BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–24

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 the Helsinki University Central Hospital for a non-neoplastic intrabronchial inflammatory polyp. The Coordinating Ethical Review Board, Helsinki and Uusimaa Hospital District (HUS 281/E6/03), has approved the collection of samples and the patient has given informed consent to use his tissue.

Total RNA was extracted from fresh-frozen tissue specimens using miRNeasy 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$\times$ Taqman Gene Expression Master Mix and 1X Taqman assay for a gene of interest (Applied Biosystems) and 0.7 to 1.3 μl of cDNA. Moreover, *SIM2s* and *SIM2l*, and *ACTB* as a reference gene, were studied using 1X 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 *SIM2l* 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 *SIM2l*. 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.

*These authors contributed equally to this work.

†To whom correspondence should be addressed.

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 - unsychronized
control - untreated
control - vector
control - vehicle
control diet
control for EGF
control for heregulin
control polyamide and dihydrotestoterone
control siRNA
control,
empty vector
mock
mock infected
mock transfected
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 gene sets 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 $\boldsymbol{X}$ be a set of n GSEA comparisons. Each comparison $\boldsymbol{x}_i$ is a tuple

$$\boldsymbol{x}_i = (\boldsymbol{x}_{ij})_{j=1}^{S} ,$$

where $\boldsymbol{x}_{ij}$ 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 $\boldsymbol{x}_{ij}$ is specified as

$$\boldsymbol{x}_{ij} = \left(x_{ij}^{(S)}, \boldsymbol{x}_{ij}^{(G)}\right) ,$$

where $x_{ij}^{(S)}$ is a binary variable that describes the activation level of gene set j and $\boldsymbol{x}_{ij}^{(G)}$ 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 $\boldsymbol{x}_{ij}^{(G)}$ 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) $\phi_{t,j} \sim \text{Beta}\left(\alpha_\phi\right)$
   b. **for** each gene $g \in \{1, \ldots, G\}$
      (1) $\psi_{t,g} \sim \text{Beta}\left(\alpha_\psi\right)$
2. **for** each module $m \in \{1, \ldots, M\}$
   a. $\boldsymbol{\eta}_m \sim \text{Dirichlet}\left(\alpha_\eta\right)$
3. **for** each GSEA comparison $i \in \{1, \ldots, n\}$
   a. $\boldsymbol{\theta}_i \sim \text{Dirichlet}\left(\alpha_\theta\right)$
   b. **for** each gene set $j \in \{1, \ldots, S\}$
      (1) $u_{i,j} \sim \text{Discrete}\left(\boldsymbol{\theta}_i\right)$
      (2) $v_{i,j} \sim \text{Discrete}\left(\boldsymbol{\eta}_{u_{i,j}}\right)$
      (3) $x_{i,j}^{(S)} \sim \text{Bernoulli}\left(\phi_{v_{i,j},j}\right)$
      (4) **if** $x_{i,j}^{(S)} = 1$ **then for** gene $g \in \sigma_j$
         (1) $x_{i,j,g}^{(G)} \sim \text{Bernoulli}\left(\psi_{v_{i,j},g}\right)$

In the above process, σ_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 $x_{i,j}^{(S)} = 0$ is uniform, and that the conditional distribution given $x_{i,j}^{(S)} = 1$ 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:

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{X}) =& P(\boldsymbol{\theta})P(\boldsymbol{\eta})P(\boldsymbol{\phi})P(\boldsymbol{\psi})P(\boldsymbol{u}|\boldsymbol{\theta}) \\ & P(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{\eta})P(\boldsymbol{X}|\boldsymbol{v}, \boldsymbol{\phi}, \boldsymbol{\psi}). \end{aligned}$$

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

$$\begin{aligned} P(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}|\boldsymbol{X}) =& \frac{P(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{X})}{P(\boldsymbol{X})} \\ \propto& P(\boldsymbol{\theta})P(\boldsymbol{\eta})P(\boldsymbol{\phi})P(\boldsymbol{\psi})P(\boldsymbol{u}|\boldsymbol{\theta})P(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{\eta}) \\ & P(\boldsymbol{X}|\boldsymbol{v}, \boldsymbol{\phi}, \boldsymbol{\psi}). \end{aligned}$$

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

$$P(\boldsymbol{u}, \boldsymbol{v}|\boldsymbol{X}) = \int P(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}|\boldsymbol{X}) d\boldsymbol{\theta}\, d\boldsymbol{\eta}\, d\boldsymbol{\phi}\, d\boldsymbol{\psi} , \qquad (1)$$

where the variables $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$ are collapsed, or integrated out. We will refer to the distribution of $\boldsymbol{u}$ and $\boldsymbol{v}$ given $\boldsymbol{X}$ as the *collapsed* posterior distribution. The collapsed posterior distribution $P(\boldsymbol{u}, \boldsymbol{v}|\boldsymbol{X})$ 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

$$\begin{aligned} P(\boldsymbol{u}, \boldsymbol{v}|\boldsymbol{X}) =& \int P(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{u}, \boldsymbol{v}|\boldsymbol{X}) d\boldsymbol{\theta}\, d\boldsymbol{\eta}\, d\boldsymbol{\phi}\, d\boldsymbol{\psi} \\ \propto& \int P(\boldsymbol{\theta})P(\boldsymbol{u}|\boldsymbol{\theta}) d\boldsymbol{\theta} \int P(\boldsymbol{\eta})P(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{\eta}) d\boldsymbol{\eta} \\ & \int P(\boldsymbol{\psi})P(\boldsymbol{\phi})P(\boldsymbol{X}|\boldsymbol{\psi}, \boldsymbol{\phi}) d\boldsymbol{\phi}\, d\psi \\ =& P(\boldsymbol{u})P(\boldsymbol{v}|\boldsymbol{u})P(\boldsymbol{X}|\boldsymbol{v}, \boldsymbol{u}). \end{aligned}$$

