Targeted exploration and analysis of large cross-platform human transcriptomic compendia

Qian Zhu[1,2], Aaron K Wong[1,2], Arjun Krishnan[2], Miriam R Aure[3], Alicja Tadych[2], Ran Zhang[2,4], David C Corney[2,4], Casey S Greene[5,6], Lars A Bongo[7], Vessela N Kristensen[3,8,9], Moses Charikar[1], Kai Li[1] & Olga G Troyanskaya[1,2,10]

We present SEEK (search-based exploration of expression compendia; http://seek.princeton.edu/), a query-based search engine for very large transcriptomic data collections, including thousands of human data sets from many different microarray and high-throughput sequencing platforms. SEEK uses a query-level cross-validation–based algorithm to automatically prioritize data sets relevant to the query and a robust search approach to identify genes, pathways and processes co-regulated with the query. SEEK provides multigene query searching with iterative metadata-based search refinement and extensive visualization-based analysis options.

The accumulation of human gene expression data in public repositories, such as The Cancer Genome Atlas[1] and Gene Expression Omnibus[2], offers unprecedented opportunities for data-driven characterization of biological pathways that underlie human diseases. Unsupervised, exploratory approaches are particularly suitable for data-driven discovery and in settings with insufficient or biased training data. However, traditional unsupervised methods, such as clustering and biclustering[3,4], do not readily extend to compendia containing thousands of data sets from different expression technologies and platforms. Query-based search can enable biomedical researchers to effectively explore and analyze the large collection of expression data sets to identify coexpressed genes. With these results, scientists can explore functional relationships and make inferences about pathway function with regard to query genes of interest. However, existing search approaches are limited to smaller compendia in model organisms[5,6] or, in human, to identifying similar arrays[7] or performing gene-level search on a single microarray platform[8].

We present SEEK, a robust cross-platform search system capable of handling large human expression data sets across multiple expression platforms, including microarray and high-throughput sequencing technologies, and automatically prioritizing data sets relevant to the user's single- or multiple-gene query to identify genes co-regulated with the query (**Supplementary Figs. 1–6**). SEEK provides biomedical researchers with a systems-level, unbiased exploration of diverse human pathways, tissues and diseases represented in the entire heterogeneous human compendium. The system integrates thousands of data sets on the fly using a novel cross-validation–based data set–weighting algorithm, which robustly identifies relevant data sets and leverages them to retrieve genes co-regulated with the query. It supports sophisticated biological search contexts defined by multigene queries and enables cross-platform analysis, with the current compendium including 155,025 experiments spanning 5,210 data sets from 41 different microarray and RNA-seq platforms (**Fig. 1a** and **Supplementary Data 1**). It has been implemented in a user-friendly interactive web interface (http://seek.princeton.edu/), which includes expression visualization and interpretation modules (**Fig. 1a**). This interface facilitates hypothesis generation by providing (i) intuitive expression visualizations of the retrieved coexpressed genes, (ii) explorations of individual data sets to establish associations between coexpressed genes and biological variables and (iii) further refinement of the search results, such as limiting data sets to a specific tissue or disease.

The search algorithm (Online Methods) allows for multigene queries and includes a gene connectivity, or 'hubbiness'[9,10], correction procedure, a novel cross-validation data set–weighting method, and a summarization procedure to calculate the final score for each gene. Prior to application of the search algorithm, the data compendium is preprocessed to make correlation distributions comparable across data sets. Then a hubbiness correction procedure is applied to remove biases caused by generically well-coexpressed genes not specific to the user's area of interest, which is defined by the query. The data set–weighting algorithm then prioritizes relevant data sets according to the query. The idea is to upweight data sets from which a subset of the query genes can retrieve the remaining query genes well on the basis of normalized, hubbiness-corrected coexpression in that data set (cross-validation–based weighting). This approach is effective even when not all query genes are coexpressed. Finally, the

[1]Department of Computer Science, Princeton University, Princeton, New Jersey, USA. [2]Lewis-Sigler Institute of Integrative Genomics, Princeton University, Princeton, New Jersey, USA. [3]Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Oslo, Norway. [4]Department of Molecular Biology, Princeton University, Princeton, New Jersey, USA. [5]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, USA. [6]Institute for Quantitative Biomedical Sciences, Dartmouth College, Hanover, New Hampshire, USA. [7]Department of Computer Science, University of Tromsø, Tromsø, Norway. [8]Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway. [9]Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital, Akershus, Norway. [10]Simons Center for Data Analysis, Simons Foundation, New York, New York, USA. Correspondence should be addressed to M.C. (moses@cs.princeton.edu), K.L. (li@cs.princeton.edu) or O.G.T. (ogt@cs.princeton.edu).

