Bioinformatics Advance Access published November 20, 2011

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.

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 REx (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 REx 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*, short isoform (*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

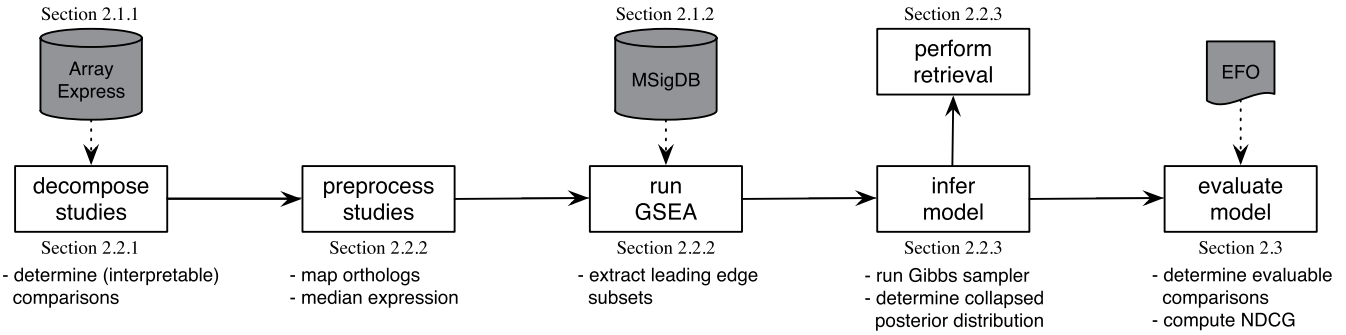

Fig. 1. Flowchart outlining the key steps of the REx information retrieval framework. "MSigDB" is the Molecular Signature Database, "NDCG" is the Normalized Discounted Cumulative Gain measure.

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.

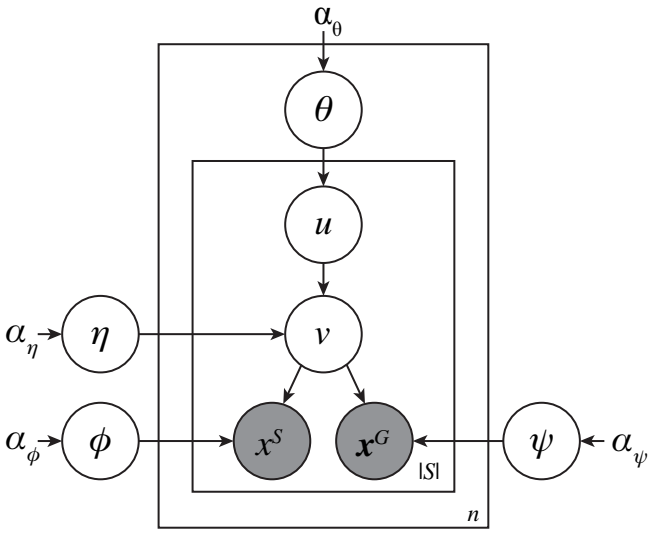

Fig. 2. Plate diagram of the proposed graphical model. Rectangles indicate sets of variables, with the cardinality of the set marked in the bottom right corner. Gray nodes correspond to observed data.

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, $\boldsymbol{\phi}_t$ and $\boldsymbol{\psi}_t$. The vector $\boldsymbol{\phi}_t$ has length equal to the number of gene sets and models the binary activation status of each gene set; the vector $\boldsymbol{\psi}_t$ 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 i has a discrete distribution over so-called modules, parameterized by a vector $\boldsymbol{\theta}_i$; each module m has a discrete distribution over submodules, parameterized by a vector $\boldsymbol{\eta}_m$. The selection of a submodule t is made by first choosing a module m using $\boldsymbol{\theta}_i$ and then choosing submodule t using $\boldsymbol{\eta}_m$. The variables u and v in Figure 2 indicate the chosen module and submodule, respectively. Finally, we endow each $\boldsymbol{\theta}_i$ and $\boldsymbol{\eta}_m$ with conjugate symmetric Dirichlet prior distributions, and each $\phi_{t,j}$ and $\psi_{t,g}$ with conjugate symmetric Beta prior distributions, parameterized by α_θ, α_η, α_ϕ, and α_ψ, 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). For reasonably uninformative priors, such as the ones used in this paper, this choice does not markedly decrease generality.

