same study and test the effect of potassium and thapsigargin in human cerebrovascular smooth muscle cells. Both potassium and thapsigargin lead to elevated levels of Ca\(^{2+}\), by activating Ca\(^{2+}\) influx channels and depleting intracellular Ca\(^{2+}\) storage, respectively (Pulver-Kaste et al., 2006). Abnormal levels of Ca\(^{2+}\) can promote tumor cell proliferation and resistance to apoptosis (Feng et al., 2010), which potentially explains the connection to MPM. A hallmark for epithelial and biphasic MPM is the expression of the calcium binding protein calretinin, which is used in the identification of the tumors, although it remains unclear what might be its putative role in carcinogenic processes (Henzi et al., 2009).

The third most relevant comparison is an investigation of an RNAi knockdown of SIM2s (single-minded homolog 2, short isoform) in a human colon carcinoma cell line at 18 hours. SIM2, located on chromosome 21, encodes a basic helix-loop-helix transcription factor and has two splicing isoforms, SIM2s (short) and SIM2l (long). Due to its chromosomal location, SIM2 has been associated with Down syndrome (trisomy 21). For instance, over-expression of SIM2 has been shown to induce a partial Down syndrome phenotype in mouse (Chrst et al., 2000). Due to the fact that individuals with Down syndrome have a higher risk for leukaemia but a lower risk for solid tumors than the general population (Hasle et al., 2000), there are genes on chromosome 21 that are likely candidates for tumor suppressors or oncogenes (Laffin et al., 2008). SIM2s has been found to be over-expressed in colon and prostate cancer (Alemán et al., 2005; Halvorsen et al., 2007), and under-expressed in breast cancer (Kwik et al., 2007). The connection found by REx suggests SIM2s may be differentially expressed in MPM. To the best of our knowledge, SIM2s has not yet been identified as having a role in MPM. Interestingly, Sim2 expression was found in the mesothelium of mice during embryonic development, whereas Sim2 mutant mice died within 3 days of birth from breathing failure due to the defects in the structural components surrounding the pleural cavity, such as pleural mesothelium tearing. After severe dyspnea, disruption of the pleural mesothelium basement membrane was observed in Sim2 mutants (Goshu et al., 2002). In the MPM study analyzed by our model (Gordon et al., 2005), SIM2s was slightly under-expressed in comparison to a pleural control (fold-change = 0.87). We tested via RT-PCR measurements whether SIM2s under-expression could be observed in an independent set of 10 MPM patients. This set consisted of six epithelial MPM and four biphasic MPM (both histological subtypes are included in the original MPM study (Gordon et al., 2005) analyzed by our model). We also quantified the expression of genes known to be closely related to SIM2s, namely its transcriptional targets MYOM2 (Woods et al., 2008), MMP3 (Kwik et al., 2007), MMP2, and SNAI2 (Laffin et al., 2008). Finally, we also measured the expression of MMP14, which has been recently observed to be differentially expressed in MPM (Crispi et al., 2009), as well as the expression of SNAI1 and SIM2l. The log-ratio results are presented in Figure 3.

SIM2s was significantly under-expressed (p < 0.05) in MPM patients in comparison to a pleural control. MMP3 and SIM2l expression was detected in all MPM specimens (except MMP3 in one biphasic sample) but not in the pleural control. While this implied differential expression of those genes in MPM, the lack of expression in the pleural control precluded us from obtaining numerical fold-change values. Although we did not confirm significant over-expression of MMP14 reported earlier (Crispi et al., 2009), the expression levels of MMP14 were significantly correlated with the expression of MMP2 (r = 0.74, p < 0.05), in accordance with the fact that MMP14 is required for MMP2 activation (Crispi et al., 2009). However, we did observe significant over-expression of SNAI2 (p < 0.05). Over-expression of SNAI2 and MMP3 is consistent with their potential role as repressive transcriptional targets of SIM2s (Laffin et al., 2008; Kwik et al., 2007). Finally, over-expression of MYOM2 is consistent with the fact that it can be activated by both short and long isoforms of SIM2 (Woods et al., 2008); in the analyzed MPM samples we observed SIM2l over-expression.

The fact that we observed statistically significant SIM2s under-expression in an independent set of MPM patients suggests that SIM2s may be a relevant gene in MPM. Currently, no known role for SIM2s in MPM has been described. However, it has been observed that SIM2s RNAi silencing in MCF-7 cells induces an epithelial-mesenchymal transition (EMT)-like phenotype and estrogen receptor (ER) α-negative tumors in mouse via an MCF-7 xenograft assay (Laffin et al., 2008). Over-expression of EMT-related genes, including SNAI2, has been recently observed in mixed MPM (Casarsa et al., 2011), which is consistent with our RT-PCR
results. The importance of estrogen signaling in MPM is an open question, although recent studies indicate that ERβ levels have prognostic value in MPM (Pinton et al., 2009). The GADD45A gene, which has been observed to be up-regulated in the SIM2s depletion study analyzed by our model (Aleman et al., 2005), is a transcriptional target of ERβ (Paruthiyil et al., 2011). It is thus tempting to hypothesize that SIM2s expression may be connected to the estrogen signaling network. An important line of evidence comes directly from REx. The top three gene sets reported for both the SIM2s and MM studies are “metabolism of xenobiotics by cytochrome p450”, “androgen and estrogen metabolism”, and “arachidonic acid metabolism”. Cytochrome p450 (CYP) enzymes are known to mediate estrogen metabolism (Tsuihya et al., 2005). The genes in the xenobiotics and arachidonic acid metabolism gene sets significantly overlap, as per a one-tailed Fisher’s exact test ($p < 0.05$).

Together, our results and existing work indicate that SIM2s may have a relevant role in MPM, potentially via the EMT network and estrogen signaling.

### 3.3 Functionally Coherent Differential Expression Patterns

We computed for every module and submodule a group of top gene sets and genes, as described in Methods. We assessed the functional profile of each group of genes by testing for the overrepresentation of Gene Ontology (GO) (Ashburner et al., 2000) biological process terms.

Supplementary Figures S1 and S2 display the associations between enriched functional categories as described by gene sets and modules and submodules, respectively. Modules are enriched on a wide span of biological processes such as apoptosis (e.g., module 1), metabolism (e.g., module 31), neoplasia (e.g., module 38), respiration (e.g., module 13), toll-like receptor signaling (e.g., module 28), and transcription (e.g., module 8). There is also an overall trend for modules to focus either on metabolic gene sets or disease-related gene sets, although modules from one group typically include gene sets from the other group.

Next, we studied how the modules combine submodules. Supplementary Figure S3 displays a heatmap of the distribution of submodules within the modules. It shows that each module is primarily focussed on a small number of submodules, with some of the submodules being effectively used by several modules. It also shows that while some submodules are predominant in at least one module, other submodules act as module fine-tuners, not being highly probable in any module.

These results demonstrate that our latent variable model is able to extract meaningful patterns of differential co-expression of gene sets and map them to core subsets of the most differentially expressed genes in those gene sets.

### 3.4 Retrieval Performance

We evaluated REx quantitatively by using each comparison in turn as a query, and measuring how well related comparisons were retrieved using the NDCG based on the EFO. This complements the qualitative evaluation through case studies and experimental validation. To put our quantitative results into context, we also computed the NDCG for other retrieval approaches, namely a Term Frequency - Inverse Document Frequency (TF-IDF) model (Manning et al., 2008) with cosine similarity based on a count representation for the GSEA results as described in previous work (Caldas et al., 2009), a Spearman rank correlation approach based on the fold-change ratios of the expression data, a Pearson correlation approach using the inferred distributions over modules of the comparisons, our own earlier method (Caldas et al., 2009), and a random baseline.