RECEIVED 5 SEPTEMBER 2014; ACCEPTED 12 NOVEMBER 2014; PUBLISHED ONLINE 12 JANUARY 2015; DOI:10.1038/NMETH.3249

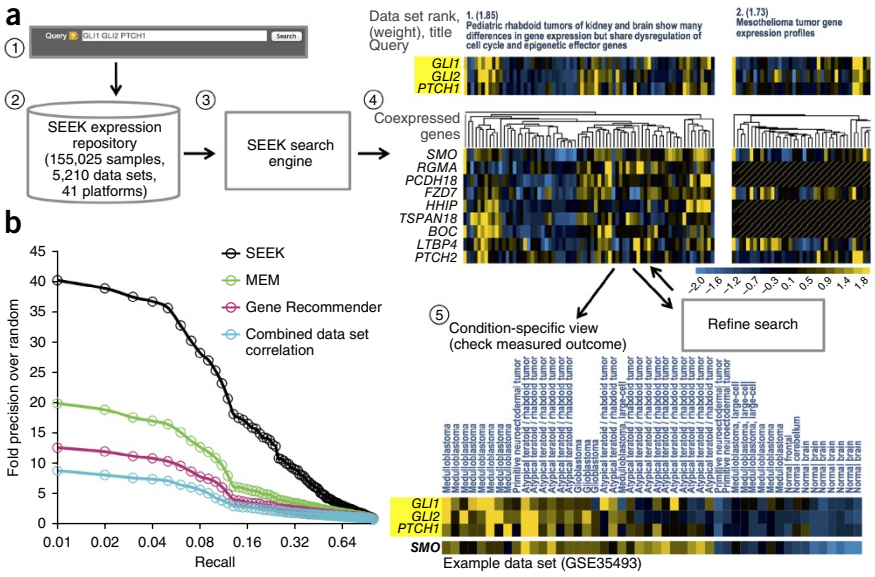

Figure 1 | SEEK system overview and systematic functional evaluation. (**a**) Users begin by providing a query gene set of interest to define a biological context of their search (step 1). SEEK searches the entire compendium and returns genes that are coexpressed with the query and the top relevant data sets (steps 2 and 3). The web user interface provides visualizations of gene coexpressions across prioritized data sets (step 4) and enables users to iteratively refine their search (**Fig. 2**) and further analyze the results through a condition-specific view (step 5) (**Supplementary Note 4**). (**b**) Gene-retrieval evaluations across 995 diverse GO biological process terms for the SEEK, MEM, Gene Recommender and combined data set correlation algorithms (**Supplementary Note 2**). Queries of diverse sizes (2–20 genes) were selected randomly among each term's genes to evaluate the precision of retrieving the remaining genes in each term. Individual term performances (**Supplementary Data 2**) and additional detailed comparative evaluations (**Supplementary Figs. 1** and **2**) are provided.

integrated gene scores are calculated on the basis of the data set weights and genes' coexpression patterns in each data set to provide a final gene ranking.

SEEK is based on measuring coexpressions, which minimizes biases toward prior knowledge, and accurately extracts functional information without need to explicitly model outcome variables such as treatment and control experiments (**Fig. 1b** and prior works[5,6,8,11]). The use of coexpression thus enables the robust integration of a large number of data sets from diverse tissues, cell lines and disease origins, generated from diverse platforms, and such an approach can be extended to make functional comparisons across organisms. A key challenge here is that the search results can be polluted by batch effects[12], poor-quality data sets or even good-quality data sets irrelevant to the user's query context. Yet the detailed, targeted correction of these issues in each data set or modeling of each outcome variable is impossible in the context of a large, multiplatform compendium. SEEK's data set–weighting algorithm addresses this challenge by enabling multigene query support for constructing expressive search contexts and by using a discriminative algorithm for identifying which data sets are relevant and accurate in representing query-related biological processes. This algorithm automatically downweights low-quality data sets (**Supplementary Fig. 7** and **Supplementary Note 1**) and provides accurate retrieval of functionally related genes and data sets (**Fig. 1b** and **Supplementary Figs. 1** and **2**).