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:

$$P(\boldsymbol{u}) = \left(\frac{\Gamma(M\alpha_\theta)}{\Gamma(\alpha_\theta)^M}\right)^n \prod_{i=1}^{n} \frac{\prod_{m=1}^{M} \Gamma(\alpha_\theta + c_{im})}{\Gamma(M\alpha_\theta + c_{i\cdot})},$$

$$P(\boldsymbol{v}|\boldsymbol{u}) = \left(\frac{\Gamma(T\alpha_\eta)}{\Gamma(\alpha_\eta)^T}\right)^M \prod_{m=1}^{M} \frac{\prod_{t=1}^{T} \Gamma(\alpha_\eta + c_{mt})}{\Gamma(T\alpha_\eta + c_{m\cdot})},$$

$$P(\boldsymbol{X}|\boldsymbol{u},\boldsymbol{v}) = \prod_{t=1}^{T}\left(\prod_{s=1}^{S}\frac{B(\alpha_\phi + c_{ts+}, \alpha_\phi + c_{ts-})}{B(\alpha_\phi,\alpha_\phi)}\right) \left(\prod_{g=1}^{G}\frac{B(\alpha_\psi + c_{tg+}, \alpha_\psi + c_{tg-})}{B(\alpha_\psi,\alpha_\psi)}\right).$$

We use dot $(\cdot)$ 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 α_θ, α_η, α_ϕ, and α_ψ. Finally, the above probabilities depend on $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{X}$ 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_{mt} is the number of times that submodule t was chosen by the v variable, given that the corresponding u variable chose module m; c_{ts+} and c_{ts-} 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+} and c_{tg-} 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 found to be active.[1]

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

$$P(u_{is}=m, v_{is}=t|u_{-(is)}, v_{-(is)}, \boldsymbol{X}) \propto (\alpha_\theta + c_{im}^{-s})\frac{\alpha_\eta + c_{mt}^{-(ds)}}{T\alpha_\eta + c_{m\cdot}^{-(ds)}} \frac{(\alpha_\phi + c_{ts+}^{-(ds)})^{x_{is}^{(S)}}(\alpha_\phi + c_{ts-}^{-(ds)})^{1-x_{is}^{(S)}}}{2\alpha_\phi + c_{ts}^{-(ds)}} \left(\prod_{g\in\sigma(s)}\frac{(\alpha_\psi + c_{tg+}^{-(ds)})^{x_{isg}^{(G)}}(\alpha_\psi + c_{tg-}^{-(ds)})^{1-x_{isg}^{(G)}}}{2\alpha_\psi + c_{tg}^{-(ds)}}\right)^{x_{is}^{(S)}}.$$

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}^{-s} 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 $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$, based on their predictive distributions over new observations (Griffiths and Steyvers, 2004). The estimates are the following (we omit the derivations):

$$\theta_{im} = \frac{\alpha_\theta + c_{im}}{M\alpha_\theta + c_{i\cdot}}, \tag{2}$$

$$\eta_{mt} = \frac{\alpha_\eta + c_{mt}}{T\alpha_\eta + c_{m\cdot}}, \tag{3}$$

$$\phi_{ts} = \frac{\alpha_\phi + c_{ts+}}{2\alpha_\phi + c_{ts}}, \tag{4}$$

$$\psi_{tg} = \frac{\alpha_\psi + c_{tg+}}{2\alpha_\psi + c_{tg}}. \tag{5}$$

$$\tag{6}$$

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 (Buntine *et al.*, 2004; Steyvers and Griffiths, 2007).

The expectation relating query comparison i to comparison j is defined as

$$\mathrm{rel}(i,j) \stackrel{\mathrm{def}}{=} \sum_{\boldsymbol{u},\boldsymbol{v}}\int P(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{u},\boldsymbol{v}|\boldsymbol{X}) \log P_j(\boldsymbol{x}_i|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{u},\boldsymbol{v}) d\boldsymbol{\theta}\, d\boldsymbol{\eta}\, d\boldsymbol{\phi}\, d\boldsymbol{\psi},$$

where $\log P_j(\boldsymbol{x}_i|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{u},\boldsymbol{v})$ 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 $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$. This log-probability is given by

$$\log P_j(\boldsymbol{x}_i|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{u},\boldsymbol{v}) = \sum_{s=1}^{S}\log\sum_{m=1}^{M}\sum_{t=1}^{T}\theta_{jm}\eta_{mt}\phi_{ts}^{x_{is}^{(S)}} (1-\phi_{ts})^{1-x_{is}^{(S)}} \left(\prod_{g\in\sigma(s)}\psi_{tg}^{x_{isg}^{(G)}}(1-\psi_{tg})^{x_{isg}^{(G)}}\right)^{x_{is}^{(S)}}.$$