The box plots of the NDCG results are shown in Figure 4. For succinctness, we show only the NDCG results of the best-performing combination of modules and submodules; the results for alternative number of modules and submodules are shown in Supplementary Figure S4. For the random baseline, we computed for each query comparison the median NDCG over 1000 random permutations of all other comparisons. In order to obtain a rigorous measure of the difference in performance between methods, we ran a two-tailed Wilcoxon signed-rank test over the NDCG values of every pair of methods, correcting for multiple hypothesis testing via a $q$-value threshold of $q < 0.05$. The random baseline performs significantly worse than all non-random approaches ($q < 0.05$). The difference between Pearson correlation and the remaining non-random approaches is also significant ($q < 0.05$), as is the difference between Spearman correlation and the remaining approaches ($q < 0.05$). To confirm whether this difference corresponds to worse or better performance, we repeated the same procedure but this time using a one-tailed Wilcoxon signed-rank test. Pearson correlation performed significantly worse than all other non-random approaches, while Spearman correlation performed significantly better. We then analyzed the magnitude of the difference in NDCG values between our method and Spearman correlation. The NDCG values obtained by REx are on average 99%
is analogous to the gene set collection in the Molecular Signature Database. Although we did not consider integration of multiple data types (e.g., Guan et al., 2010), our proposed information retrieval and meta-analysis framework provides a sound basis for that task. For instance, since our method is primarily based on the activation of gene sets, studies with different data types can readily be merged as long as the same collection of gene sets can be used, i.e. transcriptomics and proteomics data sets could be integrated and used for retrieval without major changes to the method.

There is a wide spectrum of practical applications for REx. For instance, implemented in repositories of gene expression data, the method could be used to complement existing knowledge-based approaches for study retrieval. When considering this scenario, where new studies are frequently added to a repository, the unsupervised learning algorithms employed by REx would benefit from the ability to perform online learning, which is an interesting and relevant area of future research. With such algorithms in place, the links between studies provided by REx could also serve as navigational aids in exploratory settings to guide users to relevant studies in very large repositories.

Overall, as we have showed in this paper, the relatively unexplored paradigm of data-driven information retrieval in transcriptomics data offers the possibility of obtaining novel biological findings based on existing data, and holds the potential to ultimately accelerate biomedical research in areas as diverse as drug repurposing or biomarker development.

5 FUNDING

This work was supported by the Finnish Funding Agency for Technology and Innovation [40101/07]; the Pattern Analysis, Statistical Modeling and Computational Learning Network of Excellence [ICT 216886]; the Portuguese Science and Technology Foundation [SFRH/BD/35974/2007 to JC]; the Academy of Finland [115372 to EK]; Finnish Doctoral Programme in Computational Sciences [doctoral fellowship to AF]; Finnish Center of Excellence in Adaptive Informatics Research [JC, AF, and SK] and the European Molecular Biology Laboratory (doctoral fellowship to NG).

6 ACKNOWLEDGMENTS

We thank Päivi Tuominen and Tiina Marjomaa for excellent technical assistance; Ele Holloway for advice on ArrayExpress data curation; and Leo Lahti for helpful comments.

REFERENCES


Supplementary Information: Data-Driven Information Retrieval in Heterogeneous Collections of Transcriptomics Data Links SIM2s to Malignant Pleural Mesothelioma

José Caldas 1,* Nils Gehlenborg 2,3,4,* Eeva Kettunen 5, Ali Faisal 1, Mikko Rönty 6, Andrew G. Nicholson 7, Sakari Knuutila 8, Alvis Brazma 2, Samuel Kaski 1,9,†

1 Aalto University School of Science, Department of Information and Computer Science, Helsinki Institute for Information Technology HIIT, Helsinki, Finland
2 Functional Genomics Group, European Bioinformatics Institute, Cambridge, United Kingdom
3 Graduate School of Life Sciences, University of Cambridge, Cambridge, United Kingdom
4 Current address: Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA
5 Health and Work Ability, Biological Mechanisms and Prevention of Work-Related Diseases, Finnish Institute of Occupational Health, Helsinki, Finland
6 HUSLAB, Department of Pathology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
7 Department of Histopathology, Royal Brompton Hospital, London, United Kingdom
8 Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
9 Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

SUPPLEMENTARY METHODS

Text S1: Tumor Specimens and RT-PCR

Tumor tissue specimens were obtained from ten malignant pleural mesothelioma (MPM) patients that were diagnosed with mesothelioma tumor at Royal Brompton and Harefield NHS Trust, United Kingdom. Of those, six were epithelial and four were biphasic MPMs. The diagnosis was confirmed using a standard panel of immunohistochemical markers including calretinin, cytokeratin 5/6, BerEP4, and CEA. The study protocol was approved by the Ethical Review Board of the Royal Brompton and Harefield Hospitals NHS Trust. The tumor content of each sample was verified by a pathologist at the Helsinki University Central Hospital. The characteristics of tumor patients are presented in Supplementary Table S3.

As a control we used a microscopically normal scraped pleural tissue lining of the lung of a 39-year-old, previously healthy male patient operated at Harefield Hospitals NHS Trust. The tumor content of each sample was approved by the Ethical Review Board of the Royal Brompton and Harefield NHS Trust, United Kingdom. Of those, six were epithelial and four were biphasic MPMs. The diagnosis was confirmed using a standard panel of immunohistochemical markers including calretinin, cytokeratin 5/6, BerEP4, and CEA. The study protocol was approved by the Ethical Review Board of the Royal Brompton and Harefield Hospitals NHS Trust. The tumor content of each sample was verified by a pathologist at the Helsinki University Central Hospital. The characteristics of tumor patients are presented in Supplementary Table S3.

As a control we used a microscopically normal scraped pleural tissue lining of the lung of a 39-year-old, previously healthy male patient operated at the Helsinki University Central Hospital for a non-neoplastic intrabronchial characteristics of tumor patients are presented in Supplementary Table S3. To whom correspondence should be addressed.

* These authors contributed equally to this work.
† To whom correspondence should be addressed.

Total RNA was extracted from fresh-frozen tissue specimens using mRNAeasy Mini Kit (Qiagen, Valencia, CA, USA). The eluted RNA was quantified by NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA) and the quality of RNA was assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). Total RNA was treated with Ambion® DNA-free™DNase (Applied Biosystems, Warrington, Cheshire, UK) according to the manufacturer’s protocol and was converted to cDNA using the High Capacity RNA-to-cDNA Kit (Applied Biosystems). Gene expression levels of MMP2, MMP3, MMP14, SNAI1, SNAI2, and MYOM2 were studied using the Applied Biosystems assays Hs01548727_m1, Hs00968305_m1, Hs00237119_m1, Hs00195591_m1, Hs00161904_m1, and Hs00187676_m1. Human PPIA (cyclophilin A) assay was used as an endogenous control. The PCR reactions of 20 μl consisted of 1× Taqman Gene Expression Master Mix and 1× Taqman assay for a gene of interest (Applied Biosystems). Gene expression levels were studied using 1× Power SYBR Green PCR Master Mix (Applied Biosystems) in a 20 μl PCR reaction consisting of 150 nM of each primer and 0.7 to 1.3 μl of cDNA. Moreover, SIM2s and SIM2i, and ACTB as a reference gene, were studied using 1× Power SYBR Green PCR Master Mix (Applied Biosystems) in a 20 μl PCR reaction consisting of 150 nM of each primer and 0.7 to 1.3 μl of cDNA. SIM2s and SIM2i share the first nine exons and the first part of the tenth exon. SIM2s has a specific part in the end of the tenth exon whereas exon 11 is exclusively used for SIM2i. Specific primers have been presented in Halvorsen et al. (Halvorsen et al., 2007) and were obtained from TIB MOLBIOL Syntheselabor GmbH (Berlin, Germany). PCR reactions were performed in Applied Biosystems 7500 Real-Time PCR System and analyzed using the ddCt method with SDS v1.4 software (Applied Biosystems). Supplementary Table S4 contains the final fold-change values for every patient and gene. Statistical significance of differential expression was assessed by performing a two-sided Student’s t-test on the final log-ratio values of each gene.
Text S2: Neutral Factors
ba_individual
ba_age
ba_replicated
ba_population
ba_familyhistory
ba_envhistory

