PLOS ONE

RESEARCH ARTICLE

Toward Computational Cumulative Biology by Combining Models of Biological Datasets

Ali Faisal[1], Jaakko Peltonen[1], Elisabeth Georgii[1], Johan Rung[2], Samuel Kaski[1,3]*

1. Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, Espoo, Finland, **2.** European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, United Kingdom, **3.** Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Helsinki, Finland

*samuel.kaski@aalto.fi

OPEN ACCESS

Citation: Faisal A, Peltonen J, Georgii E, Rung J, Kaski S (2014) Toward Computational Cumulative Biology by Combining Models of Biological Datasets. PLoS ONE 9(11): e113053. doi:10.1371/journal.pone.0113053

Editor: Xiaoning Qian, University of South Florida, United States of America

Received: July 4, 2014

Accepted: October 17, 2014

Published: November 26, 2014

Data Availability: The authors confirm that, for approved reasons, some access restrictions apply to the data underlying the findings. The authors confirm that all gene expression data underlying the findings are fully available without restriction from ArrayExpress (E-MTAB-62, E-CBIL-30, E-GEOD-12648, E-GEOD-4667, E-GEOD-8441, E-GEOD-10760, E-GEOD-1295, E-GEOD-474, E-GEOD-9105, E-GEOD-11686, E-GEOD-1786, E-GEOD-6011, E-GEOD-9397, E-GEOD-11971, E-GEOD-3307, E-GEOD-7146, E-GEOD-9676). The citation data underlying the findings (citation graph, h-indexes and impact factors) are available from Thomson Reuters. Reason for restriction of public deposition of citation data: We are not allowed to make the citation data available because it is third party - Copyright Thomson Reuters, 2011. To re-create the citation graph, interested users need to contact Thomson Reuters, Emma Dennis of the research analytics team at "ts.researchservices@thomson.com" and request for license for "raw tagged data". The H-indexes and impact factors are also available through the same contact. More details at: http://thomsonreuters.com/terms-of-use/.

Funding: Academy of Finland (http://www.aka.fi), Finnish Centre of Excellence in Computational Inference Research COIN, 251170, to AF, JP, EG, SK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

A main challenge of data-driven sciences is how to make maximal use of the progressively expanding databases of experimental datasets in order to keep research cumulative. We introduce the idea of a modeling-based dataset retrieval engine designed for relating a researcher's experimental dataset to earlier work in the field. The search is (i) data-driven to enable new findings, going beyond the state of the art of keyword searches in annotations, (ii) modeling-driven, to include both biological knowledge and insights learned from data, and (iii) scalable, as it is accomplished without building one unified grand model of all data. Assuming each dataset has been modeled beforehand, by the researchers or automatically by database managers, we apply a rapidly computable and optimizable combination model to decompose a new dataset into contributions from earlier relevant models. By using the data-driven decomposition, we identify a network of interrelated datasets from a large annotated human gene expression atlas. While tissue type and disease were major driving forces for determining relevant datasets, the found relationships were richer, and the model-based search was more accurate than the keyword search; moreover, it recovered biologically meaningful relationships that are not straightforwardly visible from annotations—for instance, between cells in different developmental stages such as thymocytes and T-cells. Data-driven links and citations matched to a large extent; the data-driven links even uncovered corrections to the publication data, as two of the most linked datasets were not highly cited and turned out to have wrong publication entries in the database.

Introduction

Molecular biology, historically driven by the pursuit of experimentally characterizing each component of the living cell, has been transformed into a data-driven science [1–6] with just as much importance given to the computational and statistical analysis as to experimental design and assay technology. This has brought to the fore new computational challenges, such as the processing of massive new sequencing data, and new statistical challenges arising from the problem of having relatively few (n) samples characterized for relatively many (p) variables—the "large p, small n" problem. High-throughput technologies often are developed to assay many parallel variables for a single sample in a run, rather than many parallel samples for a single variable, whereas the statistical power to infer properties of biological conditions increases with larger sample sizes. For cost reasons, most labs are restricted to generating datasets with the statistical power to detect only the strongest effects. In combination with the penalties of multiple hypothesis testing, the limitations of "large p, small n" datasets are obvious. It is, therefore, not surprising that much work has been devoted to address this problem.

Some of the most successful methods rely on increasing the effective number of samples by combining with data from other, similarly designed, experiments, in a large meta-analysis [7]. Unfortunately, this is not straightforward, either. Although public data repositories, such as the ones at NCBI in the United States and the EBI in Europe, serve the research community with ever-growing amounts of experimental data, they largely rely on annotation and meta-data provided by the submitter. Database curators and semantic tools such as ontologies provide some help in harmonizing and standardizing the annotation, but the user who wants to find datasets that are combinable with her own most often must resort to searches in free text or in controlled vocabularies, which would need significant downstream curation and data analysis before any meta-analysis can be done [8].

Ideally, we would like to let the data speak for themselves. Instead of searching for datasets that have been described similarly, which may not correspond to a statistical similarity in the datasets themselves, we would like to conduct that search in a data-driven way, using as the query the dataset itself or a statistical (rather than a semantic) description of it. This is implicitly done, for example, in multi-task learning, a method from the machine learning field [9,10], where several related estimation tasks are pursued together, assuming shared properties across tasks. Multi-task learning is a form of global analysis, which builds a single unified model of the datasets. But as the number of datasets keeps increasing and the amount of quantitative biological knowledge keeps accumulating, the complexity of building an accurate unified model becomes increasingly prohibitive.

Addressing the "large p, small n" problem requires taking into account both the uncertainty in the data and the existing biological knowledge. We now consider the hypothesized scenario where future researchers increasingly develop hypotheses in terms of (probabilistic) models of their data. Although far from

realistic today, a similar trend exists for sequence motif data, which are often published as Hidden Markov models, for instance in the Pfam database [11].