We approximate the expectation by collecting a number of samples after the sampler has converged. For each of those samples, we compute estimates for $\boldsymbol{\theta}$, $\boldsymbol{\eta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\psi}$, as described above; then, we compute $\log P_j(\boldsymbol{x}_i|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{u},\boldsymbol{v})$; finally, we approximate the integral by averaging over the computed values of $\log P_j(\boldsymbol{x}_i|\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{\psi},\boldsymbol{u},\boldsymbol{v})$. In 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 comparisons over equally spaced 20 iterations among the last 500 iterations. Approximating the relevance log-probability by an average over several samples does not suffer from the label-switching problem.

Text S6: Retrieval Evaluation

To evaluate REx 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

[1] Our use of c variables incurs in a slight abuse of notation, as only the subscript symbols in a variable, rather than the name of the variable itself, indicate the actual meaning of the variable. For instance, c_{im} and c_{mt} refer to different classes of assignments because different subscript symbols are used. However, in the current context, this slight abuse of notation yields a clear interpretation and avoids an unnecessary profusion of different variable symbols.

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 (Järvelin 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 $rel(q,r) = J'(Q,R)$ 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 (Järvelin 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 $rel(q,r_i)$, with $i = 1,\ldots,C$ where C is the number of interpretable comparisons, is defined as

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel(q,r_i)} - 1}{log_2(i+1)},$$

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

$$NDCG_p = \frac{DCG_p}{IDCG_p}.$$

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 $rel(q,r)$.

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 knock-out 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).

The third most relevant comparison found by REx is a comparison between normal and insulin-injected human muscle tissue, creating hyperinsulinemic conditions. Hyperinsulinemia may be involved in an association between pancreatic cancer and diabetes (Wang *et al.*, 2003).

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 recent study in *PCSK9* knock-out mice, which showed that the mice were hypoinsulinemic, hyperglycemic and glucose-intolerant and the study suggests that pancreatic islet cells require *PCSK9* for normal functioning (Mbikay *et al.*, 2010). However, potentially contradictory results have been reported as well (Langhi *et al.*, 2009).

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 (T1D) 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 99% 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

Allen-Jennings, A. E. *et al.* (2001). The roles of ATF3 in glucose homeostasis. *J. Biol. Chem*, **276**, 29507–29514.

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

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and Bayesian inference of topic hierarchies. *J. ACM*, **57**(2), 1–30.

Bourquin, J.-P. *et al.* (2006). Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proc. Nat. Acad. Sci. U.S.A*, **103**(9), 3339–44.

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.

Chahin, N. *et al.* (2007). Two causes of demyelinating neuropathy in one patient: CMT1A and POEMS syndrome. *Can. J. Neurol. Sci.*, **34**(3), 380–385.

Engelman, J. A. (2009). Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nat. Rev. Cancer*, **9**, 550–562.

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

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

Funel, N. *et al.* (2009). PMP22 gene duplication in pancreatic ductal adenocarcinoma. *JOP. J. Pancreas*, **10**, 616.

Gabriel, C. M. *et al.* (2000). Anti-PMP22 antibodies in patients with inflammatory neuropathy. *J. Neuroimmunol.*, **104**, 139–146.

Giambonini-Brugnoli, G. *et al.* (2005). Distinct disease mechanisms in peripheral neuropathies due to altered peripheral myelin protein 22 gene dosage or a Pmp22 point mutation. *Neurobiol. Dis.*, **18**, 656–668.

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

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.

Hibbs, M. A. *et al.* (2007). Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23**(20), 2692–2699.

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.

Hubbard, T. J. P. *et al.* (2009). Ensembl 2009. *Nucleic Acids Res.*, **37**(Database issue), D690–7.

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

Huo, J. S. *et al.* (2006). Profiles of growth hormone (GH)-regulated genes reveal time-dependent responses and identify a mechanism for regulation of activating transcription factor 3 by GH. *J. Biol. Chem.*, **17**, 4132–4141.

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

Jasra, A. *et al.* (2005). MCMC and the label switching problem in Bayesian mixture models. *Stat. Sci.*, **20**, 50–67.

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

Kasprzyk, A. *et al.* (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**(1), 160–9.

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

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

Langhi, C. *et al.* (2009). PCSK9 is expressed in pancreatic δ-cells and does not alter insulin secretion. *Biochem. Bioph. Res. Co.*, **390**, 1288–1293.

Lee, Y. H. *et al.* (2005). Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese pima indians: increased expression of inflammation-related genes. *Diabetologia*, **48**, 1776–1783.

Li, J. *et al.* (2005). Expression analysis of PMP22/Gas3 in premalignant and malignant pancreatic lesions. *J. Histochem. Cytochem.*, **53**, 885–893.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, **89**, 958–966.

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.

Mbikay, M. *et al.* (2010). PCSK9-deficient mice exhibit impaired glucose tolerance and pancreatic islet abnormalities. *FEBS Lett.*, **584**, 701–706.

Meyer zu Hörste, G. *et al.* (2006). Myelin disorders: Causes and perspectives of charcot-marie-tooth neuropathy. *J. Mol. Neurosci.*, **28**, 77–88.