Text S3: Control factors
0
0 cm away from the tumor boundary
0 d
0 days
0 Gy
0 h
0 hours
0 hours per day access
0 IU
0 IU_per_ml
0 m
0 M
0 mg_per_kg
0 mg_per_kg_per_day
0 mg/kg
0 mg/kg x 2 doses per day
0 mg/kg/day
0 mM
0 mol_per_L
0 ng/ml
0 ng/mL
0 nM
0 nmol
0 ppm
0 U/kg
0 U/ml
0 ug
0 ug_per_kg
0 ug_per_mL
0 ug/kg
0 uM
0 umol
0 umol_per_kg
0 umol/kg
aortic banding - sham
av-fistula - sham
control - 37 degree_C
control - albumin
control - BSA
control - ethanol
control - IL-1b
control - interferon-gamma
control - keratinocyte growth factor
control - unsynchronized
control - untreated
control - vector
control - vehicle
control diet
control for EGF
control for heregulin
control polyamide and dihydrotestosterone
control siRNA
control
empty vector
mock
mock infected
mock transfeected
myocardial infarction - sham
non-smoker
none
normal
normal 2
normal 9
normal contralateral cartilage
normal diet
normal donor
normal growth media (bone marrow)
normal growth media (C85)
normal terminal duct lobular unit
normal tissue from invasive ductal carcinoma patient
normal tissue from invasive lobular carcinoma patient
normal1
normal2
normal3
normal4
normal5
placebo
reference
SCA1 wild_type
SCA7 wild_type
sham
sham
sham denervation
sham fracture
sham injury
sham surgery
sham surgery - contralateral right hind limb
sham surgery - ipsilateral left hind limb
sham surgery cartilage
uninduced
uninfected
untreated
wild type
wild type
wild type SOD1 transgenic
wild type T cell receptor
wild type

Text S4: Gene Set Enrichment Analysis
In order to make data from different species commensurable and to apply GSEA, mouse and rat genes have to be mapped to human genes before any analysis can be performed. We obtained a precomputed mapping from array features to Ensembl Gene that is generated by Ensembl (Hubbard et al., 2009) and provided by ArrayExpress. Based on this mapping from array features to species-specific genes, human orthologs of mouse and rat genes were identified by querying an ortholog mapping provided by the Ensembl BioMart (Ensembl Release 56) (Vilella et al., 2009; Kasprzyk et al., 2004).
In a final step, all human Ensembl Gene identifiers were mapped to human gene symbols. In data sets where multiple array features map to the same gene, and therefore multiple expression profiles exist for a gene, they are collapsed into a single profile by computing the median expression profile across the corresponding features.
The validity of such approaches to deal with cross-species transcriptomics data in gene set enrichment analyses is supported by several previous publications (Sweet-Cordero et al., 2005; Bourquin et al., 2006), which successfully applied similar ortholog mapping approaches.
GSEA essentially consists of computing a running sum on the sorted list of differentially expressed genes; this running sum increases when a gene belongs to the gene set and decreases otherwise; the final statistic is
the maximum of this running sum. This procedure essentially amounts to computing a weighted Kolmogorov-Smirnov (KS) statistic. Significance is empirically assessed by randomly permuting the phenotype labels of the conditions and re-computing the KS statistic on the resulting sorted list of differentially expressed genes. We normalize the KS statistic for each gene set by dividing it by the mean of the respective randomly derived KS statistics; the 50 top scoring gene sets are selected according to this normalized score. Finally, the leading edge subset corresponds to the genes in the gene set that appear before the running sum achieves its maximum.

The use of a threshold that ignores significance values for extracting differentially expressed genes is heuristic, but has been previously successfully used in meta-analysis studies (Segal et al., 2004), as well as in our own previous work (Caldas et al., 2009). We have observed that selecting gene sets based on a standard q-value cut-off of q < 0.05 yields comparisons with a mean number of 0.87 gene sets (s.d. = 3.78), with 80.37% of the comparisons having zero gene sets, which is an excessively sparse encoding.

**Text S5: Unsupervised Learning Model**

**Generative Process** Let the observed data $X$ be a set of n GSEA comparisons. Each comparison $x_i$ is a tuple

$$ x_i = (x_{ij})_{j=1}^S, $$

where $x_{ij}$ contains the information about gene set $j$ and its corresponding leading edge subset, with $S$ being the total number of gene sets. In detail, each $x_{ij}$ is specified as

$$ x_{ij} = (x_{ij}^{(S)}, x_{ij}^{(G)}) , $$

where $x_{ij}^{(S)}$ is a binary variable that describes the activation level of gene set $j$ and $x_{ij}^{(G)}$ is a vector of binary variables, each of them asserting if a particular gene belongs to the leading edge subset of the gene set (the length of this vector is equal to the number of genes in gene set $j$). If a gene set is inactive, i.e., $x_{ij}^{(S)} = 0$, then the value of $x_{ij}^{(G)}$ is arbitrary and meaningless, as it is not taken into account by the model. The generative process is as follows:

1. for each submodule $t \in \{1, \ldots, T\}$
   a. for each gene set $j \in \{1, \ldots, S\}$
      1) $\phi_{t,j} \sim \text{Beta}(\alpha_{t,j})$
   b. for each gene $g \in \{1, \ldots, G\}$
      1) $\psi_{t,g} \sim \text{Beta}(\alpha_{t,g})$
2. for each module $m \in \{1, \ldots, M\}$
   a. $\eta_{m} \sim \text{Dirichlet}(\alpha_{m})$
3. for each GSEA comparison $i \in \{1, \ldots, n\}$
   a. $\theta_i \sim \text{Dirichlet}(\alpha_{0})$
   b. for each gene set $j \in \{1, \ldots, S\}$
      1) $u_{i,j} \sim \text{Discrete}(\theta_i)$
      2) $v_{i,j} \sim \text{Discrete}(\eta_{m_{i,j}})$
      3) $x_{ij}^{(S)} \sim \text{Bernoulli}(\phi_{v_{i,j},j})$
      4) if $x_{ij}^{(S)} = 1$ then for gene $g \in \sigma_j$
         1) $x_{ij}^{(G)} \sim \text{Bernoulli}(\psi_{v_{i,j},g})$

In the above process, $\sigma_j$ is the group of genes belonging to gene set $j$, and the keyword if means that the conditional distribution of the activation status of the genes in the gene set given that $x_{ij}^{(S)} = 0$ is uniform, and that the conditional distribution given $x_{ij}^{(S)} = 1$ follows the described Bernoulli distribution assumptions.

The main aim of stipulating Beta and Dirichlet priors is to take advantage of conjugacy properties that allow us to integrate out many of the model variables and use a collapsed Gibbs sampler.

**Gibbs Sampler** The joint probability of the latent and observed variables factorizes as follows:

$$ P(\theta, \eta, \phi, \psi, u, v, X) = P(\theta)P(\eta)P(\phi)P(\psi)P(u|\theta)P(v|u, \eta)P(X|v, \phi, \psi). $$

The aim of our inference engine is to obtain an approximate posterior distribution for the latent variables $u$ and $v$ given the observed data, as well as estimates of the variables $\theta$, $\eta$, $\phi$, and $\psi$. Using Bayes’ rule, this posterior distribution of all latent variables is given by

$$ P(\theta, \eta, \phi, \psi, u, v, X) = \frac{P(\theta)P(\eta)P(\phi)P(\psi)P(u|\theta)P(v|u, \eta)P(X|v, \phi, \psi)}{\int \cdots \int P(\theta)P(\eta)P(\phi)P(\psi)P(u|\theta)P(v|u, \eta)P(X|v, \phi, \psi) du dv d\psi d\theta d\eta}. $$

Our inference engine is based on a common approach known as the collapsed Gibbs sampler (Liu, 1994), which has been successfully used in several mixture models (Griffiths and Steyvers, 2004). Instead of directly approximating the joint posterior distribution of all the variables, the sampler approximates the posterior distribution

$$ P(u, v|X) = \int \cdots \int P(\theta, \eta, \phi, \psi, u, v, X) du dv d\psi d\theta d\eta,$$

where the variables $\theta$, $\eta$, $\phi$, and $\psi$ are collapsed, or integrated out. We will refer to the distribution of $u$ and $v$ given $X$ as the collapsed posterior distribution. The collapsed posterior distribution $P(u, v|X)$ is approximated by means of a Gibbs sampler. Finally, after the sampler has converged, a single sample is used to estimate the variables that were integrated out (Griffiths and Steyvers, 2004). The standard reason for using only one sample to estimate the variables is the existence of the well-known label switching problem (Jasra et al., 2005).

