Data and text mining

Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma

José Caldas1,†, Nils Gehlenborg2,3,†,‡, Eeva Kettunen4, Ali Faisal1, Mikko Rönty5, Andrew G. Nicholson6, Sakari Knuutila7, Alvis Brazma2 and Samuel Kaski1,8,∗

1Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Helsinki, Finland, 2Functional Genomics Group, European Bioinformatics Institute, 3Graduate School of Life Sciences, University of Cambridge, Cambridge, UK, 4Health and Work Ability, Biological Mechanisms and Prevention of Work-Related Diseases, Finnish Institute of Occupational Health, 5Department of Pathology, HUSLAB, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland, 6Department of Histopathology, Royal Brompton Hospital, London, UK, 7Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital and 8Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

Received on July 5, 2011; revised on October 10, 2011; accepted on November 13, 2011

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 real-time polymerase chain reaction in an independent set of mesothelioma samples.

Availability: Supplementary data and source code are available from http://www.sbi.ac.uk/fg/research/resx.

Contact: samuel.kaski@aalto.fi

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on July 5, 2011; revised on October 10, 2011; accepted on November 13, 2011

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 datasets 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 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 (Caldas et al., 2009; Engreit et al., 2011; Fujibuchi et al., 2007; Hu and Agarwal, 2009; Huang et al., 2010; Hunter et al., 2001; Kapushesky et al., 2009; Kupershmidt et al., 2010; Lamb et al., 2006; Segal et al., 2004) 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: (i) a decomposition of the experimental design of studies into differential expression (pairwise comparison) of genes or gene sets; (ii) 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; (iii) a method to extract biological patterns of interest from the encoded studies; and (iv) 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 (Fujibuchi et al., 2007; Hu and Agarwal, 2009; Hunter et al., 2001; Kupershmidt et al., 2010; Lamb et al., 2006).

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 versus 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 versus disease) in a multifactorial study while ignoring additional experimental variables may introduce confounding factors.

In this article, 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.

![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.](image)

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 <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 underexpression 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.
We applied the following criteria:

2.1 Data

2.1.1 Gene expression studies. Datasets 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 datasets that include a preprocessed expression matrix and sufficiently curated annotation. The datasets 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 datasets 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 specimen and RT-PCR. Tumor tissue specimens were obtained from 10 MPM patients who were diagnosed with mesothelioma tumor at Royal Brompton and Harefield NHS Trust, UK. Of these, six were epithelial and four were biphasic MPMs. As a control, we used a microscopically normal, scarred 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, and SIM2 via RT-PCR. We provide the full details of our experimental procedure in Supplementary Material S1.

2.2 Information retrieval framework

2.2.1 Study decomposition. The collected and preprocessed datasets 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 Material 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 datasets. The full list of control factor values is shown in Supplementary Material 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 downregulated 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 Material 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 dataset 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 \( r \); then, the binary activation status of gene set \( j \) is a sample from a Bernoulli distribution parameterized by \( \phi_{jr} \), while for each gene \( g \) in that gene set we generate its leading edge subset membership by sampling from a Bernoulli distribution with parameter \( \psi_{tg} \). 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 the so-called modules, parameterized by a vector \( \theta \); each module \( m \) has a discrete distribution over submodules, parameterized by a vector \( g_m \). The selection of a submodule \( r \) is made by first choosing a module
m using $\theta$, and then choosing submodule $t$ using $q_t$. The variables $a$ and $v$ in Figure 2 indicate the chosen module and submodule, respectively. Finally, we endow each $\theta_t$ and $q_t$ with conjugate symmetric Dirichlet prior distributions, and each $\phi_t$ and $\psi_t$ with conjugate symmetric Beta prior distributions, parameterized by $a_t$, $a_t$, $a_q$ and $a_q$, 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 article, this choice does not markedly decrease generality.