In this paper, we report on a feasibility study that uses the scenario in which many experiments have been modeled beforehand, potentially by the researcher generating the data or automatically by the database storing the model together with the data. We ask *what could be done with these models towards cumulatively building knowledge from data in molecular biology*? Speaking about models generally and assuming the many practical issues can be solved technically, we arrive at our answer: we propose creating *a modeling-driven dataset retrieval engine*, which a researcher can use for positioning her own measurement data into the context of the earlier biology. The engine will point out relationships between experiments in the form of the retrieval results, which is a naturally understandable interface. The retrieval will be based on data, instead of the state-of-the-art practice of using keywords and ontologies, which will make unexpected and previously unknown findings possible. The retrieval will use the models of the datasets, which, by our assumption above, incorporate the knowledge of the researchers producing the data about what is important in the data, but the retrieval will be designed to be more scalable than building one unified grand model of all data. This also implies that the way the models are utilized needs to be approximate. Compared to existing data-driven retrieval methods [3,5], whole datasets, incorporating the experimental designs, will be matched, instead of individual observations. The remaining question is how to design the retrieval so that it both reveals the interesting and important relationships and is fast to compute.

The model we present is a first step towards this goal. We assume that a new dataset can be explained by a combination of the models for the earlier datasets and a novelty term. This is a mixture modeling or regression task, in which the weights can be computed rapidly; the resulting method scales well to large numbers of datasets, and the speed of the mixture modeling does not depend on the sizes of the earlier datasets. The largest weights in the mixture model point at the most relevant earlier datasets. The method is applicable to several types of measurement datasets, assuming that suitable models exist. Unlike traditional mixture modeling, we do not limit the form of the mixture components; thus, we bring in the knowledge built into the stored models of each dataset. We apply this approach to a large set of experiments from EBI's ArrayExpress gene expression database [12], treating each experiment in turn as a new dataset, queried against all earlier datasets. Under our assumptions, the retrieval results can be interpreted as studies that the authors of the study generating the query set could have cited, and we show that the actual citations overlap with the retrieval results. The discovered links between datasets additionally enable forming a ''hall of fame'' of gene expression studies, containing the studies that would have been influential, assuming the retrieval system existed. The links in the ''hall of fame'' verify and complement the citation links: in our study, they revealed corrections to the citation data, as two frequently retrieved studies were not highly cited and turned out to have erroneous publication entries in the database. We provide an online

resource for exploring and searching this "hall of fame": http://research.ics.aalto.fi/mi/setretrieval.

Earlier work on relating datasets has provided partial solutions along this line, with the major limitation of being restricted to pairwise dataset comparisons, in contrast to the proposed approach of decomposing a dataset into contributions from a set of earlier datasets. Russ and Futschik [13] represented each dataset by pairwise correlations of genes, and used them to compute dataset similarities. This dataset representation is ill suited for typical functional genomics experiments, as a large number of samples is required to sensibly estimate gene correlation matrices. In addition, it makes the dataset comparison computationally expensive, as the representation is bulkier than the original dataset. In other works, specific case-control designs [14] or known biological processes [15] are assumed; we generalize by using decompositions over arbitrary models.

In summary, our work is the first approach that allows data-driven retrieval of relevant datasets by decomposing a query dataset into contributions from several earlier datasets, without requiring specific designs for the earlier datasets or their models. Unlike existing state-of-the-art retrieval, our approach is not limited to available dataset annotation. Unlike the Pfam database [11], we not only store models but use them in retrieval. Unlike existing data-driven approaches [3,5] that match individual observations, we match whole datasets incorporating their experimental designs. We fully decompose datasets instead of only computing pairwise similarities, as in [13], and we allow decomposition over arbitrary models available for the datasets instead of requiring restricted settings, such as specific case-control designs [14] or known biological processes [15]. Unlike a hypothetical approach where a unified model of all data is built, our approach is fast and scalable to large data.

Combination of Stored Models for Dataset Retrieval

Our goal is to infer data-driven relationships between a new "query" dataset q and earlier datasets. The query is a dataset of N_q samples $\{x_i^q\}_{i=1}^{N_q}$; in the ArrayExpress study, the samples are gene expression profiles, with the element x_{ij}^q being expression of the gene set j in the sample i of the query q, but the setup is general and applicable to other experimental data, as well. Assume further a dataset repository of N_S earlier datasets, and assume that each dataset s_j, $j=1,\ldots,N_S$, has already been modeled with a model denoted by M^{s_j}, later called a base model. The base models are assumed to be probabilistic generative models, *i.e.*, principled data descriptions capturing prior knowledge and data-driven discoveries under specific distributional assumptions. Base models for different datasets may come from different model families, as chosen by the researchers who authored each dataset. In this paper, we use two types of base models, which are discrete variants of principal component analysis (*Results*), but any probabilistic generative models can be applied.

As an illustrative setting, suppose that the dataset repository contains several datasets arising from base experiments, so that each base experiment studies one known important biological effect, the experiment has been designed so that the effect is present in the resulting dataset, and together the base experiments cover the set of known important biological effects. In the special example case of metagenomics with known constituent organisms, an obvious set of base experiments would be the set of genomes of those organisms [16]. A new experiment could then be expressed as a combination of the base experiments, and potential novel effects. More generally, such as in a broad gene expression atlas, it would be hard, if not impossible, to settle on a clean, well-defined, and up-to-date base set of experiments to correspond to each known effect, so we chose to *use the comprehensive collection of experiments in the current databases as the base experiments*. The problem setting then changes from searching for a unique explanation of the new experiment to the down-to-earth and realistic task of data-driven retrieval of a set of relevant earlier experiments, relevant in the sense of having induced one or more of the known or as-of-yet unknown biological effects.