Due to the specific variable dependencies in our model, the collapsed posterior distribution is given by

$$ P(u, v|X) = \int \cdots \int P(\theta, \eta, \phi, \psi, u, v, |X) | du dv d\psi d\theta d\eta $$

$$ = \int P(\psi|\phi)P(X|\psi, \phi)dv d\psi $$

$$ = P(u|\theta)P(v|u)P(X|v, u). $$

The above integrals can be computed analytically because conjugate Dirichlet-multinomial and beta-binomial distributions were used. After some algebra, the integrals are as follows:

$$ P(u|\theta) = \left(\frac{\Gamma(M\alpha_{0})}{\Gamma(\alpha_{t})^{M}}\right)^{n} \prod_{i=1}^{n} \sum_{m=1}^{M} \frac{\Gamma(M\alpha_{t})}{\Gamma(\alpha_{m} + \epsilon_{t})} $$

$$ P(v|u, \eta) = \left(\frac{\Gamma(T\alpha_{m})}{\Gamma(\alpha_{m})^{T}}\right)^{M} \prod_{m=1}^{M} \sum_{t=1}^{T} \frac{\Gamma(T\alpha_{m})}{\Gamma(\alpha_{t} + \epsilon_{m})}. $$

Supplementary Information: Information Retrieval of Transcriptomics Data
\[ P(X | u, v) = \prod_{t=1}^{T} \left( \prod_{m=1}^{S} \frac{B(\alpha_u + c_{ts^u}, \alpha_v + c_{ts^v})}{B(\alpha_u, \alpha_v)} \right) \times \left( \prod_{g=1}^{G} \frac{B(\alpha_u + c_{tg^u}, \alpha_v + c_{tg^v})}{B(\alpha_u, \alpha_v)} \right) . \]

We use dot (.) notation for vector summation. In the above equations, \( n \) is the number of GSEA comparisons, \( M \) is the number of modules, \( T \) is the number of submodules, \( S \) is the number of gene sets, and \( G \) is the number of genes. The scalar hyperparameters are in turn designated as \( \alpha_u, \alpha_v, \alpha_g, \) and \( \alpha_p \). Finally, the above probabilities depend on \( u, v, \) and \( X \) only through specific statistics that we designate as \( c \). Concretely, \( c_{im} \) is the number of times that module \( m \) was chosen in GSEA comparison \( i \); \( c_{mnt} \) is the number of times that submodule \( t \) was chosen by the \( v \) variable, given that the corresponding \( u \) variable chose module \( m \); \( c_{tg^u} \) and \( c_{tg^v} \) are respectively the number of times that gene set \( s \) was active or inactive given that the corresponding variable \( v \) was assigned to submodule \( t \); finally, \( c_{tg^u} \) and \( c_{tg^v} \) represent the number of times that gene \( g \) was active or inactive given that the corresponding variable \( v \) was assigned to submodule \( t \) and the associated gene set was also active.

Succinctly, our Gibbs sampler approximates the collapsed posterior distribution \( P(u, v | X) \) by iteratively sampling from the joint posterior distribution of \( u_{(i-1)} \) and \( v_{(i-1)} \) conditional on all other variables, i.e., by iteratively sampling from \( P(u_{(i)}, v_{(i)} | u_{(1)}, v_{(1)}, \ldots, u_{(i-1)}, v_{(i-1)}, X) \). We use the notation \( u_{(i)} \) to refer to the set of \( u \) variables except the variable \( u_{(i)} \); the same notation applies to the \( v \) variables.

The sampling equations can be derived in a straightforward manner from the equations in the previous section. Omitting the derivations, they are as follows:

\[ P(w_{im} = m, v_{is} = t | u_{(i-1)}, v_{(i-1)}, X) \propto \frac{(\alpha_u + c_{im})^{\alpha_u} (\alpha_v + c_{im})^{\alpha_v}}{\Gamma(\alpha_u + \alpha_v)} \times \left( \left( \prod_{g \in \sigma(s)} \frac{(\alpha_u + c_{tg^u})^{x_{sG}^u} (\alpha_v + c_{tg^v})^{1-x_{sG}^v}}{2^{\alpha_u + c_{tg^u}} + 2^{\alpha_v + c_{tg^v}}} \right) \right) . \]

Regarding the \( c \) variables, we use the minus (−) symbol to indicate that a certain element in the data should not be taken into account when computing those variables. For instance, \( c_{im}^{−} \) indicates the number of times module \( m \) was chosen in GSEA comparison \( i \), without considering the module chosen for gene set \( s \).

After the Gibbs sampler convergence period, we use a single sample to derive estimates for \( \theta, \eta, \phi, \) and \( \psi \), based on their predictive distributions over new observations (Griffiths and Steyvers, 2004). The estimates are the following (we omit the derivations):

\[ \theta_{im} = \frac{\alpha_u + c_{im}}{\alpha_u + \alpha_v + c_{im}}. \]
\[ \eta_{im} = \frac{\alpha_v + c_{im}}{\alpha_u + \alpha_v + c_{im}}. \]
\[ \phi_{tg} = \frac{\alpha_u + c_{tg^u}}{\alpha_u + c_{tg^u} + \alpha_v + c_{tg^v}}. \]
\[ \psi_{tg} = \frac{\alpha_v + c_{tg^v}}{\alpha_u + c_{tg^u} + \alpha_v + c_{tg^v}}. \]

We ran the Gibbs sampler for 2000 iterations, setting all hyperparameters to 0.1, with the number of modules and submodules varying between five and 45 modules and between 30 and 70 submodules. We obtained point estimates using the last obtained sample.

The low hyperparameter values correspond to non-informative priors. In this context, the use of symmetric prior distributions facilitates the sampling equations and implementation, and does not have an impact on the sampling process due to the use of low hyperparameter values.

In recent studies (Blei et al., 2010), hyperparameter sampling schemes based on Metropolis-Hastings steps have been proposed, which is an interesting possibility for future work.

Relevance measure In order to obtain a measure of relevance between a GSEA comparison \( j \) and a query GSEA comparison \( i \), we compute an approximation to the expected log-probability that the observed data in query comparison \( i \) is generated via the model parameters associated with comparison \( j \). This approach has been used before in the context of natural language processing (Bunton et al., 2004; Steyvers and Griffiths, 2007).

The expectation relating query comparison \( i \) to comparison \( j \) is defined as:

\[ \log P_j(x_i | \theta, \eta, \phi, \psi, u, v, X) \]

where \( \log P_j(x_i | \theta, \eta, \phi, \psi, u, v) \) is the log-probability of the observed data in query comparison \( i \), assuming that it has the parameters of comparison \( j \), i.e., assuming that \( \theta = \theta_j \). This log-probability is given by:

\[ \log P_j(x_i | \theta, \eta, \phi, \psi, u, v) = \sum_{s=1}^{S} \log \sum_{m=1}^{M} \sum_{t=1}^{T} \theta_{jm} \eta_{im} \phi_{tg} \psi_{tg} \left( \frac{\alpha_u + c_{tg^u}}{\alpha_u + c_{tg^u} + \alpha_v + c_{tg^v}} \right)^{x_{sG}^u} \left( \frac{\alpha_v + c_{tg^v}}{\alpha_u + c_{tg^u} + \alpha_v + c_{tg^v}} \right)^{1-x_{sG}^v} . \]

We approximate the expectation by collecting a number of samples after our context, we discarded the first 1500 iterations of the sampler as the burn-in period; we then averaged the relevance measure between any two samples.

Text S6: Retrieval Evaluation
To evaluate REs quantitatively we used a collection of 219 evaluable comparisons, i.e., comparisons in which one of the phenotypes is a control. For those comparisons, we were able to map the non-control phenotype to
the Experimental Factor Ontology (EFO) (Malone et al., 2010). The EFO represents the relationships between the terms that are used as experimental factor values to describe the biological conditions investigated in studies contained in ArrayExpress.

For the purpose of evaluating the retrieval performance, we define the relevance between two experimental factor values as the fraction of overlap between the two corresponding (shortest) EFO paths. Using this approach, the relevance between two comparisons is non-binary, which precluded us from using classical information retrieval performance measures such as average precision (Manning et al., 2008) to measure performance. We instead used the normalized discounted cumulative gain (NDCG) measure (Jarvelin and Kekäläinen, 2002; Manning et al., 2008), which effectively handles non-binary relevance scores.

Similar evaluation methodologies have been described, for instance by Hibbs et al. (2007), who employed the Gene Ontology (Ashburner et al., 2000) to cross-validate a gene-centric search engine for expression data, and Hu and Agarwal (2009), who used the Medical Subject Headings (MeSH) hierarchy to evaluate the quality of the disease-disease connections identified by a meta-analysis approach.