Peng, H. and Hagopian, W. (2006). Environmental factors in the development of type I diabetes. *Rev. Endocr. Metab. Dis.*, **7**, 149–162.

Rahib, L. *et al.* (2007). Glycerol kinase deficiency alters expression of genes involved in lipid metabolism, carbohydrate metabolism, and insulin signaling. *Eur. J. Hum. Genet.*, **15**, 646–657.

Ranheim, T. *et al.* (2008). Genome-wide expression analysis of cells expressing gain of function mutant D374Y-PCSK9. *J. Cell Physiol.*, **217**, 459–467.

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

Stewart, L. K. *et al.* (2008). Quercetin transiently increases energy expenditure but persistently decreases circulating markers of inflammation in C57BL/6J mice fed a high-fat diet. *Metabolism*, **57**, S39–S46.

Steyvers, M. and Griffiths, T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Roday to Meaning*. Lawrence Erlbaum.

Sweet-Cordero, A. *et al.* (2005). An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nature Genetics*, **37**(1), 48–55.

Talantov, D. *et al.* (2005). Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clin. Cancer Res.*, **11**, 7234–7242.

Vilella, A. J. *et al.* (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, **19**(2), 327–35.

Wang, F. *et al.* (2003). The relationship between diabetes and pancreatic cancer. *Mol. Cancer*, **2**, 4.

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.

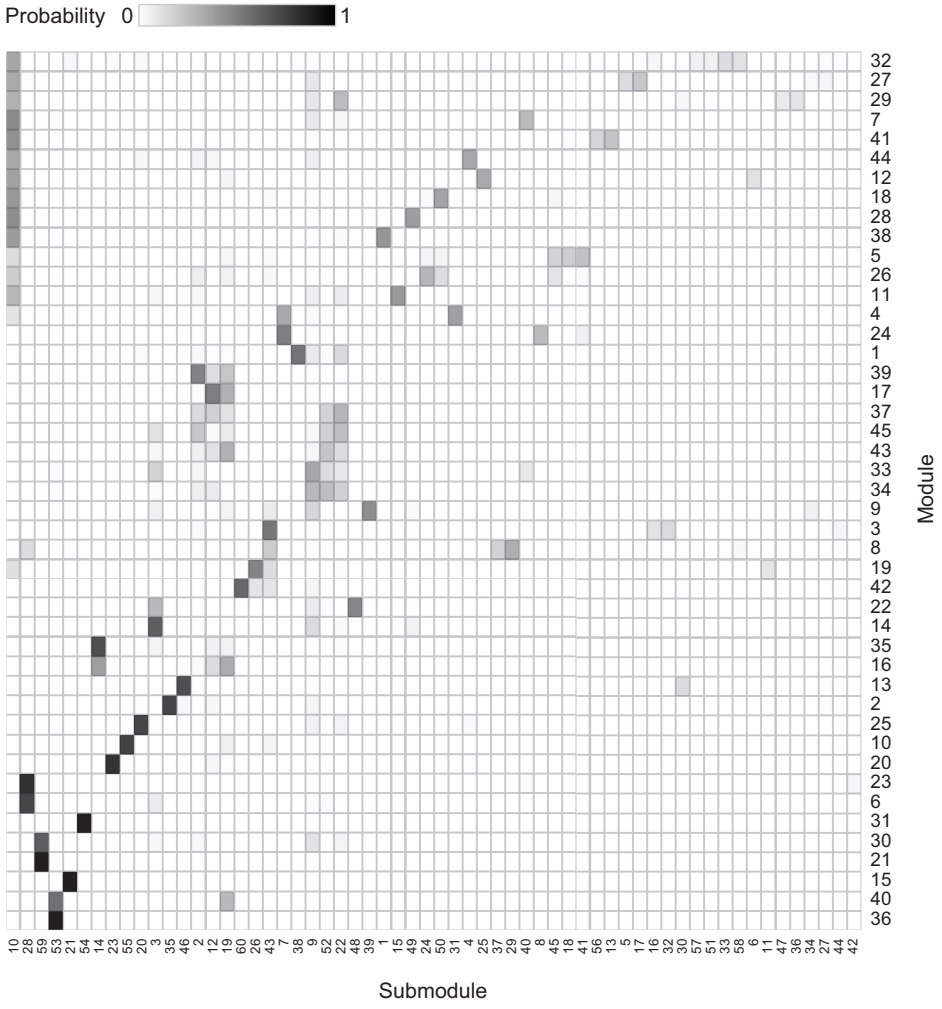


Fig. S3. Heatmap of distributions from modules to submodules.