We combined the earlier datasets by a method that is probabilistic but simple and fast. We built a *combination model* for the query dataset as a mixture model of base distributions $p(x|M^{s_j})$, which have been estimated beforehand. In our scenario, generative models M^{s_j} are available in the repository along with datasets s_j; note that the M^{s_j} need not all have the same form. In the mixture model parameterized by $\Theta^q = \{\theta_j^q\}_{j=1}^{N_S+1}$, the likelihood of observing the query is

$$p(\{x_i^q\}_{i=1}^{N_q}; \Theta^q) = \prod_{i=1}^{N_q} \left[\left(\sum_{j=1}^{N_S} \theta_j^q p(x_i^q|M^{s_j}) \right) + \theta_{N_S+1}^q p(x_i^q|\psi) \right] \quad (1)$$

where θ_j^q is the mixture proportion or *weight* of the jth base distribution (model of dataset s_j), and $\theta_{N_S+1}^q$ is the weight for the novelty term. The novelty is modeled by a background model ψ, a broad nonspecific distribution covering overall gene-set activity across the whole dataset repository. All weights are non-negative and $\sum_{j=1}^{N_S+1} \theta_j^q = 1$. In essence, this representation assumes that biological activity in the query dataset can be approximately explained as a combination of earlier datasets and a novelty term.

The remaining task is to infer the combination model Θ^q for each query q given the known models M^{s_j} of datasets in the repository. We infer a maximum a posteriori (MAP) estimate of the weights $\Theta^q = \{\theta_j^q\}_{j=1}^{N_S+1}$. Alternatively, we could sample over the posterior, but MAP inference already yielded good results. We optimize the combination weights to maximize their (log) posterior probability

$$\begin{aligned} \log p(\{\theta_j^q\}|\{x_i^q\},\{M^{s_j}\}) &\propto \log p(\{x_i^q\}|\{M^{s_j}\},\{\theta_j^q\}) + \log p(\{\theta_j^q\}) \\ &\propto \sum_i \log \left[\left(\sum_{j=1}^{N_S} \theta_j^q p(x_i^q|M^{s_j}) \right) + \theta_{N_S+1}^q \, p(x_i^q|\psi) \right] - \lambda \sum_{j=1}^{N_S+1} {\theta_j^q}^2 \end{aligned} \quad (2)$$

where $p(\{\theta_j^q\}) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$ is a naturally non-sparse L_2 prior distribution for the weights with a regularization term λ. The cost function (2) is strictly concave (Text S1), and standard constrained convex optimization techniques can be used to find the optimized weights. Algorithmic details for the Frank-Wolfe algorithm and a proof of convergence are provided in Text S1. After computing the MAP estimate, we rank the datasets for retrieval according to decreasing combination weights.

This modeling-driven approach has several advantages: 1) the approximations become more accurate as more datasets are submitted to the repository, naturally increasing the number of base distributions; 2) it is fast, as only the models of the datasets are needed, not the large datasets themselves; 3) any model types can be included, as long as likelihoods of an observed sample can be computed; hence, all expert knowledge built into the models in the repository can be used; 4) relevant datasets are not assumed to be similar to the query in any naïve sense, as they only need to explain a part of the query set; 5) the relevance scores of datasets have a natural quantitative meaning as weights in the probabilistic combination model.

Scalability

As the size of repositories such as ArrayExpress doubles every two years or even more rapidly [17], fast computation with respect to the number N_S of background datasets is crucial for future-proof search methods. The first method above already has a fast linear computation time in N_S (Text S1), and an approximate variant can be run in sublinear time. For that, the model combination will be optimized only over the k background datasets most similar to the query, which can be found in time $O(N_S^{1/(1+\epsilon)})$ where $\epsilon \geq 0$ is an approximation parameter [18], by suitable hashing functions.

Results

Data-driven retrieval of experiments is more accurate than standard keyword search

We benchmarked the combination model against state-of-the-art dataset retrieval by keyword search, in the scenario in which a user queries with a new dataset against a database of earlier released datasets represented by models. The data were from a large human gene expression atlas [12], containing 206 public datasets with 5372 samples that have been systematically annotated and consistently normalized. To make use of prior biological knowledge, we preprocessed the data by gene set enrichment analysis [19], representing each sample by an integer vector telling for each gene set the number of leading edge active genes [20] (*Methods*). As base models, we used two model types previously applied in gene expression analysis [3,6,20,21]: a discrete principal component analysis method called Latent Dirichlet Allocation [22,23], and a simpler variant called mixture of unigrams [24] (Text S1). Of the two types, for each dataset, we

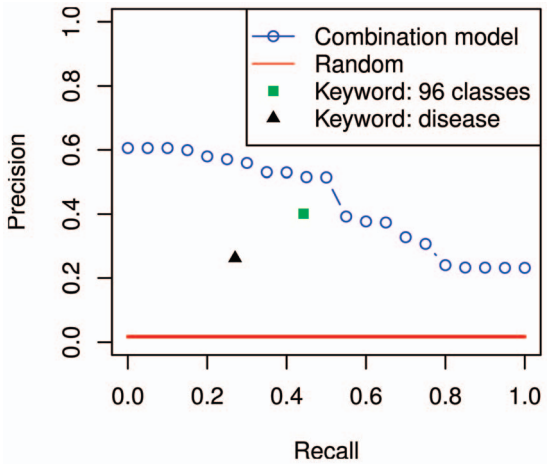

Figure 1. Data-driven retrieval outperforms the state of the art of keyword search on the human gene expression atlas [12]. Blue: Traditional precision-recall curve where progressively more datasets are retrieved from left to right. All experiments sharing one or more of the 96 biological categories of the atlas were considered relevant. In keyword retrieval, either the category names (''Keyword: 96 classes'') or the disease annotations (''Keyword: disease'') were used as keywords. All datasets having at least 10 samples were used as query datasets, and the curves are averages over all queries.

doi:10.1371/journal.pone.0113053.g001

chose the model yielding the larger predictive likelihood (Text S1). For each query (q), the earlier datasets (s_j) were ranked in descending order of the combination proportion (θ_j^q; estimated from Eq. (2)). That is, base models that explained a larger proportion of the gene set activity in the query were ranked higher. The approach yields good retrieval: the retrieval result was consistently better than with keyword searches applied to the titles and textual descriptions of the datasets (Fig. 1), which is a standard approach for dataset retrieval from repositories [25].