SEEK was accurate and robust in a large-scale gene-retrieval assessment across a diverse array of biological contexts. Specifically, we constructed over 129,000 queries spanning 995 human Gene Ontology (GO) biological process gene sets (by choosing subsets of genes from each process) and evaluated the ability of the algorithm to retrieve the remaining genes in the process (Online Methods). This setup was designed to simulate realistic situations in which the query genes are biologically coherent but are not necessarily well coexpressed and in which users are interested in identifying genes functionally related to the query (in this case, members of the same biological process). SEEK's performance was robust across a wide range of pathways (**Supplementary Data 2**), and it consistently outperformed previous search approaches, including the only query-based human search system, MEM[8]; Gene Recommender[6] (not available for human as a resource); and the correlations on the combined data set (**Fig. 1b** and **Supplementary Note 2**). Furthermore, our evaluation demonstrated that SEEK's support for multigene queries enhances the algorithm's ability to effectively weight relevant data sets from the compendium (**Supplementary Fig. 1a**) and that the algorithm is robust with respect to query noise (**Supplementary Fig. 2**).

Notably, our evaluation demonstrated the benefits of robust search of a compendium with thousands of expression data sets, as SEEK's performance improved with the inclusion of more microarray and RNA-seq data sets in the compendium, assessed by subsampling our large compendium to create smaller subsets (**Supplementary Fig. 3** and **Supplementary Data 3**). Furthermore, being able to integrate the full scale of the existing human gene expression data allows the approach to support focused queries covering diverse areas of biology (**Supplementary Fig. 4**), providing strong performance across varied processes including erythrocyte differentiation (44-fold improvement of precision over random (FIOR) at 10% recall) and glutamate signaling (104-fold) (**Supplementary Fig. 4**). In contrast, using the most relevant single data set for the same query yielded weak performance of just 3- and 6-FIOR for the two processes, respectively, thus demonstrating the value of using the entire compendium.

We illustrated the power of SEEK and multigene queries by using SEEK to identify genes dysregulated in the Hedgehog (Hh) pathway and the corresponding tissues and disease states where the Hh pathway is hyperactivated. We used Hh genes *GLI1*, *GLI2* and *PTCH1* as the query, where transcription factors GLI1 and GLI2 have been suggested as pathway markers of Hh signaling[13]. By examining this query in the context of a large compendium of expression data sets (**Fig. 2a** and **Supplementary Fig. 5**), we observed a wide prevalence of aberrant Hh signaling across many diseased tissues (**Supplementary Fig. 5**). The top-ranked data sets had substantially higher weights, indicating the presence of a strong query-related signal in these data (**Supplementary**