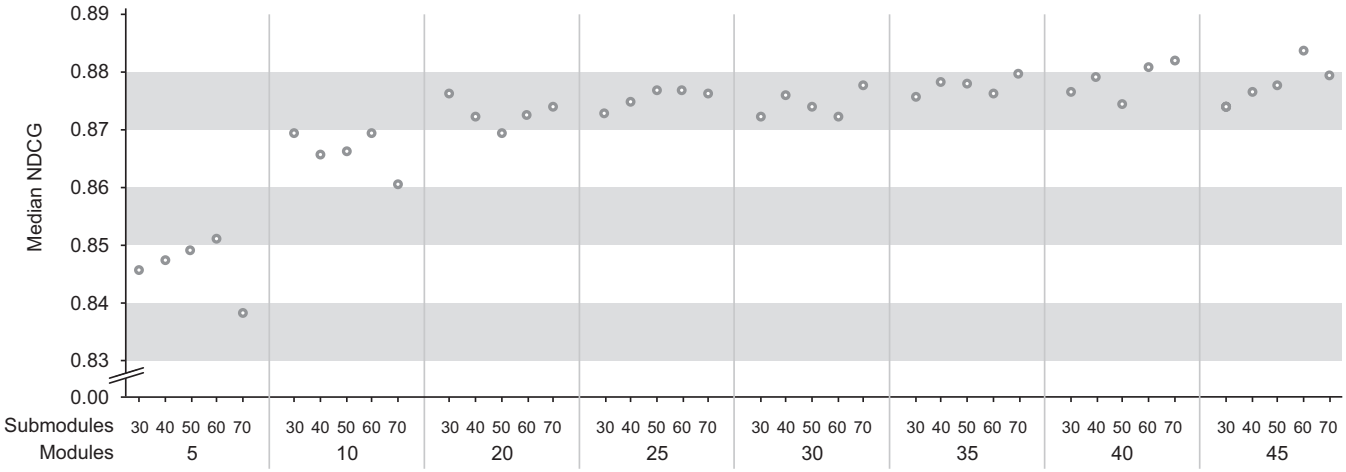
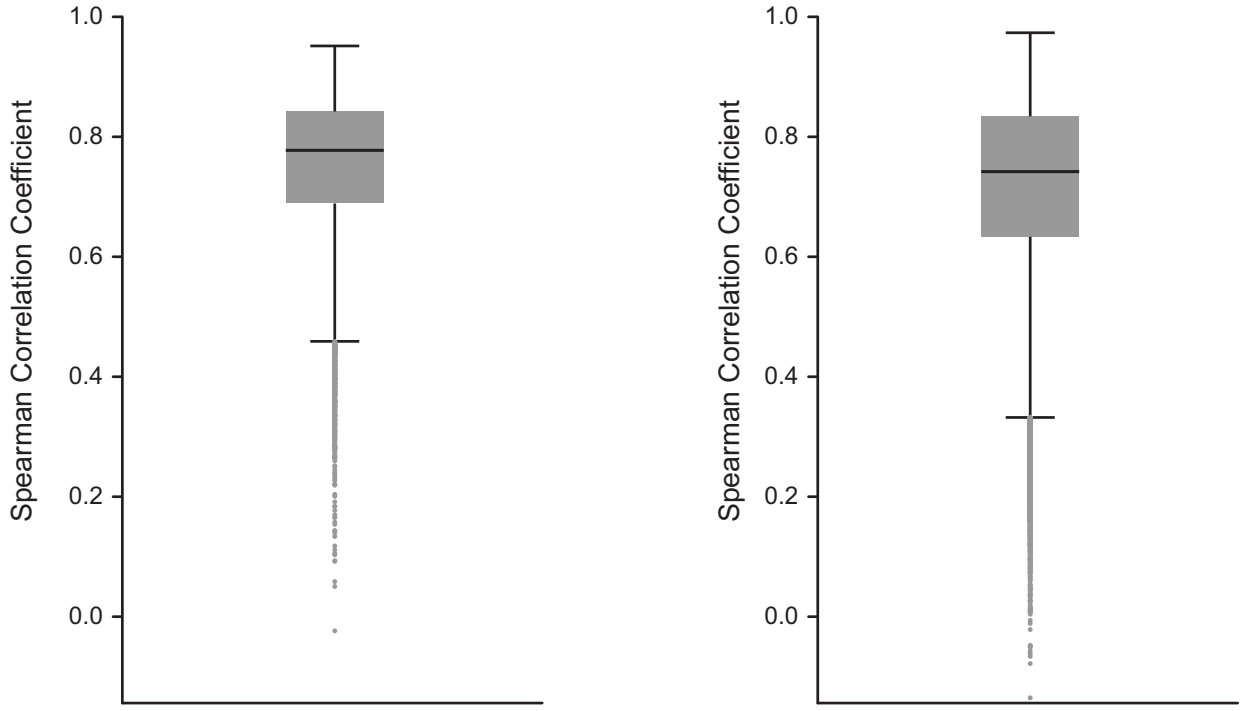


Fig. S4. Median NDCG scores for various alternatives in the number of modules and submodules. The highest NDCG was found for 45 modules and 60 submodules.



(a) Query results.

(b) Comparison-to-gene-set probabilities for GSEA Comparisons.