We checked that the result was not only due to laboratory effects by discarding, in a follow-up study, all retrieved results coming from the same laboratory. The mean average precision decreased slightly (from 0.44 to 0.42; precision-recall curve in Fig. S2) but still supports the same conclusion.

Network of computationally recommended dataset connections reveals biological relationships

When each dataset in turn is used as a query, the estimated combination weights form a ''relevance network'' between datasets (Fig. 2, left), where each dataset is linked to the relevant earlier datasets (for details, see *Methods* and an interactive searchable version at http://research.ics.aalto.fi/mi/setretrieval). The network structure is dominated but not fully explained by the tissue type. Normal and neoplastic solid tissues (cluster 1) are clearly separate from cell lines (cluster 2) and from hematopoietic tissue (cluster 4); the same main clusters were observed in [12]. Note that the model has not seen the tissue types but has found them from the data. Upon closer inspection of the clusters, some finer structure is evident. The muscle and heart datasets (gray) form an interconnected subnetwork

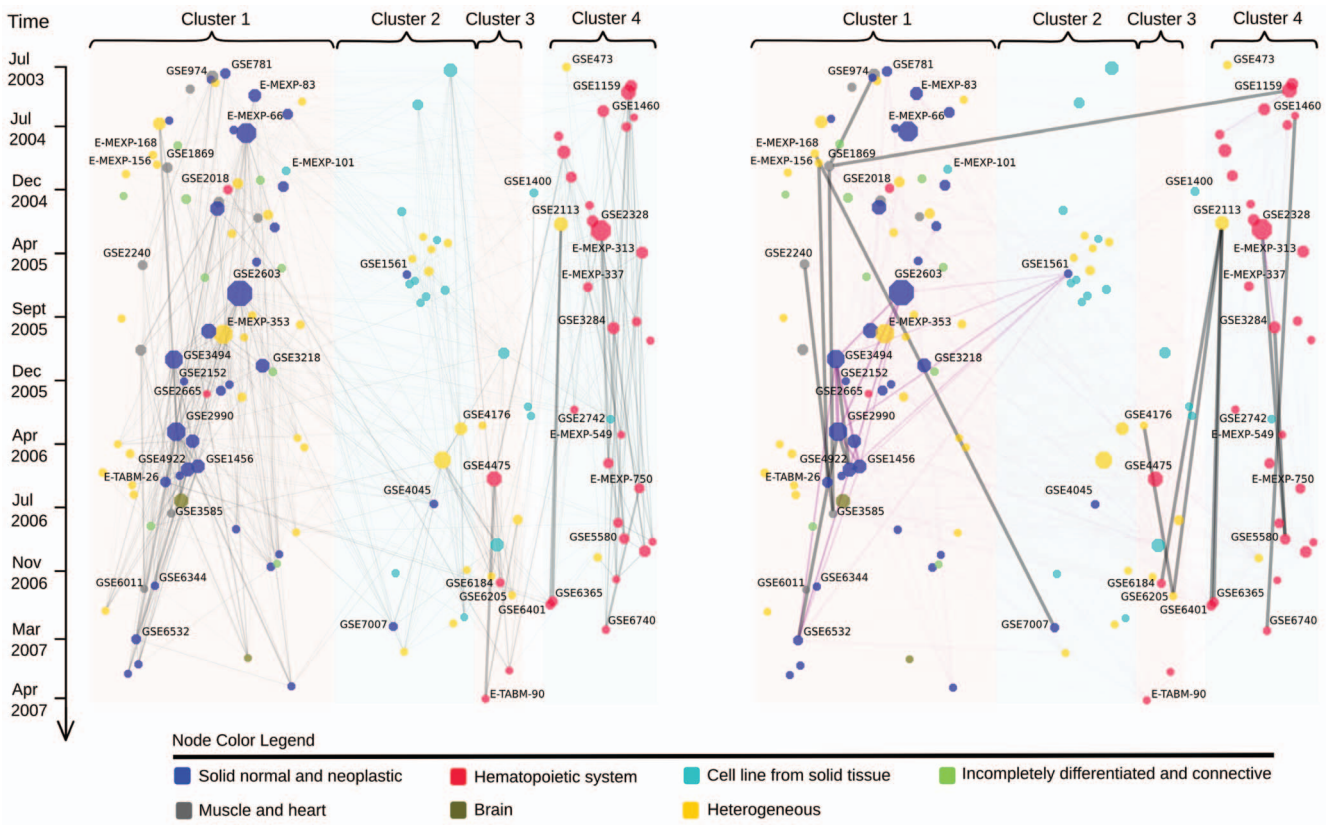

Figure 2. Relevance network of datasets in the human gene expression atlas; data-driven links from the model (left) and citation links (right). Left: each dataset was used as a query to retrieve earlier datasets; a link from an earlier dataset to a later one means the earlier dataset is relevant as a partial model of activity in the later dataset. Link width is proportional to the normalized relevance weight (combination weight θ_j^q; only links with $\theta_j^q \geq 0.025$ are shown, and datasets without links have been discarded). Right: links are direct (gray) and indirect (purple) citations. Node size is proportional to the estimated influence, *i.e.*, the total outgoing weight. Colors: tissue types (six meta tissue types [12]). The node layout was computed from the data-driven network (details in *Methods*).

doi:10.1371/journal.pone.0113053.g002

in the left edge of the image: nodes near the bottom of the image (downstream) are explained by earlier (upstream) nodes, which in turn are explained by nodes even further upstream. As another example, in cluster 4, myeloma and leukemia datasets are concentrated on the left side of the cluster, whereas the right side mostly contains normal or infected mononuclear cells.