**Modified Jaccard Coefficient** The classic Jaccard coefficient \( J \) is a distance measure used to determine the similarity between two sets \( Q \) and \( R \) as:

\[
J(Q, R) = \frac{|Q \cap R|}{|Q \cup R|}
\]

Applied to the EFO, the sets \( Q \) and \( R \) are defined as the ontology terms on the shortest paths between the root and term \( q \) and the root and term \( r \). The modified Jaccard coefficient \( J' \) used here is defined as:

\[
J'(Q, R) = \begin{cases} 1, & \text{if } r \text{ is a child of } q; \\ J(Q, R), & \text{otherwise}. \end{cases}
\]

In the context of the retrieval method described above, the modified Jaccard coefficient is used to determine the graded relevance \( \text{rel}(q, r) = J'(Q, R) \) of a retrieved comparison mapped to term \( r \), when the query is mapped to term \( q \). The relevance will be at a maximum of 1 when both the retrieved comparison and the query map to the same ontology term or when the retrieved comparison maps to a child of the query term. Accordingly, this relevance measure is not symmetric and can yield different results when \( q \) and \( r \) are exchanged.

**Normalized Discounted Cumulative Gain** Since the modified Jaccard coefficient provides a graded relevance between 0 and 1, precision and recall measures cannot be applied to evaluate the performance of our method. “Cumulative gain” evaluation methods (Jarvelin and Kekäläinen, 2002) are a family of methods that are based on graded relevance judgements. Applied to the retrieval of comparisons, these methods measure how much the investigator gains when a comparison with a particular relevance is found at a particular rank in the result list for a query.

The Discounted Cumulative Gain for a ranked list of graded relevance judgements \( \text{rel}(q, r_i) \), with \( i = 1, \ldots, C \) where \( C \) is the number of interpretable comparisons, is defined as

\[
DCG_{p} = \sum_{i=1}^{p} \frac{2^{\text{rel}(q, r_i)} - 1}{\log_{2}(i+1)}
\]

with \( p \) being the position in where ranked list is cut. To evaluate our method, \( p \) was set to \( C \), which means that the complete ranked list was taken into account. The interpretation of the DCG is that retrieved comparisons of equal relevance become less valuable, or provide less gain, the farther away from the top of the list they occur.

In order to compare the DCG across retrieval methods, it has to be normalized. The Normalized Discounted Cumulative Gain (NDCG) is the DCG relative to the best possible or Ideal Discounted Cumulative Gain (IDCG) and is defined as

\[
NDCG_{p} = \frac{DCG_{p}}{\text{IDCG}_{p}}.
\]

We obtained the IDCG by computing the DCG for the list of comparisons ranked by their relevance according to the gold standard, here expressed by \( \text{rel}(q, r) \).

**SUPPLEMENTARY RESULTS**

**Text S7: Case Study 1 - Benign Nevi, Malignant Melanoma, and Cardiomyopathies**

When querying the database with the comparisons *benign nevi vs. normal* (Supplementary Table S5) and *malignant melanoma vs. normal* (Supplementary Table S6), we observed that the top 25 results in both cases contain a range of different cardiomyopathies (viral, idiopathic, familial, ischemic, post-partum, hypertrophic).

Furthermore, the two comparisons retrieve each other, indicating a link between the two conditions, and also several cancer-related conditions, such as transfection with Ewing sarcoma family fusion gene, breast cancer, and carcinoma in situ lesion.

The link between benign nevi and malignant melanoma is well established and has been studied extensively (Talantov et al., 2005). However, the link between these conditions and (cardio)myopathies has been reported only once before in a recent study by Hu and Agarwal (2009), where the authors used gene expression data in an approach conceptually similar to the Connectivity Map (Lamb et al., 2006) to identify links between human diseases. In their paper, Hu et al. suggest that the link between benign nevi/malignant melanoma and cardiomyopathies is an inverse relationship that was found due to the cell growth properties of the benign nevi/malignant melanoma and the muscular weakness or wasting properties of the cardiomyopathies.

REx provides further evidence that the relationship is indeed inverse, as many of the top 25 gene sets that are affected by the conditions are upregulated in benign nevi/malignant melanoma and downregulated in cardiomyopathies (data not shown). For example, the most relevant gene set for benign nevi is the phosphoinositide 3-kinase (PI-3-K) pathway, which, among other things, is involved in cell survival and cell proliferation (Engelman, 2009). The second most relevant gene set for the benign nevi comparison is the Ras pathway, which is upregulated in this comparison. The Ras pathway activates the PI 3-K pathway, resulting in the inhibition of apoptosis.

In the malignant melanoma comparison, the Ras pathway is also among the top 25 gene sets and upregulated. In contrast to the benign nevi and malignant melanoma comparisons, the Ras pathway is among the top 25 gene sets and downregulated in almost all cardiomyopathies.

This case study is an example of how the information provided by REx can be used to both identify and interpret links between seemingly unrelated conditions.

**Text S8: Case Study 2 - Pancreatic Ductal Adenocarcinoma, Insulin and Inflammation**

REx found a relationship between *pancreatic cancer vs. normal* and insulin-related conditions as well as obesity. The top 25 comparisons are shown in Supplementary Table S7. The pancreatic cancer in this comparison is a pancreatic ductal adenocarcinoma (PDAC).
The most relevant result when querying with PDAC is a preadipocyte cell line from mouse, in which IRS4 (insulin receptor substrate 4) has been knocked out, which is naturally expected to have an effect on insulin signaling. Also highly ranked is a knockout of IRS1 (insulin receptor substrate 1) from the same original study.

The second most relevant result when querying with PDAC is a comparison of mouse adipocytes treated with growth hormone, which has been found to stimulate the expression of ATF3 (Activating Transcription Factor 3) (Huo et al., 2006). ATF3 is known to have a role in glucose homeostasis (Allen-Jennings et al., 2001). Treatment with growth hormone for 48 hours (as in the retrieved comparison) has been found to regulate an immune response that potentially affects insulin signaling (Huo et al., 2006).

Another result that has been found to have high relevance to the query PDAC is a comparison from a study investigating the infection of HeLa cells with Coxsackie B3 virus was found as another highly relevant result related to insulin and the PDAC query. Coxsackie B viruses have been suspected to be an environmental trigger for neuropathies (Giambonini-Brugnoli et al., 2005; Gabriel et al., 2006). A further link is provided by a recent study in which the PMP22 gene (peripheral myelin protein 22) gene.

Further highly relevant results reveal, for instance, a comparison involving a human HepG2 cell line, which overexpresses D374Y-PCSK9, a mutated allele of PCSK9 (proprotein convertase subtilisin/kexin type 9), which is a known key regulator of serum cholesterol. Moreover, D374Y-PCSK9 was suspected to downregulate certain stress-response genes and inflammation pathways (Ranheim et al., 2008). A further link is provided by a study investigating the infection of HeLa cells with Coxsackie B3 virus was found as another highly relevant result related to insulin and the PDAC query. Coxsackie B viruses have been suspected to be an environmental trigger for insulin dependent diabetes mellitus type 1 (TID) since the early 1980s (Peng and Hagopian, 2006).

A comparison between wild type and glycol kinase (GK) knock-out mice is also among the top most relevant results. The authors of the study from which this comparison originates found that, among other things, the lack of GK affects the expression of several genes that are involved in insulin signaling and insulin resistance (Rahib et al., 2007). A comparison from a study investigating the infection of HeLa cells with Coxsackie B3 virus was found as another highly relevant result related to insulin and the PDAC query. Coxsackie B viruses have been suspected to be an environmental trigger for insulin dependent diabetes mellitus type 1 (TID) since the early 1980s (Peng and Hagopian, 2006).

Another result that has been found to have high relevance to the PDAC comparison is a comparison between adipocytes from obese and non-obese human subjects. The authors of the corresponding study (Lee et al., 2005) found that a large number of genes associated with inflammation and immune response are upregulated in obese subjects. This link could be due to similar processes as the ones found in the growth hormone study described above. Furthermore, this link could explain why another highly ranked result is a comparison in which the effects of the anti-inflammatory agent "Quercetin" (Stewart et al., 2008) was studied.