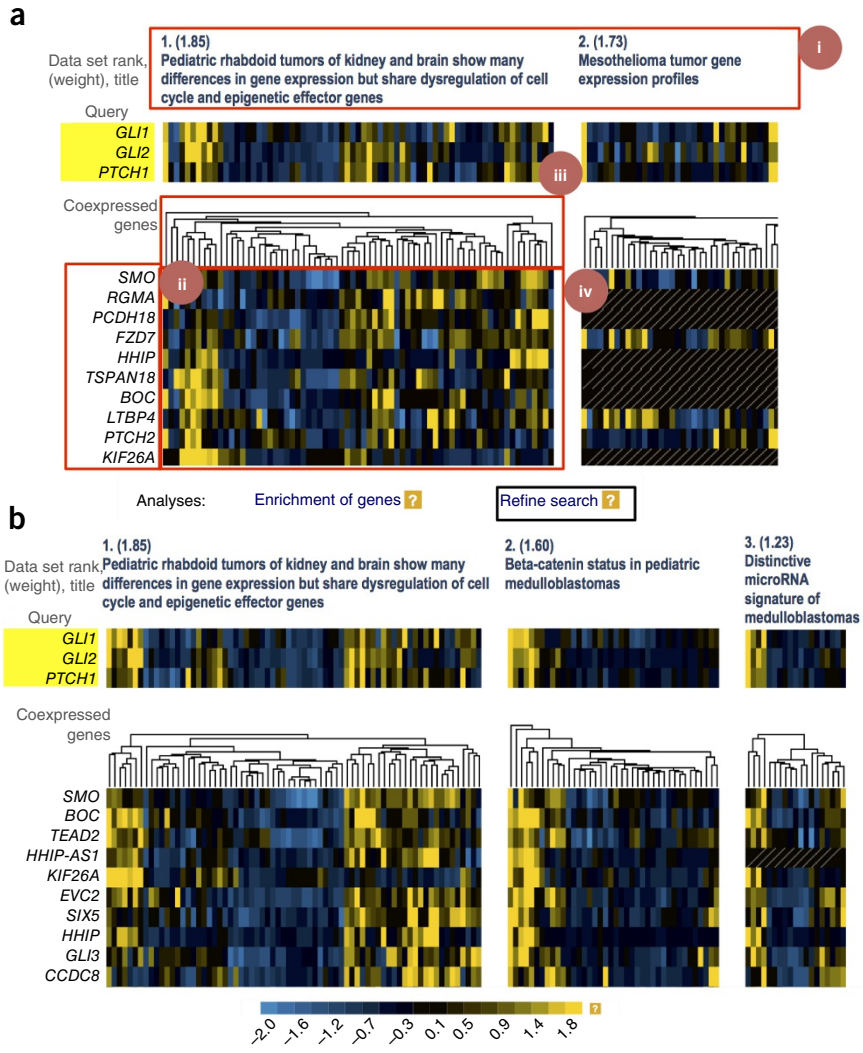

Figure 2 | Search results for the Hedgehog (Hh) query (*GLI1*, *GLI2*, *PTCH1*) and search refinement. (**a**) Data sets prioritized and genes retrieved for the query in the main result page, shown in expression view. The top-ranked data sets (i) and the coexpressed gene list (ii) are indicated. Conditions in each data set are hierarchically clustered in real time according to the expression values of the top genes retrieved from the search (iii), and an expression heat map of the genes for each data set is provided (iv). (**b**) Illustration of the search refinement function. "Refine search" enables users to narrow the scope of their search through selection criteria including tissue, cell type or disease categories; platforms; or rank of data sets from initial search (**Supplementary Note 4**). Top search results after limiting the search scope to brain data sets are shown. Brain-specific coexpressions are noted in this case with higher coexpression scores to the query and better groupings of conditions than those of the initial search. SEEK also has alternative view modes such as coexpression view and condition-specific view (**Supplementary Note 4**).

Fig. 5), and appeared to be more specific to the Hh query than to random queries (**Supplementary Fig. 6a**). These highly weighted data sets included results from studies of tumors with previously documented connections to aberrant Hh signaling, such as (i) medulloblastoma, in which overactivation of Hh has been documented[14,15], (ii) human germ cell tumors, in which Hh pathway mutations have been linked to aberrant Hh activation in human germ cells[16], and (iii) malignant rhabdoid tumors[17,18], in which mutations have been found to lead to Hh signaling activation[18]. Thus, SEEK correctly identified data sets relevant to the Hh signaling and helped explore the important role of the Hh pathway in a wide array of cancer types. The data set weighting led to accurate retrieval of other genes in the Hh pathway, including those encoding Hh pathway signaling receptors and their associated genes *SMO*, *PTCH2*, *HHIP*, *BOC*[19], the *Cos2* homolog *KIF7* (ref. 20) (**Fig. 2a** and **Supplementary Fig. 6b**) as well as additional genes associated with Hh dysregulation in cancer (**Supplementary Note 3**).

The SEEK interface can visualize the aforementioned results—including the top-ranked data sets, genes and coexpression profiles—using flexible and interactive visualizations (**Fig. 2a**). The main search result page provides users with the ability to perform extensive follow-up analyses, including functional analysis of results with a coexpression view that summarizes the query and retrieved genes' coexpression across 50 data sets at a time (**Supplementary Note 4**). Users can also examine the behavior of any gene in a given data set in detail through a condition-specific view (**Fig. 1a**), where they can examine associations between coexpressed genes and treatments or outcomes on the basis of data set metadata. An additional post-search analysis, the search refinement function, allows users to iteratively refine their search by limiting the scope of the query search to data sets of a specific disease or tissue of interest (**Fig. 2b**). This feature currently provides customized search over not merely the 2,685 cancer data sets of various tissue origins but also almost 2,000 noncancer data sets, including nearly 280 stem cell, over 100 neurodegenerative disease and 1,400 various immune and other cell type related data sets (**Supplementary Data 4**). We plan to regularly update SEEK's compendium as new microarray and RNA-seq data sets become publicly available.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The work was supported in part by the US National Institutes of Health (NIH) award R01 HG005998 and partially supported by the US National Science Foundation (NSF) CAREER award (DBI-0546275) and NIH award R01 GM071966 to O.G.T. The project was also partially supported in part by the NIH awards T32 HG003284 and P50 GM071508. M.C. was supported by NSF awards CCF 1218687 and CCF 1302518. O.G.T. receives support as part of the Canadian Institute For Advanced Research in the Genetic Networks group. We thank members of the Troyanskaya lab for comments about SEEK in the regular lab meetings and Q. Zhu for critically reading the manuscript. We thank the volunteers from Princeton and other universities, including the Canadian Institute for Advanced Research

Genetic Networks meetings attendees, for testing the SEEK web interface and providing valuable feedback.