We use a collapsed Gibbs sampler (Griffiths and Steyvers, 2004) to compute approximate posterior distributions for $\boldsymbol{u}$ and $\boldsymbol{v}$, as well as estimates for $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ 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

$$rel(q,r) \stackrel{\text{def}}{=} \int_{\boldsymbol{\Psi}} P(x_q|\boldsymbol{\Psi}_r)P(\boldsymbol{\Psi}|X)d\boldsymbol{\Psi}\,,$$

where $\boldsymbol{X}$ is the input data and $\boldsymbol{\Psi}$ is the collection of random variables upon which inference is performed (Buntine *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:

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

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 REx 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 *evaluable* comparison. A total of 219 evaluable 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 Cumulative Discounted Gain (NDCG) measure (Järvelin and Kekäläinen, 2002) to evaluate REx 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/fg/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 REx 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 (Chrast *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 (Aleman *et al.*, 2005; Halvorsen *et al.*, 2007), and under-expressed in breast cancer (Kwak *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* (Kwak *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.

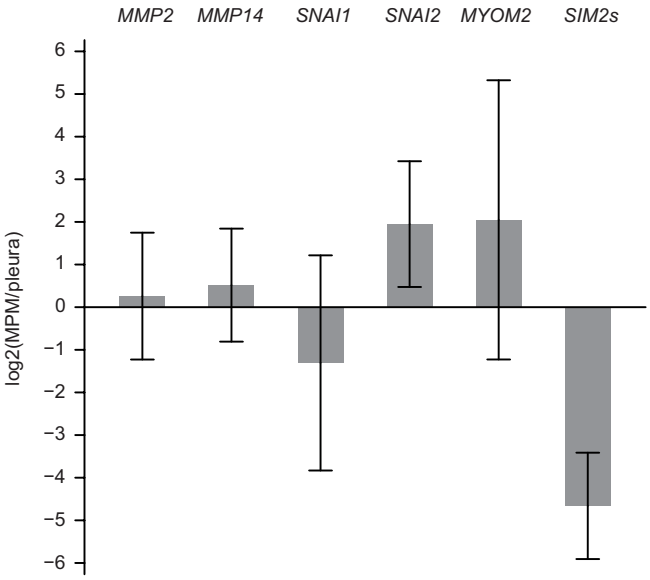

Fig. 3. Bar plots of MPM vs pleura log-ratio gene expression values obtained via RT-PCR. The height of the bars represent the log-ratio expression of the corresponding genes and error bars indicate the standard deviation.

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; Kwak *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 (Tsuchiya *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.

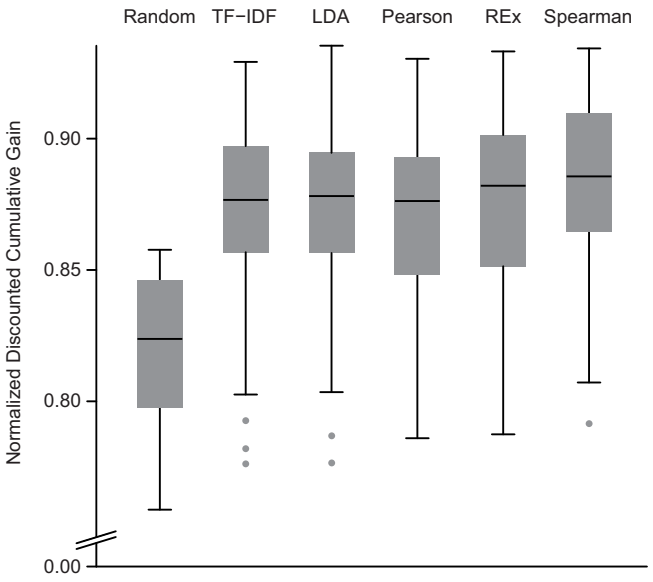

Fig. 4. Data-driven retrieval performance, NDCG results. The box plots summarize the distribution of Normalized Discounted Cumulative Gain (NDCG) results for 219 interpretable query comparisons. "LDA" corresponds to our earlier method (Caldas *et al.*, 2009).