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

$$\text{rel}(q, r) \approx \int P(X|x, \Theta, \Phi, \Psi) P(\Theta) P(\Phi) P(\Psi) \delta(a) \delta(v) \, \mathrm{d}a \, \mathrm{d}v,$$

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

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

$$P(\text{gene set } s \text{ is active|comparison } i) = \frac{\sum_{t=1}^{M} \sum_{u=1}^{T} h_{su} / \theta_{su} \phi_{su} \psi_{su}}{\max(\sum_{t=1}^{M} \sum_{u=1}^{T} h_{su} / \theta_{su} \phi_{su} \psi_{su})}$$

The full details of our model are described in Supplementary Material 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 is given by the following expression:

$$P(\text{gene set } s \text{ is active|comparison } i) = \frac{\sum_{t=1}^{M} \sum_{u=1}^{T} h_{su} / \theta_{su} \phi_{su} \psi_{su}}{\max(\sum_{t=1}^{M} \sum_{u=1}^{T} h_{su} / \theta_{su} \phi_{su} \psi_{su})}$$

Finally, our model allows computing for each comparison the marginal probability that each gene set is active. Using a one-tailed Wilcoxon rank-sum test, where the samples being compared are all the gene (gene, gene) joint probabilities that involve a particular gene versus 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/lg/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 Material S7–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 MPM

We queried the database with a comparison of MPM versus 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...
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 in a human colon carcinoma cell line at 18 h. 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, overexpression of SIM2 has been shown to induce a partial Down syndrome phenotype in mouse (Chrust et al., 2000). Due to the fact that individuals with Down syndrome have a higher risk for leukemia 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 overexpressed in colon and prostate cancer (Alemán et al., 2005; Halvorsen et al., 2007), and underexpressed in breast cancer (Kwak et al., 2007). The connection found by REx suggests that SIM2s may be differentially expressed in MPM. To the best of our knowledge, SIM2 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 underexpressed in comparison to a pleural control (fold-change = 0.87). We tested via RT-PCR measurements whether SIM2s underexpression 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 short and long isoforms of SIM2l (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 Fig. 3.

SIM2s was significantly underexpressed (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 overexpression 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 overexpression of SNAI2 (P < 0.05). Overexpression of SNAI2 and MMP3 is consistent with their potential role as repressive transcriptional targets of SIM2s (Kwak et al., 2007; Laffin et al., 2008). Finally, overexpression 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 overexpression.

The fact that we observed statistically significant SIM2s underexpression 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). Overexpression 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 upregulated in the SIM2s depletion study analyzed by our model (Alemán et al., 2005), is a transcriptional target of ERβ (Parathyrl 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 Section 2. We assessed the functional

![Fig. 3. Bar plots of MPM versus 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 SD.](http://bioinformatics.oxfordjournals.org/Downloaded from http://bioinformatics.oxfordjournals.org)
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’s rank correlation approach based on the fold-change ratios of the expression data, a Pearson’s 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’s correlation and the remaining non-random approaches is also significant (q < 0.05), as is the difference between Spearman’s 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’s correlation performed significantly worse than all other non-random approaches, while Spearman’s correlation performed significantly better. We then analyzed the magnitude of the difference in NDCG values between our method and Spearman’s correlation. The NDCG values obtained by REx are on average 99% (SD = 0.03) of the NDCG values obtained using the Spearman’s correlation approach.

The difference in performance between REx and the best-performing approach seems to be consistent, albeit small. Although Spearman’s 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 and performs slightly worse than Spearman’s correlation with a fold-change encoding, suggests that the proposed relevance measure itself is sensible, but 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 Material S11.

### 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
Although we did not consider integration of multiple data types, we thank Päivi Tuominen and Tiina Marjomaa for excellent drug repurposing or biomarker development. 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 MPM, which we validated by observing significant SIM2s underexpression 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 unites meta-analysis and information retrieval.

As the main aim of information retrieval and meta-analysis methods is to join multiple, heterogeneous datasets 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 datasets 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 article, 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.

ACKNOWLEDGMENTS

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

Funding: Finnish Funding Agency for Technology and Innovation (40101/07); Pattern Analysis, Statistical Modeling and Computational Learning Network of Excellence (ICT 216886); Portuguese Science and Technology Foundation (SFRH/BDE/39974/2007 to J.C.); Academy of Finland (115572 to E.K.); Finnish Doctoral Programme in Computational Sciences (doctoral fellowship to A.F.); Finnish Center of Excellence in Adaptive Informatics Research (J.C., A.F. and S.K.) and the European Molecular Biology Laboratory (doctoral fellowship to N.G.).

CONFLICT OF INTEREST: none declared.

REFERENCES


Hensch 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.


Information retrieval of transcriptomics data