AUTHOR CONTRIBUTIONS

Q.Z. and O.G.T. wrote the manuscript. Q.Z., O.G.T., K.L. and M.C. designed the algorithm. Q.Z. implemented the search back end and front end, and performed evaluations. A.K.W., D.C.C., A.K., R.Z., M.R.A. and C.S.G. contributed ideas, performed analyses and edited the manuscript. A.T., L.A.B. and Q.Z. performed data and metadata processing. V.N.K. and O.G.T. contributed ideas in the biological study. O.G.T., K.L. and M.C. conceived of the study and gave guidance.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. The Cancer Genome Atlas Research Network. *Nature* **455**, 1061–1068 (2008).
2. Edgar, R., Domrachev, M. & Lash, A.E. *Nucleic Acids Res.* **30**, 207–210 (2002).
3. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
4. Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. *Proc. Natl. Acad. Sci. USA* **101**, 2981–2986 (2004).
5. Hibbs, M.A. *et al. Bioinformatics* **23**, 2692–2699 (2007).
6. Owen, A.B., Stuart, J., Mach, K., Villeneuve, A.M. & Kim, S. *Genome Res.* **13**, 1828–1837 (2003).
7. Zinman, G.E., Naiman, S., Kanfi, Y., Cohen, H. & Bar-Joseph, Z. *Nat. Methods* **10**, 925–926 (2013).
8. Adler, P. *et al. Genome Biol.* **10**, R139 (2009).
9. Barabási, A.-L. & Oltvai, Z.N. *Nat. Rev. Genet.* **5**, 101–113 (2004).
10. Han, J.-D.J. *et al. Nature* **430**, 88–93 (2004).
11. Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. & Pavlidis, P. *Genome Res.* **14**, 1085–1094 (2004).
12. Leek, J.T. *et al. Nat. Rev. Genet.* **11**, 733–739 (2010).
13. Kimura, H., Stephen, D., Joyner, A. & Curran, T. *Oncogene* **24**, 4026–4036 (2005).
14. Oliver, T.G. *et al. Proc. Natl. Acad. Sci. USA* **100**, 7331–7336 (2003).
15. Berman, D.M. *et al. Science* **297**, 1559–1561 (2002).
16. Carpenter, D. *et al. Proc. Natl. Acad. Sci. USA* **95**, 13630–13634 (1998).
17. Oue, T., Yoneda, A., Uehara, S., Yamanaka, H. & Fukuzawa, M. *J. Pediatr. Surg.* **45**, 387–392 (2010).
18. Jagani, Z. *et al. Nat. Med.* **16**, 1429–1433 (2010).
19. Cohen, M.M. Jr. *Am. J. Med. Genet. A.* **123A**, 5–28 (2003).
20. Cheung, H.O.-L. *et al. Sci. Signal.* **2**, ra29 (2009).

ONLINE METHODS

Data preparation and correlation normalization. SEEK assembles its human gene expression compendium by obtaining data sets from NCBI's Gene Expression Omnibus (GEO) database[2] and the Cancer Genome Atlas (TCGA)[1]. The compendium consists of data sets from 41 platforms including 32 platforms from Affymetrix, Agilent and Illumina, and 9 RNA sequencing platforms (**Supplementary Data 1**). These platforms were chosen on the basis of the number of available data sets and the availability of raw data to perform consistent processing for each platform. The data sets were processed consistently by applying platform-specific procedures on their raw measurements (**Supplementary Note 5** and **Supplementary Data 5**) to remove the systematic differences among data sets[21]. The normalized data sets containing gene-level expression values can be accessed through the SEEK website.

The next step of data processing is calculating the Pearson correlations $r_d(x, y)$ between all pairs of genes x and y in each data set d. As correlation values arising from different genome-wide distributions are not directly comparable across data sets, a Fisher transform procedure[22] is applied to convert the distribution of correlations to a normal-like distribution:

$$f_d(x,y) = \frac{1}{2}\ln\frac{1+r_d(x,y)}{1-r_d(x,y)}$$

where $f_d(x, y)$ is the Fisher-transformed score. Then the data are translated to z scores for standardization:

$$z_d(x,y) = \frac{1}{\mathrm{std}(f_d)}[f_d(x,y) - \mathrm{avg}(f_d)]$$

where $\mathrm{avg}(f_d)$ is the average of f_d for all (x, y) pairs, and $\mathrm{std}(f_d)$ is the s.d. of f_d.