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%

(s.d. = 0.03) of the NDCG values obtained using the Spearman correlation approach.

The difference in performance between REx and the best-performing approach seems to be consistent, albeit small. Although Spearman rank correlation performs slightly better than REx, all evaluated retrieval methods have essentially a very similar performance according to the gold standard derived from the EFO. The advantage of REx is that it readily provides key information for the interpretation of the results, as illustrated by our case studies. The fact that our proposed model-based relevance measure outperforms other approaches that use the GSEA encoding, but performs slightly worse than Spearman correlation with a fold-change encoding, suggests that the proposed relevance measure itself is sensible, but that the current GSEA encoding may be suboptimal. Finally, the retrieval results obtained by our method are robust with regard to the number of modules and submodules. Changing the number of modules and/or submodules yields retrieval results that are significantly correlated with the reported ones. We also found that for multiple choices in the number of modules and submodules, the structures inferred by the model yield comparison-to-gene-set probabilities that are significantly correlated with the comparison-to-gene-set probabilities in the final model, which justifies the similarity in the query results. The details are described in Supplementary Text S10.

4 CONCLUSION

We proposed a method, REx, for performing data-driven information retrieval in a heterogeneous, large repository of transcriptomics studies. By associating studies with shared patterns of differential gene set and gene expression, REx facilitates the analysis of the retrieval results. We also proposed an ontology-based approach to evaluate the retrieval results, which will become more precise, as more and more phenotypes are mapped to the ontology. Additionally, we carried out case studies showing that the method yields biologically meaningful results. In one of the case studies, REx suggested a novel connection between differential expression of *SIM2s* and malignant pleural mesothelioma (MPM), which we validated by observing significant *SIM2s* under-expression in an independent set of MPM tumor samples. These results show that REx can be used to drive biological discovery. Finally, we pointed out that both REx and existing work are particular cases of a general framework that unifies meta-analysis and information retrieval.

As the main aim of information retrieval and meta-analysis methods is to join multiple, heterogeneous data sets in order to obtain robust and novel findings, one particularly important direction of research is how to extend the current framework to alternative or multiple data types. Generally, REx can be applied to functional genomics data types for which GSEA is a suitable analysis method (Subramanian *et al.*, 2005). While not demonstrated here, the method is immediately applicable to gene expression data generated with sequencing technologies (RNA-seq) once gene-level measurements have been derived. The same applies to data sets obtained from quantitative proteomics studies, e.g. using mass spectrometry, which can also be analyzed with GSEA. REx can also be applied to collections of metabolite profiling studies, but there are some practical challenges that would have to be addressed, such as the need for an appropriate collection of metabolite sets that 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

Aleman, M. J. *et al.* (2005). Inhibition of Single Minded 2 gene expression mediates tumor-selective apoptosis and differentiation in human colon cancer cells. *Proc. Nat. Acad. Sci. U.S.A*, **102**, 12765–12770.

Ashburner, M. *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**(1), 25–9.

Barrett, T. *et al.* (2009). NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.

Buntine, W. *et al.* (2004). A scalable topic-based open source search engine. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence 2004*, pages 228–234, Los Alamitos. IEEE Computer Society.

Caldas, J. *et al.* (2009). Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, **25**, i145–i153.

Casarsa, C. *et al.* (2011). Epithelial-to-mesenchymal transition, cell polarity and stemness-associated features in malignant pleural mesothelioma. *Cancer Lett.*, **302**, 136–143.

Chrast, R. *et al.* (2000). Mice trisomic for a bacterial artificial chromosome with the single-minded 2 gene (sim2) show phenotypes similar to some of those present in the partial trisomy 16 mouse models of Down syndrome. *Hum. Mol. Genet.*, **9**, 1853–1864.

Crispi, S. *et al.* (2009). Global gene expression profiling of human pleural mesotheliomas: Identification of matrix metalloproteinas 14 (MMP-14) as potential tumour target. *PLoS One*, **4**, e7016.

Engreitz, J. *et al.* (2011). Content-based microarray search using differential expression profiles. *BMC Bioinformatics*, **11**, 603.