A highly relevant, but unexpected, result is a comparison of wild type B cells and B cells from mice carrying a point mutation ("trembler") in the PMP22 (peripheral myelin protein 22) gene. PMP22 is involved in demyelination and dysmyelinating peripheral neuropathies (Giambonini-Brugnoli et al., 2005; Gabriel et al., 2000), which are associated with diseases such as Charcot-Marie-Tooth disease Type 1A (CMT1A) or diabetes mellitus (Chahin et al., 2007). CMT1A is usually caused by a partial duplication of the PMP22 gene (Meyer zu Hörste et al., 2006); but recently it has also been found that the PMP22 region is amplified in PDACs (Funel et al., 2009), which establishes a potential link to the query comparison. In previous studies it has been shown that the gene is actually expressed in these cancers (Li et al., 2005).

In this case study we were able to link 11 of the top 14 retrieved comparisons either directly to the query comparison (pancreatic cancer) or to related conditions (insulin signaling, diabetes mellitus, inflammation). The 11 comparisons came from 10 different studies in our collection (E-MEXP-950, E-GEOD-2556, E-GEOD-2120, E-GEOD-7146, E-MEXP-1235, E-GEOD-2508, E-GEOD-4656, E-GEOD-4262, E-GEOD-697, E-GEOD-1947), which indicates that the method indeed can identify links across different studies. Interestingly, two out of three comparisons for which links could not be found are from studies that appear to have never been published in a journal. This may indicate problems with the data or the experimental setup, which gives reason to believe that these comparisons might be false positive hits.

Text S9: Case Study 3 - Glioblastoma

When we queried the collection of comparisons with glioblastoma vs. normal, the twelve most relevant results all involve samples from nervous tissue, either from the brain or elsewhere in the central nervous system. Among the top 25 most relevant results, which are shown in Supplementary Table S8, a total of 19 comparisons involve nervous tissue. The comparisons do not show a clear pattern, as they include cancers, induced brain and spinal cord injuries, genetic modifications, treatments with various chemicals and brain disorders such as bipolar disorder and Alzheimer’s disease.

The results of this query illustrates that the retrieval of comparisons can be based on tissue specificity, rather than on conditions such as diseases or treatments. It is important to point out that general tissue-specific expression patterns most likely are not the cause for the similarity observed between these comparisons, as the differential analysis is designed to remove these effects.

Text S10: Sensitivity Analysis and Model Robustness

We assessed the robustness of the query results for evaluable queries with respect to variation in the number of modules and submodules. The model that we considered for biological and quantitative analysis includes 45 modules and 60 submodules; we refer to this model as the "final" model, and to all other models as "alternative". We computed the Spearman correlation coefficient between the relevance-sorted list of GSEA comparisons for a given query in any alternative model and the corresponding list in the final model. The box plot of the resulting correlation estimates is shown in Figure 5(a). While several correlation estimates are low, the majority is high, with the median estimate being over 0.7. We tested how many of the correlation estimates are significant, correcting for multiple hypothesis testing via Bonferroni correction (n = 8541). Slightly above 90% of the correlation estimates are significant (p < 0.01/n). We therefore conclude that the query results in a model with a reasonable number of modules and submodules are typically similar to the corresponding query results in the final model with 45 modules and 60 submodules. However, despite the correlation
between query results, models with a lower number of modules and submodules have a worse information retrieval performance than the final model, as shown in Figure S4.

We followed a similar correlation-based approach to test if varying the number of modules and submodules has an impact on the comparison-to-gene-set probability distributions. As shown in Equation (1) of the main manuscript, the comparison-to-gene-set probabilities involve summing out comparison-to-module and module-to-submodule distributions. Therefore, the comparison-to-gene-set probabilities effectively take into account most of the structures inferred by the model. For every alternative model and comparison, we computed the Spearman rank correlation coefficient between the comparison-to-gene-set probabilities and the corresponding comparison-to-gene-set probabilities in the final model. The corresponding box plot of correlation estimates is presented in Figure 5(b). It can be seen that most correlation estimates are high, with the median estimate being above 0.7. Using a Bonferroni correction (n = 270192), again more than 99% of the comparison-to-gene-set distributions are significantly correlated with the corresponding distributions in the final model (p < 0.01/n).

The above results suggest that the inferred model structures and query results are robust to variations in the number of modules and submodules. The results were obtained using a random Gibbs sampler initialization in which the random seed was different for every model. This suggests that the model is also robust with respect to initialization procedures.

Finally, we set all hyperparameters to 0.1 for two reasons: First, as discussed earlier, by making use of conjugate distributions, we are able to integrate out model variables and use a collapsed Gibbs sampler, which is known to be an efficient procedure for inference and estimation in latent variable models (Griffiths and Steyvers, 2004); second, given the large amount of data used by our model and the low hyperparameter values, the latent variable assignment of a data point during the Gibbs sampling process depends entirely on the assignments of the remaining data points, with the contribution stemming from the hyperparameter values being negligible. However, an interesting possibility for future work is to sample the hyperparameters during the Gibbs sampling process, as has been suggested in recent work (Blei et al., 2010).

REFERENCES


SUPPLEMENTARY FIGURES
Fig. S1. Significant gene sets and GO BP terms over modules. For ease of illustration, we only included the 15% most frequent gene sets and Gene Ontology (GO) Biological Process (BP) terms. A gray box indicates membership of an enriched term in a module. We sorted the gene sets/GO terms and modules according to dendrograms obtained by running hierarchical clustering on both rows and columns of the matrix, using complete linkage and Manhattan distances.
Fig. S2. Significant gene sets and GO BP terms over submodules. For ease of illustration, we only included the 15% most frequent gene sets and Gene Ontology (GO) Biological Process (BP) terms. A gray box indicates membership of an enriched term in a submodule. We sorted the gene sets/GO terms and submodules according to dendrograms obtained by running hierarchical clustering on both rows and columns of the matrix, using complete linkage and Manhattan distances.
Fig. S3. Heatmap of distributions from modules to submodules.
Fig. S4. Median NDCG scores for various alternatives in the number of modules and submodules. The highest NDCG was found for 45 modules and 60 submodules.
Fig. S5. Box plots of Spearman correlation coefficients. (a) Correlation estimates between query results in alternative models and the corresponding query results in the final model. (b) Correlation estimates between comparison-to-gene-set probabilities in alternative models and the corresponding comparison-to-gene-set probabilities in the final model.
SUPPLEMENTARY TABLES
Table S1. The multiple components of a general meta-analysis and information retrieval framework, and existing approaches to each of those components. DE stands for “differential expression”. Regarding the study decomposition component, “one vs. all” means comparing each condition against the mean of all other conditions, while “all vs. all” means comparing every pair of conditions. Regarding the relevance measure, “correlation” means any type of parametric or non-parametric (rank-based) correlation measure.
<table>
<thead>
<tr>
<th>Q</th>
<th>malignant pleural mesothelioma vs normal in <em>Homo sapiens</em> (pleura)</th>
<th>E-GEOD-2549</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>high potassium vs control in <em>Homo sapiens</em></td>
<td>E-GEOD-2883</td>
</tr>
<tr>
<td>2</td>
<td>thapsigargin vs control in <em>Homo sapiens</em></td>
<td>E-GEOD-2883</td>
</tr>
<tr>
<td>3</td>
<td>SIM2s vs control in <em>Homo sapiens</em> (18 h)</td>
<td>E-MEXP-101</td>
</tr>
<tr>
<td>4</td>
<td>hydrogen peroxide vs control in <em>Homo sapiens</em> (1 h)</td>
<td>E-GEOD-5339</td>
</tr>
<tr>
<td>5</td>
<td>0.01 ug per kg per day vs 0 mg per kg per day in <em>Rattus norvegicus</em> (17a-ethynylestradiol)</td>
<td>E-TABM-12</td>
</tr>
<tr>
<td>6</td>
<td>POR null vs wild type in <em>Mus musculus</em> (none &amp; ileum)</td>
<td>E-GEOD-4262</td>
</tr>
<tr>
<td>7</td>
<td>10 ug per kg vs 0 ug per kg in <em>Rattus norvegicus</em> (17a-ethynylestradiol &amp; 8 h)</td>
<td>E-MEXP-999</td>
</tr>
<tr>
<td>8</td>
<td>idiopathic dilated cardiomyopathy vs normal in <em>Homo sapiens</em></td>
<td>E-GEOD-1145</td>
</tr>
<tr>
<td>9</td>
<td>calorie-restricted diet vs normal diet in <em>Mus musculus</em> (4 months)</td>
<td>E-GEOD-4786</td>
</tr>
<tr>
<td>10</td>
<td>non-progressive HIV infection vs uninfected in <em>Homo sapiens</em> (CD4+ T cell)</td>
<td>E-GEOD-6740</td>
</tr>
<tr>
<td>11</td>
<td>K14deltaNB-cateninER vs wild type in <em>Mus musculus</em> (1 d)</td>
<td>E-GEOD-1579</td>
</tr>
<tr>
<td>12</td>
<td>endometriosis vs normal in <em>Homo sapiens</em> (mid secretory phase)</td>
<td>E-GEOD-6364</td>
</tr>
<tr>
<td>13</td>
<td>Aldh5a1-l vs wild type in <em>Mus musculus</em> (hippocampus)</td>
<td>E-GEOD-2866</td>
</tr>
<tr>
<td>14</td>
<td>dexamethasone vs none in <em>Homo sapiens</em> (4 h)</td>
<td>E-GEOD-3040</td>
</tr>
<tr>
<td>15</td>
<td>hydrogen peroxide vs control in <em>Homo sapiens</em> (12 h)</td>
<td>E-GEOD-5339</td>
</tr>
<tr>
<td>16</td>
<td>0.5 h vs 0 h in <em>Homo sapiens</em> (none &amp; uninfected)</td>
<td>E-GEOD-697</td>
</tr>
<tr>
<td>17</td>
<td>ketogenic diet vs control diet in <em>Rattus norvegicus</em></td>
<td>E-GEOD-1155</td>
</tr>
<tr>
<td>18</td>
<td>hydrogen peroxide vs control in <em>Homo sapiens</em> (24 h)</td>
<td>E-GEOD-5339</td>
</tr>
<tr>
<td>19</td>
<td>vanadium pentoxide vs control in <em>Homo sapiens</em> (12 h)</td>
<td>E-GEOD-5339</td>
</tr>
<tr>
<td>20</td>
<td>superficial transitional cell carcinoma with surrounding carcinoma in situ lesion vs normal in <em>Homo sapiens</em> (bladder)</td>
<td>E-GEOD-3167</td>
</tr>
<tr>
<td>21</td>
<td>retinoic acid vs none in <em>Mus musculus</em> (6 h)</td>
<td>E-GEOD-1588</td>
</tr>
<tr>
<td>22</td>
<td>RP1 knockout vs wild type in <em>Mus musculus</em> (7 d)</td>
<td>E-GEOD-128</td>
</tr>
<tr>
<td>23</td>
<td>HIV-1 infected vs normal in <em>Homo sapiens</em> (none)</td>
<td>E-GEOD-2504</td>
</tr>
<tr>
<td>24</td>
<td>Yersinia enterocolitica WA(pTTS, pP60) (control) vs uninfected in <em>Mus musculus</em> (interferon-gamma &amp; BALB/c)</td>
<td>E-GEOD-2973</td>
</tr>
<tr>
<td>25</td>
<td>BALB/c SCID vs wild type in <em>Mus musculus</em> (Nippostrongylus brasiliensis &amp; 8 d)</td>
<td>E-GEOD-3414</td>
</tr>
</tbody>
</table>