The normalization procedure has been used in previous studies[5,23] and has been found successful in transforming most correlation distributions that are generated from different platforms and technologies into a comparable normal distribution with mean 0 and variance 1 (**Supplementary Fig. 8**). Note that SEEK also works well with other correlation measures, such as Spearman and bicor[24] (**Supplementary Fig. 9**). We found that the normalized Pearson correlation provides performance better or comparable to that of Spearman and bicor in the search setting, likely because the normalization procedure and the SEEK algorithm itself reduce the effects of outliers in search performance (**Supplementary Fig. 9**).

Search algorithm. The search algorithm takes two inputs: (i) a set of query genes $Q = \{q_1, \ldots, q_x\}$ and (ii) the set of correlation z scores containing the query $z_d(g, q)$ for each data set d in the data compendium D, for all genes q in Q and for all genes g in the genome G. The outputs of the algorithm are a prioritized list of data sets and coexpressed genes relevant to Q.

The search algorithm consists of four steps. The first step is to load precomputed z scores of Pearson correlations (in the normalization step above) containing the query across D.

The second step is to perform hubbiness correction on each data set d. The correction procedure is motivated by the observation that 'hubby'[9,10] or well-connected genes in the coexpression network represent global, well-coexpressed processes[25] and can contaminate the search results regardless of query composition owing to the effect of unbalanced gene connectivity in a scale-free coexpression network[9,10,26–28], which can lead to nonspecific results in search or clustering approaches. To avoid the bias created by hubby genes that are not related to the user's query or pathway of interest, our method corrects each gene g's correlation to q in each data set d

$$\tilde{z}_d(g,q) = z_d(g,q) - \frac{1}{|G|}\sum_{x\in G} z_d(g,x) \qquad (1)$$

where $\tilde{z}$ is the hubbiness-corrected z score. By subtracting g's average correlation from the correlation of (g, q), we expect the resulting score to emphasize g's coexpression specifically with the query rather than its general connectivity. The control of coexpression hubbiness enables the detection of specific biological signals in the data that would otherwise be swamped by broad coexpression patterns of the most well-connected genes.

The third step performs cross-validation–based data set weighting. The goal is to rank data sets according to each data set's relevance to the query[5]. The result will be the first output of the search system and will also be used to compute the final gene-score vector for the last step. The main idea is to upweight data sets where a subset of the query genes can retrieve the remaining query genes well on the basis of normalized, hubbiness-corrected coexpression in that data set. Thus, it is analogous in spirit to the cross-validation procedures commonly used in machine learning, where a subset of the standard (in this case, query) 'hides' from the system to assess how well the method can predict these hidden genes.

To describe the weighting method, we first introduce some notations. The data set d is implicit in each formula below and omitted for brevity; thus $\tilde{z}(g, q)$ is the corrected z score for g to a query gene q in Q in data set d. Let $R_q = (g^{(1)}, g^{(2)}, g^{(3)}, \ldots, g^{(r)})$ be the sequence of genes at rank 1, 2, 3, …, r obtained from ordering genes by decreasing $\tilde{z}(g, q)$. That is, R_q satisfies: $\tilde{z}(g^{(1)}, q) \geq \tilde{z}(g^{(2)}, q) \geq \tilde{z}(g^{(3)}, q)\ldots$. Let $r(t, R_q)$ be the rank of gene t in the ranking R_q minus 1 (for example, $r(g^{(1)}, R_q) = 0$), and let $p < 1$ be a rate parameter, which we set at 0.99 based on empirical analysis (**Supplementary Fig. 10**). Then the weight w of the data set is

$$w = \frac{1}{|Q|}\sum_{q\in Q}\left[(1-p)\sum_{t\in Q-q} p^{r(t,R_q)}\right] \qquad (2)$$

The weighting formula performs cross-validations on q in the set Q. The goal is to detect which query genes q can best retrieve the remainder query $Q - q$; such instances of q have a high contribution to w. We shorten $r(t, R_q)$ in equation (2) to $r(t)$. The exact form of this expression for weight (i.e., sum of $p^{r(t)}$) is inspired by rank-biased precision[29] and is adapted to our setting to robustly measure the effectiveness of the data set in retrieving $Q - q$. Here, $p < 1$ is the rate parameter in rank-biased precision and is the

parameter of geometric distribution, as $r(t)$ assumes discrete values. When it is employed, $p^{r(t)}$ upweights ranks for genes t in the set $Q - q$ that are high in the rank list (i.e., $r(t)$ is small), which intuitively emphasizes those genes in the query that are highly coexpressed with each other. The measure has the desired property of upweighting pairs of query genes that are well correlated while not allowing the correlations between the uninformative, noncoherent part of the query to affect the weight of the data set because the query genes may only be partially coexpressed in a given data set. Compared to previous methods[5], our method gains robustness to heterogeneous query signals because the reward on the highly coherent query genes far outweighs the damaging effect of a few noncoherent query genes, which are poorly ranked relative to other query genes, have high $r(t)$ and have scores $p^{r(t)}$ tending to 0.