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

Component	Approaches
Study decomposition	one vs. all (Segal *et al.*, 2004; Kapushesky *et al.*, 2009); case vs. control (Lamb *et al.*, 2006; Kupershmidt *et al.*, 2010); all vs. all (Hu and Agarwal, 2009; Huang *et al.*, 2010; Engreitz *et al.*, 2011); contextualized all vs. all (Caldas *et al.*, 2009); none (Hunter *et al.*, 2001; Fujibuchi *et al.*, 2007)
Measure of DE	hypergeometric test (Segal *et al.*, 2004); t-test (Kapushesky *et al.*, 2009; Hu and Agarwal, 2009); fold-change (Lamb *et al.*, 2006; Kupershmidt *et al.*, 2010; Engreitz *et al.*, 2011); log-rank ratio (Huang *et al.*, 2010); GSEA (Caldas *et al.*, 2009); none (Hunter *et al.*, 2001; Fujibuchi *et al.*, 2007)
Pattern extraction	Unsupervised learning (Segal *et al.*, 2004; Caldas *et al.*, 2009; Engreitz *et al.*, 2011); none (Hunter *et al.*, 2001; Lamb *et al.*, 2006; Fujibuchi *et al.*, 2007; Hu and Agarwal, 2009; Huang *et al.*, 2010; Kapushesky *et al.*, 2009; Kupershmidt *et al.*, 2010)
Relevance measure	Bayes factor (Hunter *et al.*, 2001); correlation (Lamb *et al.*, 2006; Fujibuchi *et al.*, 2007; Hu and Agarwal, 2009; Huang *et al.*, 2010; Kupershmidt *et al.*, 2010; Engreitz *et al.*, 2011); Generative probability (Caldas *et al.*, 2009); none (Segal *et al.*, 2004; Kapushesky *et al.*, 2009)

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.

Q		
	malignant pleural mesothelioma vs normal in *Homo sapiens* (pleura)	E-GEOD-2549
1	**high potassium** vs control in *Homo sapiens*	E-GEOD-2883
2	**thapsigargin** vs control in *Homo sapiens*	E-GEOD-2883
3	**SIM2s** vs control in *Homo sapiens* (18 h)	E-MEXP-101
4	**hydrogen peroxide** vs control in *Homo sapiens* (1 h)	E-GEOD-5339
5	**0.01 ug per kg per day** vs 0 mg per kg per day in *Rattus norvegicus* (17a-ethynylestradiol)	E-TABM-12
6	**POR null** vs wild type in *Mus musculus* (none & ileum)	E-GEOD-4262
7	**10 ug per kg** vs 0 ug per kg in *Rattus norvegicus* (17a-ethynylestradiol & 8 h)	E-MEXP-999
8	**idiopathic dilated cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
9	**calorie-restricted diet** vs normal diet in *Mus musculus* (4 months)	E-GEOD-4786
10	**non-progressive HIV infection** vs uninfected in *Homo sapiens* (CD4+ T cell)	E-GEOD-6740
11	**K14deltaNB-cateninER** vs wild type in *Mus musculus* (1 d)	E-GEOD-1579
12	**endometriosis** vs normal in *Homo sapiens* (mid secretory phase)	E-GEOD-6364
13	**Aldh5a1-/-** vs wild type in *Mus musculus* (hippocampus)	E-GEOD-2866
14	**dexamethasone** vs none in *Homo sapiens* (4 h)	E-GEOD-3040
15	**hydrogen peroxide** vs control in *Homo sapiens* (12 h)	E-GEOD-5339
16	**0.5 h** vs 0 h in *Homo sapiens* (none & uninfected)	E-GEOD-697
17	**ketogenic diet** vs control diet in *Rattus norvegicus*	E-GEOD-1155
18	**hydrogen peroxide** vs control in *Homo sapiens* (24 h)	E-GEOD-5339
19	**vanadium pentoxide** vs control in *Homo sapiens* (12 h)	E-GEOD-5339
20	**superficial transitional cell carcinoma with surrounding carcinoma in situ lesion** vs normal in *Homo sapiens* (bladder)	E-GEOD-3167
21	**retinoic acid** vs none in *Mus musculus* (6 h)	E-GEOD-1588
22	**RP1 knockout** vs wild type in *Mus musculus* (7 d)	E-GEOD-128
23	**HIV-1 infected** vs normal in *Homo sapiens* (none)	E-GEOD-2504
24	**Yersinia enterocolitica WA(pTTS, pP60) (control)** vs uninfected in *Mus musculus* (interferon-gamma & BALB/c)	E-GEOD-2973
25	**BALB/c SCID** vs wild type in *Mus musculus* (Nippostrongylus brasiliensis & 8 d)	E-GEOD-3414

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.

Case	Histological type	Sex	Age	Asbestos exposure	Smoking	Tumor content %
1	biphasic	M	63	Yes	None	80
2	biphasic	M	64	No	None	70
3	epithelial	M	67	No	Ex	50
4	epithelial	M	57	No	Ex	70
5	epithelial	M	76	Yes	Ex	> 50
6	epithelial	F	56	Yes	None	> 50
7	epithelial	M	56	Yes	Ex	40
8	epithelial	M	71	Yes	Ex	30
9	biphasic	M	68	Yes	None	> 50
10	biphasic	M	57	Yes	None	70

Table S3. Clinical data for malignant pleural mesothelioma patients, including histological type, sex, age, asbestos exposure, smoking status, and sample tumor content.