Feng, M. *et al.* (2010). Store-independent activation of Orai1 by SPCA2 in mammary tumors. *Cell*, **143**, 84–98.

Fujibuchi, W. *et al.* (2007). CellMontage: similar expression profile search server. *Bioinformatics*, **23**, 3103–3104.

Gordon, G. J. *et al.* (2005). Identification of novel candidate oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling. *Am. J. Pathol.*, **166**, 1827–1840.

Goshu, E. *et al.* (2002). *Sim2* mutants have developmental defects not overlapping with those of *Sim1* mutants. *Mol. Cell Biol.*, **22**, 4147–4157.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *P. Natl. Acad. Sci. U. S. A.*, **101**, 5228–5235.

Guan, Y. *et al.* (2010). Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol.*, **6**, e1000991.

Halvorsen, O. J. *et al.* (2007). Increased expression of SIM2-s protein is a novel marker of aggressive prostate cancer. *Clin. Cancer Res.*, **13**, 892–897.

Hasle, H. *et al.* (2000). Risks of leukaemia and solid tumours in individuals with Down's syndrome. *Lancet*, **355**, 165–169.

Henzi, T. *et al.* (2009). SV40-induced expression of calretinin protects mesothelial cells from asbestos cytotoxicity and may be a key factor contributing to mesothelioma pathogenesis. *Am. J. Pathol.*, **174**, 2324–2336.

Hu, G. and Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PLoS One*, **4**, e6536.

Huang, H. *et al.* (2010). Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *P. Natl. Acad. Sci. U. S. A.*, **107**, 6823–6828.

Hunter, L. *et al.* (2001). GEST: a gene expression search tool based on a novel bayesian similarity metric. *Bioinformatics*, **17**, S115–S122.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM T. Inform. Syst.*, **20**(4), 422–446.

Kapushesky, M. *et al.* (2009). Gene expression atlas at the European Bioinformatics Institute. *Nucleic Acids Res.*, **38**, D690–D698.

Kupershmidt, I. *et al.* (2010). Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One*, **5**, e13066.

Kwak, H. I. *et al.* (2007). Inhibition of breast cancer growth and invasion by single-minded 2s. *Carcinogenesis*, **28**, 259–266.

Laffin, B. *et al.* (2008). Loss of singleminded-2s in the mouse mammary gland induces an epithelial-mesenchymal transition associated with up-regulation of slug and matrix metalloprotease 2. *Mol. Cell. Biol.*, **28**, 1936–1946.

Lamb, J. *et al.* (2006). The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Li, W. and McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In W. W. Cohen and A. Moore, editors, *Proceedings of the Twenty-Third International Conference on Machine Learning*, pages 577–584, New York. ACM Press.

Malone, J. *et al.* (2010). Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.

Manning, C. D. *et al.* (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Parkinson, H. *et al.* (2009). ArrayExpress update — from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.

Paruthiyil, S. *et al.* (2011). Estrogen receptor β causes a G2 cell cycle arrest by inhibiting CDK1 activity through the regulation of cyclin B1, GADD45A, and BTG2. *Breast Cancer Res. Treat.*, **to appear**.

Pinton, G. *et al.* (2009). Estrogen receptor-β affects the prognosis of human malignant mesothelioma. *Cancer Res.*, **69**, 4598–4604.

Pulver-Kaste, R. A. *et al.* (2006). Ca^{2+} source-dependent transcription of CRE-containing genes in vascular smooth muscle. *Am. J. Physiol. Heart. Circ. Physiol.*, **291**, H97–H105.

Segal, E. *et al.* (2004). A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Subramanian, A. *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

Tsuchiya, Y. *et al.* (2005). Cytochrome p450-mediated metabolism of estrogens and its regulation in human. *Cancer Lett.*, **227**, 115–124.

Woods, S. *et al.* (2008). The bHLH/Per-Arnt-Sim transcription factor SIM2 regulates muscle transcript myomesin2 via a novel, non-canonical E-box sequence. *Nucleic Acids Res.*, **36**, 3716–3727.

Zhu, Y. *et al.* (2008). GEOmetadb: Powerful alternative search engine for the gene expression omnibus. *Bioinformatics*, **24**, 2798–2800.