The last step of the algorithm calculates the final integrated gene scores to generate a master ranking of coexpressed genes that is the second output of the system (in addition to data set relevance weighting). We obtain the gene-to-query score matrix $\mathbf{M}_{G,D}$, where the entry $M_{g,d}$ is the average corrected z score of gene g to the query in data set d

$$M_{g,d} = \frac{1}{|Q|}\sum_{q \in Q} \tilde{z}_d(g,q)$$

With the data set weight vector from the previous step $\mathbf{w} = [w_1, w_2, \ldots]$, a simple formulation of the final gene-score vector $\mathbf{F}$ is given by

$$\mathbf{F} = \mathbf{M}_{G,D} \times \alpha \mathbf{w}^{\mathrm{T}}, \quad \alpha = 1/\sum_{d \in D} w_d$$

Although previous research had some success with this formulation[5], our findings show that it works well only in the presence of complete gene information with no missing genes in $\mathbf{M}_{G,D}$. When there are heterogeneous sources of data in the compendium (for example, different microarray and RNA-seq platforms), the confounding factor of missing genes and partial gene rankings must be accounted for. Our approach is to modify the procedure above by employing threshold parameters to exclude a data set from weighting if it does not contain enough query genes and to exclude a gene from the final ranking if it is not assayed by a sufficient number of data sets in the compendium (**Supplementary Note 6**).

The pseudocode for the entire SEEK search algorithm can be found in **Supplementary Note 6**. The algorithm is robust to query composition (**Supplementary Figs. 1** and **2**) and data set quality, including automatically downweighting data sets with substantial batch effects (**Supplementary Note 1** and **Supplementary Fig. 7**). Computer source codes are deposited at https://bitbucket.org/libsleipnir/sleipnir.

For single-gene queries, the search algorithm performs the same steps above except that in the data set weighting step, the algorithm assigns equal weight to all data sets. Thus, for single-gene queries, the search system will treat each data set equally and retrieve genes that are generally correlated with the query in the hubbiness-corrected space. If users wish to perform their single-gene searches in a tissue-specific or disease-specific manner, they can manually define a category of data sets using the extensive "Refine Search" interface on the SEEK website, which will restrict D in the search system input.

Estimating the significance of gene scores. We estimate a P value for each retrieved gene by comparing the integrated score of each gene with scores from a pool of 10,000 randomly generated queries with diverse query sizes varying from 1 to 100 genes. The random pool allows SEEK to estimate the significance of gene score as well as evaluate the specificity of that gene to the query genes (as opposed to random queries). For a given gene g and its final coexpression score $S_Q(g)$ generated from the user's query Q, the P value of g is estimated as the number of random queries R in which $S_R(g) > S_Q(g)$ divided by the random pool size.

Algorithm and interface implementations. The SEEK algorithm is implemented in C++ and has been integrated into the open-source C++ Sleipnir library, enabling other computational users to use and expand SEEK without website tie-in[30]. The back end employs the efficient data structures from the Sleipnir library to facilitate the process of handling large query sets of over 100 genes without memory overflow. SEEK's jobs are parallelized to make full use of the multiprocessor resources and their processing power. The SEEK web server is constructed with some of the latest web technologies including JQuery and Qtip2 libraries. Dynamic pages are generated with Java servlets running behind the Apache Tomcat server on a Red Hat CentOS Linux operating system. In addition, Ajax technology is deployed to send and retrieve data from the server asynchronously such that users can receive instant feedback on their gene enrichment analysis, expression zoom-in function and data set selection module without having to leave or refresh the page.

Metadata processing. SEEK categorizes data sets into tissue and disease groups by mining the description, title and sample-level characteristic fields in data sets' metadata. The text-mining procedure utilizes the UMLS MetaThesaurus[31] and BRENDA[32] controlled vocabularies to extract predefined concept names that are present in the individual fields. To ensure that tissue groups are accurate, we manually reviewed annotations to the frequently appearing terms generated by text mining. Similarly, we formed additional 'meta' data set groups, such as cancer and noncancer groups and the multitissue profiling group (**Supplementary Data 4**), to provide users with the ability to limit their search to such groups under the "Refine Search" feature of the website.