Case	MMP2	MMP14	SNAI1	SNAI2/SLUG	MYOM2	SIM2s
1	0.26	1.62	0.03	1.74	5.06	0.05
2	0.59	0.9	0.47	3	0.2	0.01
3	4.14	2.26	0.43	5.25	1.7	0.04
4	0.45	1.09	0.05	0.6	0.38	0.02
5	0.4	0.33	0.05	1.05	59.31	0.02
6	3.82	4.81	0.46	7.77	3.15	0.03
7	1	0.34	0.73	5.88	446.85	0.06
8	2.04	1.8	4.32	7.36	5.88	0.12
9	3.48	2.84	2.29	9.22	3.65	0.18
10	1.95	3.64	2.3	13.74	1.25	0.03

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.

Q	**benign nevi** vs normal in *Homo sapiens*	E-GEOD-3189
1	**viral cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
2	**polycystic ovary syndrome** vs normal in *Homo sapiens* (none)	E-GEOD-1615
3	**myelodysplastic syndrome** vs normal in *Homo sapiens* (“” & female)	E-GEOD-2779
4	**idiopathic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
5	**familial cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
6	**EWS-FL1 transfected** vs mock transfected in *Homo sapiens*	E-GEOD-1822
7	**ischemic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
8	**Parkinson’s disease** vs normal in *Homo sapiens* (female)	E-GEOD-7621
9	**RAD001** vs placebo in *Mus musculus* (wild_type & 48 h)	E-GEOD-1413
10	**SR-A mutant** vs wild_type in *Mus musculus* (bilateral olfactory bulbectomy & 8 h)	E-GEOD-3455
11	**diabetes mellitus** vs normal in *Rattus norvegicus* (4 weeks)	E-MEXP-515
12	**post-partum cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
13	**Rs1h null** vs wild_type in *Mus musculus*	E-GEOD-5581
14	**hypertrophic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
15	**carcinoma in situ lesion** vs normal in *Homo sapiens* (bladder)	E-GEOD-3167
16	**1 d** vs 0 in *Homo sapiens* (female)	E-GEOD-1295
17	**ochratoxin** vs none in *Rattus norvegicus* (kidney & 7 m)	E-GEOD-2852
18	**malignant melanoma** vs normal in *Homo sapiens*	E-GEOD-3189
19	**non steroidal anti-inflammatory drugs** vs none in *Homo sapiens* (osteoarthritis)	E-GEOD-7669
20	**coxsackievirus B3** vs uninfected in *Homo sapiens* (none & 3 h)	E-GEOD-697
21	**valproic acid** vs none in *Homo sapiens* (normal)	E-GEOD-1615
22	**12 h** vs 0 h in *Homo sapiens* (control)	E-GEOD-3183
23	**Beta4 nAChR subunit null** vs wild_type in *Mus musculus* (none)	E-GEOD-6614
24	**dexamethasone** vs none in *Homo sapiens* (24 h)	E-GEOD-1815
25	**Dysf-/-** vs wild_type in *Mus musculus* (left ventricular myocardium)	E-GEOD-2507

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.

Q	**malignant melanoma** vs normal in *Homo sapiens*	E-GEOD-3189
1	**ochratoxin** vs none in *Rattus norvegicus* (kidney & 7 m)	E-GEOD-2852
2	**hypertrophic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
3	**benign nevi** vs normal in *Homo sapiens*	E-GEOD-3189
4	**ischemic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
5	**idiopathic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
6	**post-partum cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
7	**24 h** vs 0 in *Homo sapiens* (none)	E-GEOD-2803
8	**EWS-FL1 transfected** vs mock transfected in *Homo sapiens*	E-GEOD-1822
9	**familial cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
10	**viral cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
11	**Parkinson's disease** vs normal in *Homo sapiens* (female)	E-GEOD-7621
12	**diabetes mellitus** vs normal in *Rattus norvegicus* (4 weeks)	E-MEXP-515
13	**breast cancer** vs normal in *Homo sapiens*	E-MEXP-882
14	**hypoxic-ischemic injury** vs none in *Mus musculus* (none)	E-GEOD-1999
15	**soluble tumor necrosis factor alpha** vs control in *Mus musculus*	E-GEOD-4518
16	**polycystic ovary syndrome** vs normal in *Homo sapiens* (none)	E-GEOD-1615
17	**myelodysplastic syndrome** vs normal in *Homo sapiens* ("" & female)	E-GEOD-2779
18	**12 h** vs 0 h in *Homo sapiens* (control)	E-GEOD-3183
19	**RAD001** vs placebo in *Mus musculus* (wild_type & 48 h)	E-GEOD-1413
20	**valproic acid** vs none in *Homo sapiens* (normal)	E-GEOD-1615
21	**SR-A mutant** vs wild_type in *Mus musculus* (bilateral olfactory bulbectomy & 8 h)	E-GEOD-3455
22	**carcinoma in situ lesion** vs normal in *Homo sapiens* (bladder)	E-GEOD-3167
23	**Hoxc13 overexpressing transgenic** vs wild_type in *Mus musculus*	E-GEOD-2374
24	**dermatomyositis** vs normal in *Homo sapiens*	E-GEOD-5370
25	**dexamethasone** vs none in *Homo sapiens* (24 h)	E-GEOD-1815

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.