Table S2. Query results for an MPM comparison (query Q). Text in parentheses after the name of the species is the context of the corresponding comparisons.
<table>
<thead>
<tr>
<th>Case</th>
<th>Histological type</th>
<th>Sex</th>
<th>Age</th>
<th>Asbestos exposure</th>
<th>Smoking</th>
<th>Tumor content %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>biphasic</td>
<td>M</td>
<td>63</td>
<td>Yes</td>
<td>None</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>biphasic</td>
<td>M</td>
<td>64</td>
<td>No</td>
<td>None</td>
<td>70</td>
</tr>
<tr>
<td>3</td>
<td>epithelial</td>
<td>M</td>
<td>67</td>
<td>No</td>
<td>Ex</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>epithelial</td>
<td>M</td>
<td>57</td>
<td>No</td>
<td>Ex</td>
<td>70</td>
</tr>
<tr>
<td>5</td>
<td>epithelial</td>
<td>M</td>
<td>76</td>
<td>Yes</td>
<td>Ex</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>6</td>
<td>epithelial</td>
<td>F</td>
<td>56</td>
<td>Yes</td>
<td>None</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>7</td>
<td>epithelial</td>
<td>M</td>
<td>56</td>
<td>Yes</td>
<td>Ex</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>epithelial</td>
<td>M</td>
<td>71</td>
<td>Yes</td>
<td>Ex</td>
<td>30</td>
</tr>
<tr>
<td>9</td>
<td>biphasic</td>
<td>M</td>
<td>68</td>
<td>Yes</td>
<td>None</td>
<td>&gt; 50</td>
</tr>
<tr>
<td>10</td>
<td>biphasic</td>
<td>M</td>
<td>57</td>
<td>Yes</td>
<td>None</td>
<td>70</td>
</tr>
</tbody>
</table>

Table S3. Clinical data for malignant pleural mesothelioma patients, including histological type, sex, age, asbestos exposure, smoking status, and sample tumor content.
Table S4. Final fold-change values, obtained by RT-PCR, for every patient and gene. Fold-change values were not computed for the genes MMP3 and SIM2l because these were not expressed in the pleural control.