Large-scale functional evaluation setup. We conducted a comprehensive evaluation of SEEK in comparison with existing algorithms Gene Recommender, MEM (multi-experiment matrix) and combined data set correlation search (**Supplementary Note 2**). We tested each system's ability to retrieve genes from the same biological process given some chosen genes from the process as queries. For the evaluation, we partitioned the genes in each of the 995 GO biological process terms (**Supplementary Data 2**) into a query building set and a testing set. The query building set consists of a random sample of 25 genes from each term if the term has more than 40 genes, or else it is made of half of the number of genes in the term. Queries were formed by repeatedly sampling genes from the set, so that each query size has

doi:10.1038/nmeth.3249

ten different queries of that size represented, and we iteratively generated queries for sizes 2, 3, 4, … up to Q genes, where $Q = 0.8|\text{query building set}|$. The testing set consists of the remaining genes in the term (after subtracting the query building set) and is used for evaluating the queries' retrieval results. A precision-recall (PR) curve is computed on a per-query basis, averaged over all queries of a term and finally averaged over all evaluated terms to derive an overall system performance plot for each method. Fold improvement of precision over random is calculated at 10% recall (FIOR@10%) and uses a random ranking of genes where genes' rank positions are shuffled. By selecting genes randomly from each process in building the queries, we mimic the situation in which the query genes are functionally related but not well coexpressed. By keeping the two sets (query building and testing) separate in the evaluation, we can reduce the performance variation between the queries of the same size within a process.

For building gold-standard GO gene sets used in evaluation, we used gene annotations with experimental evidence codes (IMP, IGI, IPI, IDA, IEP, EXP) as well as TAS (traceable author statement) and NAS (nontraceable author statement). To select the GO slim set (**Supplementary Data 3**) used for studying the effect of compendium size, we carefully examined the title and description of the GO terms in the context of the GO hierarchy and arrived at a nonredundant subset of GO terms that are both specific enough to be informative and diverse enough to represent the hierarchy; this is similar to the approach in ref. 33.

To evaluate SEEK's performance as a function of the query size, we pooled together previously built biological process queries from 995 processes and then binned them by query size (2–20 genes). We examined three categories of biological processes based on the number of annotated genes in each process: 20–40 genes, 40–100 genes and 100–300 genes. Performance refers to the fold improvement of precision over random at 10% recall in using each query to retrieve remaining genes from its corresponding process.

To evaluate the search system's robustness to noisy query genes, we selected over 1,800 five-gene and ten-gene queries from 90 KEGG pathways with 50–100 genes per pathway. Each pathway had ten queries selected of each query size. We established a 'no-noise' case, where each query was purely made of genes belonging to the same KEGG pathway, and a noisy case, where one, two and four random genes were respectively added to each query. The fraction (FIOR@10% of each noisy query)/(FIOR@10% of the corresponding no-noise query) was calculated, where FIOR@10% refers to the performance of retrieving KEGG pathway genes using the queries.

21. Ramasamy, A., Mondry, A., Holmes, C.C. & Altman, D.G. *PLoS Med.* **5**, e184 (2008).
22. Fisher, R.A. *Biometrika* **10**, 507–521 (1915).
23. Huttenhower, C. *et al. Genome Res.* **19**, 1093–1106 (2009).
24. Song, L., Langfelder, P. & Horvath, S. *BMC Bioinformatics* **13**, 328 (2012).
25. Stuart, J.M., Segal, E., Koller, D. & Kim, S.K. *Science* **302**, 249–255 (2003).
26. Horvath, S. & Dong, J. *PLoS Comput. Biol.* **4**, e1000117 (2008).
27. Ruan, J., Dean, A.K. & Zhang, W. *BMC Syst. Biol.* **4**, 8 (2010).
28. Xulvi-Brunet, R. & Li, H. *Bioinformatics* **26**, 205–214 (2010).
29. Moffat, A. & Zobel, J. *ACM Trans. Inf. Syst.* **27**, 2 (2008).
30. Huttenhower, C., Schroeder, M., Chikina, M.D. & Troyanskaya, O.G. *Bioinformatics* **24**, 1559–1561 (2008).
31. Bodenreider, O. *Nucleic Acids Res.* **32**, D267–D270 (2004).
32. Gremse, M. *et al. Nucleic Acids Res.* **39**, D507–D513 (2011).
33. Myers, C.L., Barrett, D.R., Hibbs, M.A., Huttenhower, C. & Troyanskaya, O.G. *BMC Genomics* **7**, 187 (2006).