Q		
	pancreatic cancer vs normal in *Homo sapiens*	E-MEXP-950
1	**IRS-4** vs wild_type in *Mus musculus*	E-GEOD-2556
2	**growth hormone** vs control in *Mus musculus* (48 h)	E-GEOD-2120
3	**insulin** vs none in *Homo sapiens*	E-GEOD-7146
4	**D374Y-PCSK9** vs wild_type in *Homo sapiens*	E-MEXP-1235
5	**obesity** vs normal in *Homo sapiens* (male)	E-GEOD-2508
6	**severe malarial anaemia** vs normal in *Homo sapiens*	E-GEOD-1124
7	**partial paw denervation** vs sham denervation in *Rattus norvegicus* (3)	E-GEOD-2874
8	**Gyk knockout** vs wild_type in *Mus musculus*	E-GEOD-4656
9	**quercetin** vs none in *Mus musculus* (POR null & jejunum)	E-GEOD-4262
10	**U0126** vs none in *Homo sapiens* (coxsackievirus B3 & 9 h)	E-GEOD-697
11	**E2F2-/-** vs wild_type in *Mus musculus* (48 h)	E-MEXP-1413
12	**Trembler** vs wild_type in *Mus musculus* (B cell & P4)	E-GEOD-1947
13	**IRS-1** vs wild_type in *Mus musculus*	E-GEOD-2556
14	**24 h** vs 0 h in *Homo sapiens*	E-MEXP-1194
15	**Cnr1 -/- /Cnr2 -/-** vs wild_type in *Mus musculus* (dinitrofluorobenzene)	E-GEOD-7694
16	**non-progressive HIV infection** vs uninfected in *Homo sapiens* (CD8+ T cell)	E-GEOD-6740
17	**bic-deficient** vs wild_type in *Mus musculus* (Th1)	E-TABM-232
18	**Brg1 null** vs wild_type in *Mus musculus*	E-GEOD-5371
19	**SOD1 mutant** vs control in *Mus musculus* (6 weeks & spinal cord)	E-GEOD-3343
20	**IL-22** vs control - untreated in *Homo sapiens*	E-GEOD-7216
21	**alpha-tocopherol + gamma-tocopherol** vs none in *Mus musculus* (5 months)	E-GEOD-8150
22	**ulcerative colitis** vs normal in *Homo sapiens* (female)	E-GEOD-3365
23	**IMP(1,3)A** vs mock in *Homo sapiens*	E-MEXP-548
24	**lipopolysaccharide** vs none in *Homo sapiens* (low response)	E-GEOD-3491
25	**chimpanzee diet** vs control diet in *Mus musculus*	E-GEOD-6297

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.

Q		
	glioblastoma vs normal in *Homo sapiens*	E-MEXP-567
1	**experimental autoimmune encephalomyelitis (recovery)** vs normal in *Rattus norvegicus*	E-MEXP-1025
2	**experimental autoimmune encephalomyelitis (relapsing)** vs normal in *Rattus norvegicus*	E-MEXP-1025
3	**astrocytic tumor** vs normal in *Homo sapiens*	E-MEXP-567
4	**kainate** vs control in *Rattus norvegicus* (24 h)	E-GEOD-1156
5	**neurofibriallary tangle** vs normal in *Homo sapiens*	E-GEOD-4757
6	**0.5 h** vs 0 in *Rattus norvegicus* (sham & Rattus norvegicus)	E-GEOD-2392
7	**experimental autoimmune encephalomyelitis (acute)** vs normal in *Rattus norvegicus*	E-MEXP-1025
8	**8 h** vs 0 in *Rattus norvegicus* (sham & Rattus norvegicus)	E-GEOD-2392
9	**spinal cord contusion** vs none in *Rattus norvegicus*	E-GEOD-2599
10	**lateral fluid percussion-induced injury** vs sham in *Rattus norvegicus* (Rattus norvegicus & 8 h)	E-GEOD-2392
11	**R6/1 transgenic** vs wild_type in *Mus musculus* (27 weeks)	E-GEOD-3621
12	**R6/1 transgenic** vs wild_type in *Mus musculus* (18 weeks)	E-GEOD-3621
13	**9 d** vs 0 d in *Mus musculus* (embryoid body)	E-GEOD-2972
14	**3H-1,2-dithiole-3-thione** vs none in *Rattus norvegicus*	E-GEOD-3173
15	**kainate** vs control in *Rattus norvegicus* (240 h)	E-GEOD-1156
16	**diabetes mellitus** vs normal in *Rattus norvegicus* (vanadyl sulfate)	E-GEOD-3068
17	**1.5 d** vs 0 d in *Mus musculus* (embryoid body)	E-GEOD-2972
18	**adenoviral vector** vs none in *Mus musculus*	E-GEOD-3172
19	**FrCasE** vs mock infected in *Mus musculus*	E-MEXP-459
20	**spinal nerve transection** vs sham surgery in *Rattus norvegicus*	E-MEXP-976
21	**bipolar disorder** vs normal in *Homo sapiens* (male)	E-GEOD-5389
22	**severe spinal cord injury** vs normal in *Rattus norvegicus* (spinal cord (T10) & 2 d)	E-GEOD-464
23	**creatine** vs control in *Mus musculus*	E-GEOD-5140
24	**moderate spinal cord injury** vs normal in *Rattus norvegicus* (spinal cord (T10) & 2 d)	E-GEOD-464
25	**monocular deprivation right eyelid sutured** vs control in *Mus musculus*	E-GEOD-4537

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.