There is a substantial number of links both across clusters and across tissue categories. Among the top 30 cross-category links, 25 involve heterogeneous datasets containing samples from diverse tissue origins. The strongest link connects GSE6365, a study on multile myeloma, with GSE2113, a larger study from the same lab, which largely includes the GSE6365 samples. The dataset E-MEXP-66 is a hub connected to all of the clusters and to nodes in its own cluster that have different tissue labels. It contains samples studying Kaposi sarcoma, and it also includes control samples from skin endothelial cells from blood vessels and the lymph system. Blood vessels and cells belonging to the lymph system are expected to be present in almost any solid tissue biopsy as well as in samples based

on blood samples. The strongest link between two homogeneous datasets of different tissue types connects GSE3307, which compares skeletal muscle samples from healthy individuals with 12 groups of patients affected by various muscle diseases, to GSE5392, which measures the transcriptome profiles of the normal brain and a brain with bipolar disorder. Interestingly, the shortening of telomeres has been associated both with bipolar disorder [26] and muscular disorder [27]. Treatment of bipolar disorder has been found to also slow down the onset of skeletal muscle disorder [28].

Next, we investigated ‘‘outlier" datasets where the tissue type does not match the main tissue types of a cluster, implying that they might reveal commonalities between cellular conditions across tissues. Cluster 1 contained three outlier datasets: two hematopoietic datasets and one cell line dataset. The two hematopoietic outlier datasets are studies related to macrophages and are both strongly connected to GSE2004, which contains samples from the kidney, liver, and spleen, sites of long-lived macrophages. The first hematopoietic outlier, GSE2018, studies bronchoalveolar lavage cells from lung transplant receipts; the majority of these cells are macrophages. The dataset has strong links to solid tissue datasets, including GSE2004, and the diverse dataset E-MEXP-66. The second hematopoietic outlier, GSE2665, is also strongly connected to GSE2004 and measures the expression of the lymphatic organs (sentinel lymph node) that contain sinusoidal macrophages and sinusoidal endothelial cells. The third outlier, E-MEXP-101, studies a colon carcinoma cell line and has connections to other cancer datasets in cluster 1.

Top dataset links overlap well with citation graph

We compared the model-driven network to the actual citation links (Fig. 2, right) to find out to what extent the citation practice in the research community matches the data-driven relationships. Of the top 200 data-driven edges, 50% overlapped with direct or indirect citation links (see *Methods*, Text S1 and Fig. S3). Most of the direct citations appear within the four tissue clusters (Fig. 2, right). The two cross-cluster citations are not due to the biological similarity of the datasets. The publication for GSE1869 cites the publication for GSE1159 regarding the method of differential expression detection. The GSE7007, a study on Ewing sarcoma samples, cites the study on human mesenchymal stem cells (E-MEXP-168), stating that the overall gene expression profiles differ between those samples.

We additionally compared the densely connected sets of experiments between the two networks. In the citation graph, the breast cancer datasets GSE2603, GSE3494, GSE2990, GSE4922, and GSE1456 form an interconnected clique in cluster 1, while the three leukocyte datasets GSE2328, GSE3284, and GSE5580 form an interconnected module in cluster 4. In the relevance network, the corresponding edges for both cliques are among the strongest links for those datasets, and some of them are among the top 20 strongest edges in the network (see Table S1 for the list of top 20 edges). There are also densely connected modules in the relevance network that are not strongly connected in the citation

graph; when we systematically sought cliques associated with each of the top 20 edges, the strongest edges constitute a clique among E-MEXP-750, GSE6740, and GSE473, all three studying CD4+ T helper cells, which are an essential part of the human immune system. Another interesting set is among three T-cell related datasets in cluster 3. Two of the datasets contain T lymphoblastic leukemia samples (E-MEXP-313 and E-MEXP-549), whereas E-MEXP-337 reports thymocyte profiles. Thymocytes are developing T lymphocytes that are matured in thymus, so this connection is biologically meaningful but not straightforward to find from dataset annotations. Other strongly connected cliques are discussed in Text S1.

Analysis of network hubs discovers datasets deserving more citations

Datasets that have high weights in explaining other datasets have a large weighted outdegree in the data-driven relevance network, and they are expected to be useful for many other studies. We checked whether the publications corresponding to these *central hubs* are highly cited in the research community. There is a low but statistically significant correlation between the weighted outdegree of datasets and their citation counts (Fig. 3; Spearman $\rho(169) = 0.2656$, $p < 0.001$). Both