<table>
<thead>
<tr>
<th>Case</th>
<th>MMP2</th>
<th>MMP14</th>
<th>SNA11</th>
<th>SNAI2/SLUG</th>
<th>MYOM2</th>
<th>SIM2s</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.26</td>
<td>1.62</td>
<td>0.03</td>
<td>1.74</td>
<td>5.06</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.59</td>
<td>0.9</td>
<td>0.47</td>
<td>3</td>
<td>0.2</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>4.14</td>
<td>2.26</td>
<td>0.43</td>
<td>5.25</td>
<td>1.7</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>0.45</td>
<td>1.09</td>
<td>0.05</td>
<td>0.6</td>
<td>0.38</td>
<td>0.02</td>
</tr>
<tr>
<td>5</td>
<td>0.4</td>
<td>0.33</td>
<td>0.05</td>
<td>1.05</td>
<td>59.31</td>
<td>0.02</td>
</tr>
<tr>
<td>6</td>
<td>3.82</td>
<td>4.81</td>
<td>0.46</td>
<td>7.77</td>
<td>3.15</td>
<td>0.03</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.34</td>
<td>0.73</td>
<td>5.88</td>
<td>446.85</td>
<td>0.06</td>
</tr>
<tr>
<td>8</td>
<td>2.04</td>
<td>1.8</td>
<td>4.32</td>
<td>7.36</td>
<td>5.88</td>
<td>0.12</td>
</tr>
<tr>
<td>9</td>
<td>3.48</td>
<td>2.84</td>
<td>2.29</td>
<td>9.22</td>
<td>3.65</td>
<td>0.18</td>
</tr>
<tr>
<td>10</td>
<td>1.95</td>
<td>3.64</td>
<td>2.3</td>
<td>13.74</td>
<td>1.25</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Condition</td>
<td>Species</td>
<td>Accession Number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------</td>
<td>-----------------------</td>
<td>------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>viral cardiomyopathy</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>polycystic ovary syndrome</td>
<td>Homo sapiens</td>
<td>E-GEOD-1615</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>myelodysplastic syndrome</td>
<td>Homo sapiens</td>
<td>E-GEOD-2779</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>idiopathic cardiomyopathy</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>familial cardiomyopathy</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>EWS-FL1 transfected</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>ischemic cardiomyopathy</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Parkinon's disease</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>RAD001 vs placebo in Mus musculus</td>
<td>(wild type &amp; 48 h)</td>
<td>E-GEOD-1413</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>SR-A mutant vs wild type in Mus musculus</td>
<td>(bilateral olfactory bulbectomy &amp; 8 h)</td>
<td>E-GEOD-3455</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>diabetes mellitus</td>
<td>Rattus norvegicus</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>post-partum cardiomyopathy</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Rs1h null</td>
<td>Mus musculus</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>hypertrophic cardiomyopathy</td>
<td>Homo sapiens</td>
<td>E-GEOD-1145</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>carcinoma in situ lesion</td>
<td>Homo sapiens</td>
<td>E-GEOD-3167</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1d vs 0 in Homo sapiens</td>
<td>(female)</td>
<td>E-GEOD-1295</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>ochratoxin</td>
<td>Rattus norvegicus</td>
<td>E-GEOD-2852</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>malignant melanoma</td>
<td>Homo sapiens</td>
<td>E-GEOD-3189</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>non steroidal anti-inflammatory drugs</td>
<td>Homo sapiens</td>
<td>E-GEOD-7669</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>coxsackievirus B3</td>
<td>Homo sapiens</td>
<td>E-GEOD-697</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>valproic acid</td>
<td>Homo sapiens</td>
<td>E-GEOD-1615</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>12h vs 0 in Homo sapiens</td>
<td>(control)</td>
<td>E-GEOD-3183</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Beta4 nAChR subunit null</td>
<td>Mus musculus</td>
<td>E-GEOD-6614</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>dexamethasone</td>
<td>Homo sapiens</td>
<td>E-GEOD-1815</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Dysf-/- vs wild type in Mus musculus</td>
<td>(left ventricular myocardium)</td>
<td>E-GEOD-2507</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S5. Query results for a benign nevi comparison (query Q). Text in parentheses after the name of the species is the context of the corresponding comparisons. The third column contains the ArrayExpress accession number of the source data set.
Table S6. Query results for a malignant melanoma comparison (query Q). Text in parentheses after the name of the species is the context of the corresponding comparisons.
<table>
<thead>
<tr>
<th>Q</th>
<th>pancreatic cancer vs normal in <em>Homo sapiens</em></th>
<th>E-MEXP-950</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>IRS-4 vs wild type in <em>Mus musculus</em></td>
<td>E-GEO-D2556</td>
</tr>
<tr>
<td>2</td>
<td>growth hormone vs control in <em>Mus musculus</em> (48 h)</td>
<td>E-GEO-D2120</td>
</tr>
<tr>
<td>3</td>
<td>insulin vs none in <em>Homo sapiens</em></td>
<td>E-GEO-D7146</td>
</tr>
<tr>
<td>4</td>
<td>D374Y-PCSK9 vs wild type in <em>Homo sapiens</em></td>
<td>E-MEXP-1235</td>
</tr>
<tr>
<td>5</td>
<td>obesity vs normal in <em>Homo sapiens</em> (male)</td>
<td>E-GEO-D2508</td>
</tr>
<tr>
<td>6</td>
<td>severe malarial anaemia vs normal in <em>Homo sapiens</em></td>
<td>E-GEO-D1124</td>
</tr>
<tr>
<td>7</td>
<td>partial paw denervation vs sham denervation in <em>Rattus norvegicus</em> (3)</td>
<td>E-GEO-D2874</td>
</tr>
<tr>
<td>8</td>
<td>Gyk knockout vs wild type in <em>Mus musculus</em></td>
<td>E-GEO-D4656</td>
</tr>
<tr>
<td>9</td>
<td>quercetin vs none in <em>Mus musculus</em> (POR null &amp; jejunum)</td>
<td>E-GEO-D4262</td>
</tr>
<tr>
<td>10</td>
<td>U0126 vs none in <em>Homo sapiens</em> (coxackievirus B3 &amp; 9 h)</td>
<td>E-GEO-D697</td>
</tr>
<tr>
<td>11</td>
<td>E2F2-O vs wild type in <em>Mus musculus</em> (48 h)</td>
<td>E-MEXP-1413</td>
</tr>
<tr>
<td>12</td>
<td>Trembler vs wild type in <em>Mus musculus</em> (B cell &amp; P4)</td>
<td>E-GEO-D1947</td>
</tr>
<tr>
<td>13</td>
<td>IRS-1 vs wild type in <em>Mus musculus</em></td>
<td>E-GEO-D2556</td>
</tr>
<tr>
<td>14</td>
<td>24 h vs 0 h in <em>Homo sapiens</em></td>
<td>E-MEXP-1194</td>
</tr>
<tr>
<td>15</td>
<td>Cor1 -/-/Cor2 -/- vs wild type in <em>Mus musculus</em> (dinitrofluorobenzene)</td>
<td>E-GEO-D7694</td>
</tr>
<tr>
<td>16</td>
<td>non-progressive HIV infection vs uninfected in <em>Homo sapiens</em> (CD8+ T cell)</td>
<td>E-GEO-D6740</td>
</tr>
<tr>
<td>17</td>
<td>bic-deficient vs wild type in <em>Mus musculus</em> (Th1)</td>
<td>E-TABM-232</td>
</tr>
<tr>
<td>18</td>
<td>Brng null vs wild type in <em>Mus musculus</em></td>
<td>E-GEO-D5371</td>
</tr>
<tr>
<td>19</td>
<td>SOD1 mutant vs control in <em>Mus musculus</em> (6 weeks &amp; spinal cord)</td>
<td>E-GEO-D3343</td>
</tr>
<tr>
<td>20</td>
<td>IL-22 vs control - untreated in <em>Homo sapiens</em></td>
<td>E-GEO-D7216</td>
</tr>
<tr>
<td>21</td>
<td>alpha-tocopherol + gamma-tocopherol vs none in <em>Mus musculus</em> (5 months)</td>
<td>E-GEO-D8150</td>
</tr>
<tr>
<td>22</td>
<td>ulcerative colitis vs normal in <em>Homo sapiens</em> (female)</td>
<td>E-GEO-D3365</td>
</tr>
<tr>
<td>23</td>
<td>IMP1,3A vs mock in <em>Homo sapiens</em></td>
<td>E-MEXP-548</td>
</tr>
<tr>
<td>24</td>
<td>lipopolysaccharide vs none in <em>Homo sapiens</em> (low response)</td>
<td>E-GEO-D3491</td>
</tr>
<tr>
<td>25</td>
<td>chimpanzee diet vs control diet in <em>Mus musculus</em></td>
<td>E-GEO-D6297</td>
</tr>
</tbody>
</table>

Table S7. Query results for a pancreatic cancer comparison (query Q). Text in parentheses after the name of the species is the context of the corresponding comparisons. The third column contains the ArrayExpress accession number of the source data set.
Table S8. Query results for a glioblastoma comparison (query Q). Text in parentheses after the name of the species is the context of the corresponding comparisons.

Q glioblastoma vs normal in *Homo sapiens*  
1 experimental autoimmune encephalomyelitis (recovery) vs normal in *Rattus norvegicus*  
2 experimental autoimmune encephalomyelitis (relapsing) vs normal in *Rattus norvegicus*  
3 astrocytic tumor vs normal in *Homo sapiens*  
4 kainate vs control in *Rattus norvegicus* (24 h)  
5 neurofibrillary tangle vs normal in *Homo sapiens*  
6 0.5 h vs 0 in *Rattus norvegicus* (sham & *Rattus norvegicus*)  
7 experimental autoimmune encephalomyelitis (acute) vs normal in *Rattus norvegicus*  
8 8 h vs 0 in *Rattus norvegicus* (sham & *Rattus norvegicus*)  
9 spinal cord contusion vs none in *Rattus norvegicus*  
10 lateral fluid percussion-induced injury vs sham in *Rattus norvegicus* (Rattus norvegicus & 8 h)  
11 R6/1 transgenic vs wild_type in *Mus musculus* (27 weeks)  
12 R6/1 transgenic vs wild_type in *Mus musculus* (18 weeks)  
13 9 d vs 0 d in *Mus musculus* (embryoid body)  
14 3H-1,2-dithiole-3-thione vs none in *Rattus norvegicus*  
15 kainate vs control in *Rattus norvegicus* (240 h)  
16 diabetes mellitus vs normal in *Rattus norvegicus* (vanadyl sulfate)  
17 1.5 d vs 0 d in *Mus musculus* (embryoid body)  
18 adenoviral vector vs none in *Mus musculus*  
19 FrCasE vs mock infected in *Mus musculus*  
20 spinal nerve transection vs sham surgery in *Rattus norvegicus*  
21 bipolar disorder vs normal in *Homo sapiens* (male)  
22 severe spinal cord injury vs normal in *Rattus norvegicus* (spinal cord (T10) & 2 d)  
23 creatine vs control in *Mus musculus*  
24 moderate spinal cord injury vs normal in *Rattus norvegicus* (spinal cord (T10) & 2 d)  
25 monocular deprivation right eyelid sutured vs control in *Mus musculus*