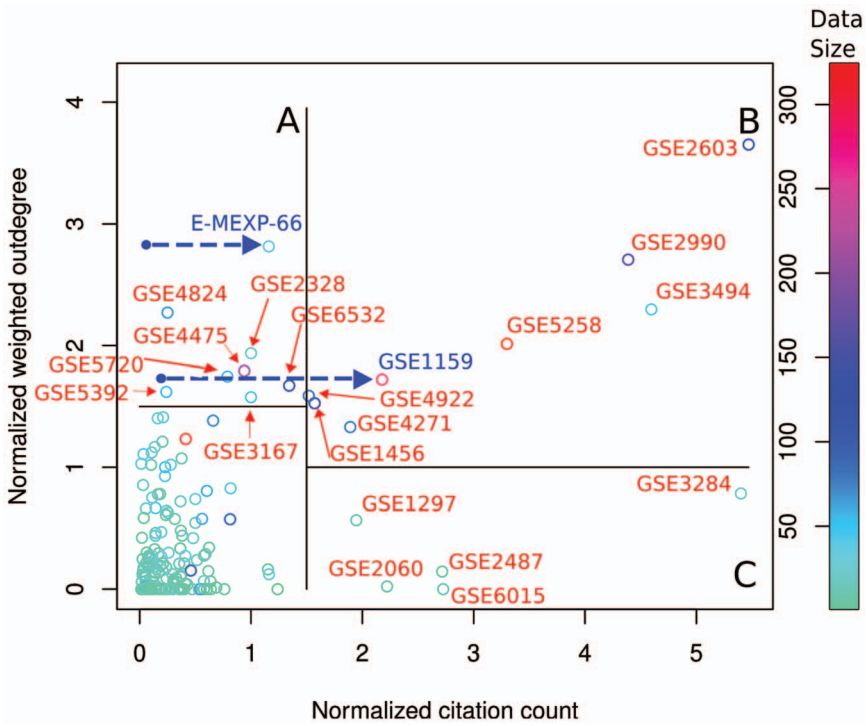

Figure 3. Data-driven prediction of usefulness of datasets vs. their citation counts. Manual checks comparing sets for which the two scores differed revealed inconsistent database records for two datasets; the blue arrows point to their corrected locations, which are more in line with the data-driven model. Regions A, B, and C: see text.

doi:10.1371/journal.pone.0113053.g003

quantities were normalized to avoid bias due to different release times of the datasets (*Methods*). We further examined whether the prestige of the publication venue (measured by impact factor) and the senior author (h-index of the last author) biased the citation counts, which could explain the low correlation between the outdegree and the citation count, and the answer was affirmative (*Methods*).

We inspected more closely the datasets where the recommended or the actual citation counts were high (Fig. 3): (A) datasets having low citation counts but high outdegrees, (B) datasets having both high citation counts and high outdegrees, and (C) datasets having high citation counts but low outdegrees. We manually checked the publication records of region A in Gene Expression Omnibus (GEO) [29] and ArrayExpress [17], to find out why the datasets had low citation counts despite their high outdegree (data-driven citation recommendations). Two of the eight datasets had an inconsistent publication record. The blue arrows in Fig. 3 point from their original position to the corrected position confirmed by GEO and ArrayExpress. Thus, the data-driven network revealed the inconsistency, and the new positions, corresponding to higher citation counts, validate the model-based finding that these datasets are good explainers for other datasets. In region B, most of the papers have been published in high-impact journals and have a relatively high number of samples (average sample size of 154) compared to region A (average sample size of 75). One of the eight datasets in the collection is the well-known Connectivity Map experiment (GSE5258). Lastly, the set C mostly contains unique targeted studies; there are five studies in the set, which are about leukocytes of injured patients, Polycomb group (PcG) proteins, senescence, Alzheimer's disease, and the effect of cAMP agonist forskolin, a traditional Indian medicine. The studies have been published in high-impact forums, and a possible reason of their low outdegree is their specific cellular responses, which are not very common in the atlas.

Discussion

Our main goal was to test the feasibility of the scenario where researchers let the data speak for themselves when relating new research to earlier studies. The conclusion is positive: even a relatively straightforward and scalable mixture modeling approach found both expected relationships such as tissue types, and relationships not easily found with keyword searches, including cells in different developmental stages or treatments resembling conditions in other cell types. While biologists could find such connections by bringing expert knowledge into keyword searches, the ultimate advantage of the data-driven approach is that it also yields connections beyond current knowledge, giving rise to new hypotheses and follow-up studies. For example, it seems surprising that the skeletal muscle dataset GSE6011 is linked also to kidney and brain datasets. Closer inspection yielded possible partial explanations. Some kidney areas are rich in blood vessels, lined by smooth muscle. Studies have shown common gene signatures between skeletal muscle and brain. Abnormal expression of the protein dystrophin leads to Duchenne muscular dystrophy, exhibited by a

majority of samples in GSE6011; the brain is another major expression site for dystrophin [30]. Interestingly, the top three potentially novel datasets, where only less than 50% of the expression pattern is modelled by earlier datasets (i.e., $\theta^{q}_{N_{S+1}} > 0.5$), are GSE2603 (a central breast cancer set), the Connectivity Map data (GSE5258), and the Burkitt's Lymphoma set (GSE4475, a cancer fundamentally distinct from other types of lymphoma). The first two are also recovered by the citation data (as they have relatively high citation counts and appear in region B in Fig. 3), unlike the third (which is part of region A in Fig. 3).

Our case study focused on a global analysis of the relevance network obtained for a representative dataset collection, allowing for comparisons with the citation graph. The data-driven relationships corresponded to actual citations when available but were richer and were able to spot out errors in citation links. Another intended use of the retrieval method is to support researchers in finding relevant data on a particular topic of interest. We performed a study with additional skeletal muscle datasets (Table S2) to obtain insights into relationships among skeletal muscle datasets (Text S1) as well as between skeletal muscle and other datasets (Text S1 and Table S3), and we showed that the retrieval method lessens the need for laborious manual searches (Text S1 and Fig. S4).

In this work, we made simplifying assumptions: we only employed two model families, included biological knowledge only as pre-chosen gene sets, and assumed all new experiments to be mixtures of earlier ones, instead of finding common effects in them and combining them either as mixtures or sums. We expect the results to improve considerably with more advanced future alternatives, with the research challenge being to maintain scalability. Generalizability of the search across measurement batches, laboratories, and measurement platforms is a challenge. Our feasibility study showed that for carefully preprocessed datasets (of the microarray atlas [12]), data-driven retrieval is useful even across laboratories. Our method is generally applicable to any single platform, and it takes into account the expert knowledge built into models of datasets for that platform; abstraction-based data representations, such as the gene set enrichment representation we used, have the potential to facilitate cross-platform analyses. As data integration approaches develop further [31,32], it may be possible to do searches even across different omics types; here, integration of meta data (pioneered in a specific semi-supervised framework [33]), several ontologies (MGED ontology, experimental factor ontology, and ontology of biomedical investigations [34]) and text mining results [35,36] are obviously useful first steps.

Materials and Methods

Gene expression data

We used the human gene expression atlas [12] available at ArrayExpress under accession number E-MTAB-62. The data were preprocessed by gene set enrichment analysis (GSEA) using the canonical pathway collection (C2-CP) from

the Molecular Signatures Database [19]. Each sample was represented by its top enriched gene sets [20] (Text S1).

Node layout and normalized relevance weight

The weight matrix contains a weight vector for each query dataset, encoding the amount of variation in that query explained by each earlier dataset. As query datasets from early years have only a few even earlier sets available, there is a bias towards the edges being stronger for the datasets from early years. To remove the bias we normalized, for the visualizations, the edge strengths of each query data set by the number of earlier datasets. To visualize the relationship network over time in Fig. 2, we needed a layout algorithm that positions the datasets on the horizontal axis highlighting structure and avoiding tangling. We used a *cluster-emphasizing* Sammon's mapping; Sammon's mapping [37] is a nonlinear projection method or multidimensional scaling algorithm that aims at preserving the interpoint distances (here $1-\theta_j^q$). By clustering the network (with unsupervised Markov clustering [38]) and increasing between-cluster distances by adding a constant ($c=1$) to them, the mapping was made to emphasize clusters and hence untangle the layout.

Citation graph

Direct citations between dataset-linked publications were extracted from the Web of Science (26 Jul 2012) and PubMed (17 Oct 2012). We additionally considered two types of indirect edges. Firstly, we introduced links between datasets whose publications share common references. This covers, for instance, related datasets whose publications appeared close in time, making direct citation unlikely. A natural measure of edge strength is given by the number of shared references. Secondly, we connect datasets whose articles are cited together, because co-citation is a sign that the community perceives the articles as related. Here, the edge strength was taken to be the number of articles co-citing the two dataset publications; these edges dominate the indirect links in the citation graph. For this analysis, we used citation data, available for 171 datasets and provided by Thomson Reuters as of 13 September 2012.

Normalization of citation counts and weighted outdegrees

As early datasets have many more papers that can cite them and many more later datasets that they can help model, both the citation counts and estimated weighted outdegrees are expected to be upwards biased for them. For Fig. 3, we normalized the quantities; for each dataset, we normalized the outdegree by the number of newer datasets and the citation count by the time difference between publishing the data and the newest dataset in the atlas. To make sure the normalization did not introduce side effects, we additionally checked that the same conclusions were reached without the citation count normalization (Fig. S1;

plotted as stratified subfigures for each 1-year time window). The citation counts were extracted from PubMed on 16 May 2012.

Citation counts are strongly influenced by external esteem of the publication forum and the senior author

We stratified the data sets according to the numbers of data-driven citation recommendations, and studied whether the impact factor of the forum or the h-index of the last author were predictive of the actual citation count in each stratum. The strata were the top and bottom quartiles, and for each, we compared the top and bottom quartiles of the actual citation counts (resulting in comparing the four corners of Fig. 3). For low outdegree (low recommended citation count), the h-index was lower for less cited datasets ($t_{11} = 2.78, p = 0.0086$; mean value 24.20 vs 54.62), and the impact factor was lower ($t_7 = 2.6, p = 0.016$; mean value 4.38 vs 21.13). Similarly, for the high recommended citation count, the impact factor for the little-cited datasets was lower ($t_{19} = 3.99, p = 4.0^{-4}$; mean value 6.45 vs 21.91), while the difference in h-index was not significant. All t statistics and p-values were computed by one-sided independent sample Welch's t-tests. The h-indices and impact factors were collected from Thomson Reuters Web of Knowledge and Journal Citation Reports 2011, respectively, on 23rd July 2012.

Supporting Information

Figure S1. Stratified data-driven prediction of usefulness of datasets vs. their citation counts. Black solid lines mark the boundary for potentially interesting datasets; the boundaries are set to hold the same percentiles of data as in Fig. 3 in the main paper. *ImpFac* stands for Impact Factor of the publication venue.
doi:10.1371/journal.pone.0113053.s001 (TIFF)

Figure S2. Removal of laboratory effects changes the retrieval performance only slightly, as measured by the precision-recall curves. *Original*: Replicated from Fig. 1 of the main paper; *Lab. effects removed*: all retrieval results from the same laboratory as the query data have been discarded.
doi:10.1371/journal.pone.0113053. s002 (TIFF)

Figure S3. Overlap of data-driven recommendations with the actual citation graph: Precision @k for top edges that explain more than 2.5% variation. The gold standard is the extended citation graph, which is built as the union of edges from 1) the original directed graph, 2) between any two articles that are cited together by some other article, and 3) between any two articles that have at least one common reference.
doi:10.1371/journal.pone.0113053.s003 (TIFF)

Figure S4. Retrieval performance evaluation of the data-driven model against keyword search in the skeletal muscle case study. The precision-recall curves are averaged across the 16 skeletal muscle datasets having at least 10 samples.
doi:10.1371/journal.pone.0113053.s004 (TIFF)

Table S1. Top 20 strongest edges in the relevance network.
doi:10.1371/journal.pone.0113053.s005 (XLSX)

Table S2. ArrayExpress accession numbers of 16 skeletal muscle datasets used in the retrieval case study in addition to the human gene expression atlas [12]. All datasets were measured with the human genome platform HG-U133A, the same used in the atlas.
doi:10.1371/journal.pone.0113053.s006 (XLSX)

Table S3. Skeletal muscle queries with at least one retrieved non-skeletal muscle dataset, sorted according to decreasing precision.
doi:10.1371/journal.pone.0113053.s007 (XLSX)

Text S1. More details on methods and results.
doi:10.1371/journal.pone.0113053.s008 (PDF)

Acknowledgments

We thank Matti Nelimarkka and Tuukka Ruotsalo for helping with citation data. Certain data included herein are derived from the following indices: Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, prepared by Thomson Reuters, Philadelphia, Pennsylvania, USA, Copyright Thomson Reuters, 2011.

Author Contributions

Conceived and designed the experiments: AF JP EG SK. Performed the experiments: AF EG. Analyzed the data: AF EG JR. Wrote the paper: AF JP EG JR SK.

References

1. **Greene CS, Troyanskaya OG** (2011) PILGRM: An interactive data-driven discovery platform for expert biologists. Nucleic Acids Res 39: W368–374.
2. **Tanay A, Steinfeld I, Kupiec M, Shamir R** (2005) Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. Mol Syst Biol 1: e1–10.
3. **Caldas J, Gehlenborg N, Kettunen E, Faisal A, Rönty M, et al.** (2012) Data-driven information retrieval in heterogeneous collections of transcriptomics data links *SIM2s* to malignant pleural mesothelioma. Bioinformatics 28: i246–i253.
4. **Adler P, Kolde R, Kull M, Tkachenko A, Peterson H, et al.** (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. Genome Biol 10: R139.
5. **Schmid PR, Palmer NP, Kohane IS, Berger B** (2012) Making sense out of massive data by going beyond differential expression. Proc Natl Acad Sci U S A 109: 5594–5599.
6. **Gerber GK, Dowell RD, Jaakkola TS, Gifford DK** (2007) Automated discovery of functional generality of human gene expression programs. PLoS Comput Biol 3: e148.
7. **Tseng GC, Ghosh D, Feingold E** (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res 40: 3785–3799.
8. **Rung J, Brazma A** (2012) Reuse of public genome-wide gene expression data. Nature Rev Genet 14: 89–99.

9. **Baxter J** (1997) A Bayesian/information theoretic model of learning to learn via multiple task sampling. Machine Learning 28: 7–39.
10. **Caruana R** (1997) Multitask learning. Machine Learning 28: 41–75.
11. **Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al.** (2012) The Pfam protein families database. Nucleic Acids Research 40: D290–D301.
12. **Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, et al.** (2010) A global map of human gene expression. Nat Biotechnol 28: 322–324.
13. **Russ J, Futschik ME** (2010) Comparison and consolidation of microarray data sets of human tissue expression. BMC Genomics 11: 305.
14. **Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ** (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol 6: e1000662.
15. **Huttenhower C, Troyanskaya OG** (2008) Assessing the functional structure of genomic data. Bioinformatics 24: i330–338.
16. **Meinicke P, Asshauer KP, Lingner T** (2011) Mixture models for analysis of the taxonomic composition of metagenomes. Bioinformatics 27: 1618–1624.
17. **Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al.** (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37: D868–72.
18. **Gionis A, Indyk P, Motwani R** (1999) Similarity search in high dimensions via hashing. In: Proc 25th VLDB Conf. San Francisco, CA: Morgan Kaufmann, pp. 518–529.
19. **Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al.** (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.
20. **Caldas J, Gehlenborg N, Faisal A, Brazma A, Kaski S** (2009) Probabilistic retrieval and visualization of biologically relevant microarray experiments. Bioinformatics 25: i145–i153.
21. **Engreitz JM, Morgan AA, Dudley JT, Chen R, Thathoo R, et al.** (2010) Content-based microarray search using differential expression profiles. BMC Bioinformatics 11: 603.
22. **Pritchard JK, Stephens M, Donnelly P** (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
23. **Blei DM, Ng AY, Jordan MI, Lafferty J** (2003) Latent Dirichlet allocation. J Mach Learn Res 3: 993–1022.
24. **Nigam K, McCallum A, Thrun S, Mitchell T** (2000) Text classification from labeled and unlabeled documents using EM. Machine Learning 39: 103–134.
25. **Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y** (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. Bioinformatics 24: 2798–2800.
26. **Martinsson L, Wei Y, Xu D, Melas PA, Mathé AA, et al.** (2013) Long-term lithium treatment in bipolar disorder is associated with longer leukocyte telomeres. Transl Psychiatry 3: e261
27. **Mourkioti F, Kustan J, Kraft P, Dav JW, Zhao MM, et al.** (2013) Role of telomere dysfunction in cardiac failure in Duchenne muscular dystrophy. Nature Cell Bio 15: 895–904.
28. **Kitazawa M, Trinh DN, LaFerla FM** (2008) Inflammation induces tau pathology in inclusion body myositis model via glycogen synthase kinase-3 beta. Ann Neurol 64: 15–24.
29. **Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al.** (2011) NCBI GEO: archive for functional genomics data sets-10 years on. Nucleic Acids Res 39: D1005–D1010.
30. **Culligan K, Glover L, Dowling P, Ohlendieck K** (2001) Brain dystrophin-glycoprotein complex: Persistent expression of beta-dystroglycan, impaired oligomerization of Dp71 and up-regulation of utrophins in animal models of muscular dystrophy. BMC Cell Biol 2: 2.
31. **Tripathi A, Klami A, Orešič M, Kaski S** (2011) Matching samples of multiple views. Data Min Knowl Discov 23: 300–321.

32. **Virtanen S, Klami A, Khan SA, Kaski S** (2012) Bayesian group factor analysis. In:, Lawrence N, Girolami M, , editors. International Conference on Artificial Intelligence and Statistics. Vol. 22 of *JMLR W&CP*, pp., 1269–1277.

33. **Wise A, Oltvai Z, Bar-Joseph Z** (2012) Matching experiments across species using expression values and textual information. Bioinformatics 28: i258–i264.

34. **Zheng J, Stoyanovich J, Manduchi E, Liu J, Stoeckert CJ** (2011) Annotcompute: annotation-based exploration and meta-analysis of genomics experiments. Database: Oxford. doi:10.1093/database/bar045

35. **Jensen LJ, Saric J, Bork P** (2006) Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet 7: 119–129.

36. **Rzhetsky A, Seringhaus M, Gerstein M** (2008) Seeking a new biology through text mining. Cell 134: 9–13.

37. **Sammon JW** (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comput 18: 401–409.

38. **van Dongen S** (2000) Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht.