DETECTING GENE SIMILARITIES USING LARGE-SCALE CONTENT-BASED SEARCH SYSTEMS

QIAN ZHU

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE

BY THE DEPARTMENT OF

COMPUTER SCIENCE

ADVISOR: OLGA TROYANSKAYA

SEPTEMBER 2016

ABSTRACT

The accumulation of public gene expression datasets offers numerous opportunities for researchers to utilize these data to characterize gene functions, understand pathway actions, and formulate hypotheses about the molecular basis of human diseases. Yet, exploring this extremely large gene expression data collection has been challenging, due to a lack of effective tools in reusing existing datasets and exploring these datasets for targeted analyses. An important challenge is discovering robust gene signatures of biological processes and diseases, where this depends on the ability to detect similar genes that share gene expression patterns across a large set of conditions. This thesis discusses query-based systems that are intended for large-scale integration and exploration of gene similarities. It also discusses their key biological applications.

In the first part, I present SEEK, a search system and a novel algorithm for searching similar (or coexpressed) genes around a multigene query of interest. The search algorithm combines coexpressed genes using a sensitive dataset weighting algorithm for effective weighting of coexpression results. Notably, through the robust search of thousands of human datasets, the retrieval of functionally co-annotated genes always improves with the inclusion of more datasets, showing the promise of the large compendia. In the second part, I extend the work of SEEK to the expression compendia of 5 commonly studied model organisms. The new system ModSEEK enables accurate searches in a wider experimental variety, and has been extensively evaluated. In the third

part, I propose a novel framework for integrating and comparing coexpression context across a pair of organisms. I leverage both comparative genomics orthology data and functional genomics coexpression data, in an unsupervised framework to identify pairs of genes in an orthologous group that are similarly highly coexpressed to an orthologous query in two organisms. I show that such functionally similar pairs of genes can be used to improve the performance of single-organism gene retrieval searches. In the final part, I demonstrate how coexpressed genes can be used to identify important transcription factors and dysregulated processes underlying breast cancer subtypes. This part highlights the promise of coexpressed genes in providing an understanding of cancer dysregulations.

ACKNOWLEDGMENTS

First, I would like to thank my advisor Olga Troyanskaya for her continued support in this project and for teaching me just so many things, from writing scientific papers to designing great slides for presentation, offering advice on my career and pointing out areas of improvement. Without her guidance, none of this thesis would be possible. I want to thank her for her keen insights, and for introducing me to several wonderful collaborators that I have had the pleasure working during my graduate study.

I want to thank my thesis committee members Mona Singh, and Andrea LePaugh for taking the time to serve on my committee. I want to thank Kai Li, my coadvisor, and Moses Charikar, who is involved in the search project, for providing valuable feedback in the early stage of the search project. I want to thank Vessela Kristensen for her continued guidance and encouragement, and contagious enthusiasm, which makes collaboration with her a great experience.

I want to thank the current and past members of Olga Troyanskaya's lab, including: Aaron Wong, Christopher Park, Young-suk Lee, Alicja Tadych, Arjun Krishnan, Casey Greene, Jian Zhou, Ana Bell, Chandra Theesfeld, Ruth Dannenfelser, Vicky Yao, Ran Zhang, Max Homilius, Dima Gorenshteyn, and Yuanfang Guan. These are the people with whom I had the pleasure collaborating for the work covered in this thesis, or with whom I shared many great moments in the lab. Thanks to all of you for making my time in the lab always memorable and enjoyable.

Last but not least, I want to thank my wife Xi Wang. Thanks to her constant encouragement and support, I am able to finish writing this thesis. My parents have always been role models for me as I grew up. Their unconditional love and support has been absolutely crucial for the completion of my PhD study.

Finally, I acknowledge the various funding sources from Princeton University and grants from US National Institutes of Health (NIH) and US National Science Foundation (NSF): NIH R01 HG005998, NIH R01 GM071966, and NSF DBI-0546275.

Contents

LIST OF FIGURES

1 INTRODUCTION

1.1 Background

1.1.1 Rapid accumulation of gene expression data

The beginning of the 21st century is marked by rapid developments of high-throughput genomics technologies. These technologies have generated massive amounts of data which hold great potentials for exploration and discoveries by biological researchers. At present, Gene Expression Omnibus [1] and ArrayExpress [2] are two of the most well-known web-based repositories for experimental data, and have allowed researchers from all over the world to submit diverse high-throughput experiments in standardized formats [3].

These noteworthy efforts in collecting and organizing data, along with the reduction of cost of technologies and increase of computational power, have allowed data to accumulate at an extremely rapid pace. Datasets are no longer found in small numbers, but are now registered in huge numbers containing together billions of data points. Gene expression datasets assayed by microarray [4] or by high-throughput RNA-sequencing [5] technologies represent the most abundant category of data. Primarily, gene expression datasets have been targeted for patient diagnostic purpose, which stratifies patients to different disease risk groups [6], or for identifying genes that might be disease-markers [7], or for understanding gene functions in a perturbed cellular system of model organisms [8,9]. Each expression experiment typically examines the expression (or mRNA abundance) of genes for tens of thousands of genes in an organism's genome, and is specific to a condition which can be a disease-state, cell line perturbation, or natural variation among

individuals. Needless to say, expression studies have been an extraordinarily powerful tool both in research and in the clinic[6,10].

1.1.2 Challenges in interpreting gene expression data

Interpreting microarray datasets can be challenging due to the inherently noisy nature of microarray datasets. There are uncertainties in the gene expression measurements, and technical variables such as the experimental design, number of replicates can influence the interpretability and success of a study. Therefore, typically results from a single dataset are combined with or compared to an independent study to gain credibility. In the past decade, considerable efforts have been devoted to increasing robustness and eliminating technical biases in the data. These include 1) the correction of batch effects [11], which are experimental factors such as time of day, who perform the experiments, sample batches that confound the main variable of interest, 2) increasing sample sizes for better statistical power, such as large-scale studies consisting of thousands of individuals, 3) publication of guidelines on proper processing and handling of data [12], and 4) combining the results of multiple related expression studies (also known as meta-analysis) [13].

1.1.3 Meta-analysis, integrating diverse gene expression data

Particularly, this 4th approach meta-analysis has been popularized, demonstrated to be quite successful in the last decade in increasing robustness and generalizability of expression studies and can be applied effectively in many situations. Retrospectively, early examples of meta-analysis can be traced back to 2004 with the integration of differentially expressed genes for about 100 human expression datasets (Oncomine) [14,15]. Here, differentially expressed genes are defined as genes found statistically different in terms of expression between two groups of conditions, such as case vs. control, normal vs.

disease, or treatment vs. no treatment. Oncomine seeks to integrate cancer vs. normal gene-lists across ~100 cancer microarray studies to identify common neoplastic progression gene signatures [14]. Over the years, different meta-analysis approaches have been developed[13], including methods that combine the gene *P*-values among studies[14], combine ranks[16,17], combine effect sizes[18], or directly merge raw datasets [19]. The key message to be learnt from these meta-analysis approaches is that genes derived from meta-analysis are more robust than any single expression dataset can derive.

Though these approaches have been extraordinarily useful, a key challenge has been scaling up the meta-analysis to larger data collections. For example, a caveat of Oncomine has been that experimenters need to manually identify dichotomous groups of conditions prior to differential expression analysis. As the number of expression datasets has exploded in the modern day, this would be difficult to do. Namely, manually curating datasets cannot be expected to keep pace with the growth of data, so alternative computational strategies must be developed for analyzing, integrating datasets in large quantities to maximize utilization of the existing data compendia.

1.1.4 Coexpression and its biological implications

The idea of finding groups of genes that are *coexpressed* (or exhibit coordinated expression) has become more and more widespread and routine. Coexpression as a *content-based* measure of gene-similarity is normally characterized by the use of expression data to define similar genes. It is alternative to other gene similarity measures such as semantic similarity[20] or sequence/phylogenetic similarity [21]. Coexpression carries very valuable information about genes. Early microarray studies on the yeast organism have revealed that when genome-wide expression profiles for biological specimens taken

from different time points are compared, the expression of genes in cell cycle follow an interesting cyclical pattern according to time [22]. Genes belonging to different clusters of expression patterns are heavily enriched in different stages of the cell cycle. For example, clusters of genes have been found to be up-regulated (or have the highest expression) during the G-phase of the cell cycle, while a distinct second cluster of genes are up-regulated specifically in S-phase. These observations indicate that gene expression programs align closely with the biological activity within the cell. The expressions of genes are changed in a coordinated manner (i.e. together in a group) rather than individually and stochastically.

The task of finding coexpressed genes has important biological implications. First, it was observed that not only are genes in cell cycle coexpressed but also genes participating in other processes are well-coexpressed such as cellular respiration, ribosome biogenesis, protein synthesis, in human cancers and other organisms [23,24]. Thus, coexpressions hold the key to solving several difficult tasks, including unraveling gene functions for uncharacterized genes in the genome, identifying novel gene members of existing pathways, and characterizing multifunctional roles of existing genes. To characterize gene function using coexpression, one relies on the principle of guilt-by-association [25], whereby the unknown function of a gene can be inferred from the known gene functions of the neighboring genes (or coexpressed genes). Using this principle, coexpression analysis has been extended to multiple organisms to find conserved gene functions. This makes use of the fact that homologous genes (or similar genes) across organisms can have similar coexpressed gene neighbors [26]. As it is rather rare that compatible experiments can be located for comparing diverse organisms, coexpression

can become a useful measure for comparing organisms, as has been illustrated in prior works [27,28]. Useful measures such as the expression context conservation (ECC)[27,29], coexpression network similarity, have been developed to compare *A. thaliana* and *Oryza sativa*, based on the similarity between coexpression network neighborhoods for orthologous gene pairs. The implications of coexpression have included many other applications, including identifying genes that are in the same cellular compartment, or being part of a protein complex[30].

1.1.5 Computational approaches for searching coexpressed genes

Methods for finding coexpressed genes have been developed and evolved over time. Earlier works include clustering, such as K-means, hierarchical clustering, biclustering-based approaches (as reviewed in this paper [31]), singular value decomposition and principle component analysis [32] approaches. Whereas clustering identifies a set of genes coexpressed across all conditions of a dataset, biclustering on the other hand seeks to identify genes whose expression is coexpressed across only a subset of conditions. While these methods have been successful in many ways, they suffer a major weakness: the fact that biclustering is a computationally intractable problem imposes a severe limitation on the size of dataset that biclustering generally can tackle. As a result, biclustering approaches have been typically applied to datasets involving a few hundred conditions. It becomes apparent that coexpression methods that analyze the entirety of the gene expression collection (involving 100,000 conditions from thousands of datasets) were urgently needed to properly explore the diverse genomic landscape. Query-, context-sensitive search approaches were designed to overcome this challenge of large-scale search and analysis. The idea of query-sensitive search was pioneered by Hibbs &

Troyanskaya who first developed a system integrating over 100 datasets covering 2400 experimental conditions (SPELL [33]). The distinction with earlier approaches is the introduction of a *query*, which is made up of a gene or a gene-set defining users' context. This search system is closely analogous to what Google has done for text-based document search and image search – the output of an expression search is composed of 1) expression datasets found to be relevant to the query and 2) genes that are found coexpressed with the query based on the measure of similarity of expression in the datasets, much like searching for similar images given a query image by utilizing image features. The advantages of employing a query in expression search include 1) significant savings of search space and time, 2) targeting of the search results to a context of interest, and 3) weighting of specific datasets, which is particularly useful for computing context-specific query similarity scores. Some key concepts of a query-based expression search system are next introduced.

1.1.6 Definitions

A *dataset* is an expression matrix composed of genes (rows) and conditions (columns), and is a unit of submission in the repository Gene Expression Omnibus. It is linked to a biological question or study. Each dataset contains experimental conditions that are 1) done at a certain time by a certain laboratory, and 2) intended to investigate a biological question, therefore a dataset is typically linked to a publication. A collection or a *compendium* contains many datasets. Instruments generating datasets are called *platforms*, or technologies, such as Affymetrix Human Genome U133 Plus 2.0 Arrays, Illumina HiSeq 2000, etc. *Gene-gene correlations*, interchangeably referred to as *gene similarities*, *coexpression*, or simply *correlations*, are measures of similarities of two genes in terms

of their expression pattern across all conditions in a given dataset. The simplest ways of defining correlations are Pearson and Spearman correlation coefficients. In a multi-dataset scenario, gene-gene correlations may be weighted and aggregated, for example: $r_{x,y} = \frac{1}{\sum_{d_i \in D} w_{d_i}} \left(w_{d_1} r_{d_1,x,y} + w_{d_2} r_{d_2,x,y} + \cdots \right)$ where $r_{x,y}$ is the aggregated correlation, x, y are two genes and w_d is the weight of a dataset in the compendium. This weight can be a reflection of dataset reliability or relevance to the query of interest. Given a query $Q = \{q_1, q_2, \ldots\}$, the *correlation search* (or *gene-similarity search*) for whole-compendium is described as follows: find a rank-list of genes sorted by decreasing correlation score between each gene and the query, by some measure of similarity and by some way of aggregation of compendium datasets to arrive at the final rank-list of genes. This type of search system is *data-driven*, since every step of the search algorithm relies only on the expression matrices and the query genes with no human intervention involved. It is *content-based* since the search of similar genes is based on expression data.

1.2 Contributions of my thesis

Though several expression-based gene-similarity search systems have been developed in the past, none of them have achieved the following objectives simultaneously: 1) Support for multi-gene queries. It was assumed that the query must be single-gene in most of previous systems. While it is simplistic, single-gene query may cause ambiguous context due to genes' multifunctional roles. 2) Support for cross-platform dataset integration. In prior systems, a platform of interest must be selected in order to search the query. Usually only Affymetrix array platforms are supported. 3) Explore the entirety of the expression compendium. Data integration has been limited to a few hundred datasets, which represent a small portion of the available public data holdings. These are the main

reasons that motivated me to formulate this thesis, with the goal of developing a large-scale gene-expression similarity search system for human and later extending it to 5 other commonly used model organisms. The developed systems not only solve all of the above three objectives, they also feature distinct advantages not previously present, including 1) prioritization of all of the thousands of expression datasets based on a multi-gene query, 2) visualization of the search results (i.e. display of coexpression patterns between query and query-similar genes), 3) refinement of search results by allowing users restrict datasets at will, and 4) extensive evaluation of the system using ground-truth Gene Ontology Biological Process gene annotations. The systems already have enabled thousands of biomedical researchers with or without a computational background to easily search and navigate all of the available expression data, which means that these systems are practically putting the expression data in the hands of researchers for exploration-based discoveries.

This thesis is divided into the following chapters with contributions listed below:

- In Chapter 2, I develop a gene-similarity search system called SEEK (Search-based exploration of expression compendium) which enables coexpression mining from over 5,000 human datasets containing 100,000 conditions. This system importantly solves the challenge of weighting datasets. Manually locating relevant datasets to a query of interest is infeasible in thousands of datasets. Yet this task of identifying important datasets is critical to the effective exploration of data and for accurate coexpression mining. I develop a computational method for weighting datasets from a query of interest and utilize it in coexpressed gene discovery. This weighting method is based on cross-validation of query genes by discovering datasets which exhibit

coexpression of query genes themselves. It is robust to detect even datasets where query gene-set may not be fully coexpressed with each other but are significantly more coexpressed than random gene-sets. To evaluate this method, I demonstrate that the derived dataset weights, when being used to integrate coexpressed genes across datasets, enable accurate retrieval of genes sharing important biological contexts. The coexpressed genes were judged as accurate or not according to the ground-truth GO Biological Process gene annotations. The search performance has been evaluated to be robust across multiple platforms including microarrays and next generation sequencing technologies, and the method can handle thousands of datasets. Also, I demonstrate that the accuracy of retrieval is increased when thousands of datasets are utilized together. The search system implementing this method is available in a web-friendly interface at http://seek.princeton.edu, which features on-the-fly computation of dataset weights, computation of coexpressed genes, intuitive visualizations, and further search refinements. This work is published in Nature Methods.

- Based on the encouraging results from SEEK, in Chapter 3 I extend the search functionalities of SEEK to 5 other commonly studied model organisms: *S. cerevisiae* (yeast), *D. melanogaster* (fly), *M. musculus* (mouse), *C. elegans* (worm), *D. rerio* (zebrafish). The search system, called ModSEEK (model-organism SEEK) and available at http://seek.princeton.edu/modSeek/, permits for the first time large-scale coexpression mining for model organism biologists. The search system holds in its database many more experimental varieties previously not seen in the human version. Thus, it is suitable for experiment planning and hypothesis generations. I extensively evaluated the system for robust gene retrieval. In addition, I develop a fast estimation

method for coexpression *P*-value given a gene-set of interest. This method is based on the extreme value theory, where I show that the distribution of coexpression scores in a dataset can be modeled with a generalized pareto distribution (GPD). Estimating the parameters of GPD allows fast computation of P-values across thousands of datasets.

- In Chapter 4, I develop an unsupervised coexpression-based cross-organism gene-retrieval framework and system (http://seek.princeton.edu/modSeek/viewer_index.jsp). The rise of model organism datasets in the repository calls for a need to integrate data across organisms for more meaningful comparisons. In this chapter, I combine both functional genomics and comparative genomics gene orthology data to infer pairs of genes across organisms that share similarly high degree of coexpression with respect to a query of interest. These functionally similar linkages are particularly robust as they are derived from large-compendia integration, and specific to the query context in question. I report that this combined strategy can significantly boost the gene retrieval performances of single organisms. A manuscript detailing Chapters 3 and 4 will be submitted soon for publication.
- Lastly in Chapter 5, I describe a key biological application of SEEK: transcription factor (TF) inference and TF regulatory network mapping. The coexpressed genes returned by SEEK can serve as input for further analysis of the transcriptional mechanisms regulating coexpressed genes. Analyzing cis-regulatory elements enriched by the enhancer and promoter regions of coexpressed genes allow one to glean insights on what transcription factors may possibly bind to these genes. The ENCODE consortium currently provides experimentally characterized, genome-wide maps of cis-regulatory binding elements for hundreds of diverse human cell types and

cell lines (via ChIP-seq TF-binding data). Here I leverage this useful resource and couple it with SEEK's robust coexpressed genes to deduce a list of candidate TF regulators of coexpression. Using the breast cancer as a disease system, I use SEEK to expand disease-gene signatures for individual subtypes of breast cancer. Then I next systematically determine which ENCODE ChIP'd TFs best resemble individual disease subtype based on coverage of cis-regulatory elements among coexpressed genes. The TF regulators inferred from coexpression are indeed verified to be specifically related to breast cancer misregulations (copy number aberrations, DNA methylations, SNP) in TCGA datasets, demonstrating success of the approach. This work is supervised by Vessela Kristensen and a manuscript will be soon submitted. The work was presented at the 2015 Recomb Systems Biology/Regulatory Genomics Conference.

2 TARGETED EXPLORATION AND ANALYSIS OF LARGE CROSS-PLATFORM HUMAN TRANSCRIPTOMIC COMPENDIA

2.1 Abstract

We present SEEK (search-based exploration of expression compendia; http://seek.princeton.edu/), an expression-based search engine with the capability to handle 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.

2.2 Introduction

The accumulation of human gene expression data in public repositories, such as The Cancer Genome Atlas[34] and Gene Expression Omnibus[1], offers unprecedented opportunities for data-driven characterization of biological pathways that underlie human diseases. Exploratory, unsupervised approaches have proven to be essential to expression data analysis because they can provide a relatively unbiased view of a biological system and the flexibility for biomedical researchers to focus their analyses on an area of interest. In the past, a number of tools have been developed to enable effective

exploration of individual expression datasets[35–37]. However, existing exploratory and unsupervised approaches, such as clustering and bi-clustering[35,38] tools, do not readily extend to compendia that contain thousands of datasets from different technologies and platform. This is mostly attributable to the lack of scalability and robustness to the biological noise of these large expression collections.

Unsupervised, exploratory approaches are particularly suitable for data-driven discovery and settings with insufficient training data, as is the case for many areas of human biology. While supervised approaches have demonstrated promise in combining expression data for gene function/relationship prediction[39–41], they can be prone to biases towards prior knowledge and require ample high-specificity labeled data for training. They are generally focused on specific tasks such as function prediction. There is a clear need to enable user-driven, flexible, interactive exploratory analysis of the large compendia of gene expression data. Towards addressing this need, recent efforts have been focused on query-based search systems[33,42,43]. In general, these systems allow a biological researcher to start with a query of interest in mind and find genes that are coregulated (or coexpressed) with researcher's query genes. Query-based search provides an effective, easy to use, and flexible approach for generating hypotheses about genes and datasets in an area of interest. Previous query-based systems have been hampered by the number or type of datasets which can be handled: they are focused on a small number of datasets[33,43], or datasets pre-selected to a particular disease[15,44], or the system can only integrate datasets generated from one platform[42]. Several other approaches, such as GeneChaser[45] and ExpressionBlast[46], have been proposed for identifying datasets with

similar differential expression signals as a query expression profile, but they do not provide a coexpressed gene retrieval capability, nor a targeted expression visualization.

Enabling accurate and robust expression search over the entire diverse compendium is powerful because more than one dataset is likely to contain the biological signals relevant to the user's area of interest, as each human dataset represents a mixture of signals from diverse pathways affected by disease states, environmental factors, and clinical/experimental treatments. Thus, while it's challenging to identify such signals, it is worthwhile to address this issue so that users can harvest all the existing expression data to identify signals and make hypotheses that may not be discoverable from a single dataset. Building a query-based, unsupervised search system for the full dimension of the human compendium is difficult. The system must first deal with ambiguous probe-to-gene mappings, changing gene annotations, varying gene coverage between platforms, and diversely pre-processed datasets[12,47], and other technical challenges[11,48]. Second, the system must be able to handle the extensive and heterogeneous expression patterns. The compendium includes many different experimental conditions, cell types, and diseases[49] with substantial expression diversities[49,50] – a query-sensitive search algorithm is required to detect specific biological signals of interest to the user.

We thus developed a large-scale search system that is capable of handling expression data across multiple platforms, including NGS and microarray technologies. The system can automatically prioritize datasets that are relevant to the user's single or multi-gene query, and return genes co-regulated with the query in these datasets.

2.3 Results

2.3.1 System description.

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 (**Figs. 2.1**–**2.2**). 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. 2.1a** and **Supplementary Data 2.1**). It has been implemented in a user-friendly interactive web interface (http://seek.princeton.edu/), which includes expression visualization and interpretation modules (**Fig. 2.1a**). 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.

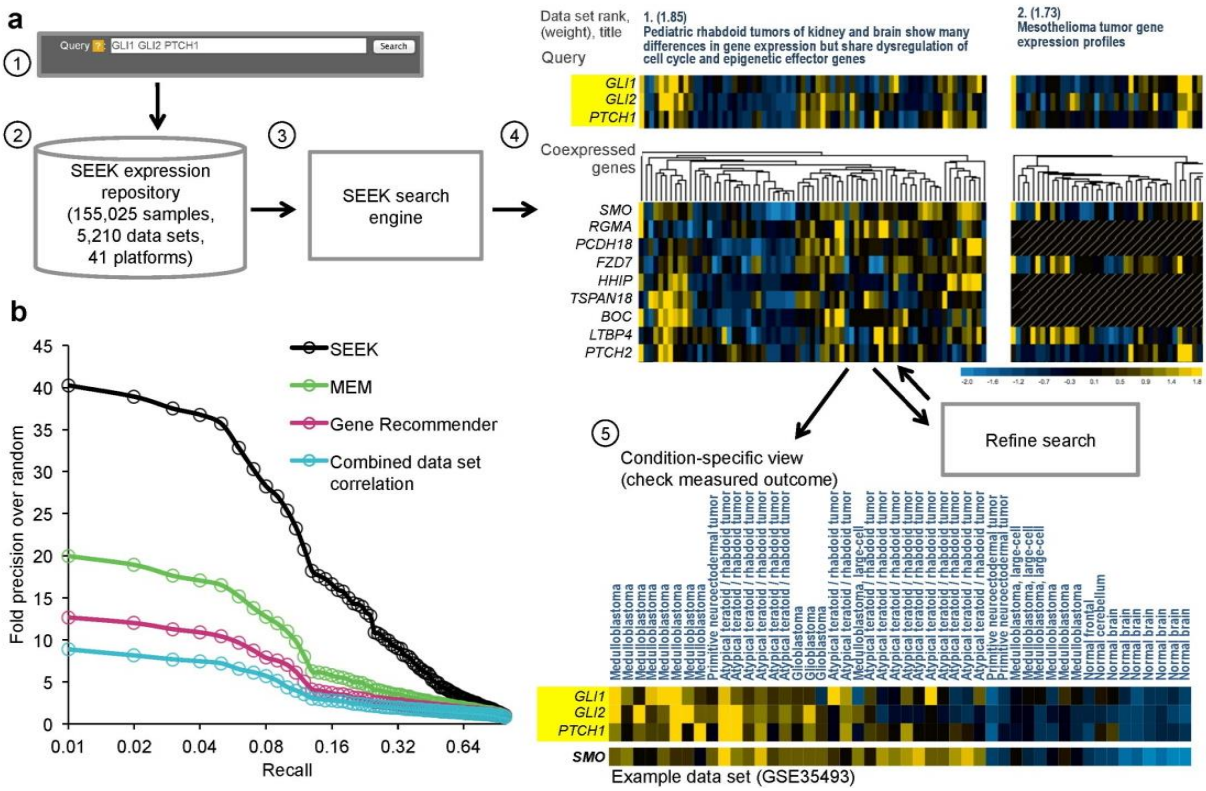


Figure 2.1 | SEEK system overview and systematic functional evaluation. (**a**) Users begin by providing a query gene set of interest to define a biological context of their search (step 1). SEEK searches the entire compendium and returns genes that are coexpressed with the query and the top relevant 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. 2.2**) and further analyze the results through a condition-specific view (step 5) (**Section A.2**). (**b**) Gene-retrieval evaluations across 995 diverse GO biological process terms for the SEEK, MEM, Gene Recommender and combined data set correlation algorithms (**Section 2.3.8**). 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 2.2**) and additional detailed comparative evaluations (**Figs. 2.2–2.3**, **Supplementary Figs. 2.1–2.2**) are provided.

The search algorithm (see Section 2.3 Methods) allows for multigene queries and includes a 'hub' gene[51,52] bias 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 hub gene correction procedure is applied to remove biases caused by generically well-coexpressed genes not specific to the user's query of interest. The data set–weighting algorithm then prioritizes relevant data sets according to the query of interest on the basis of normalized, hub-corrected coexpression in each data set. The idea is to upweight data sets from which a subset of the query genes can retrieve the remaining query genes well using coexpression (cross-validation–based weighting). This approach is effective even when not all query genes are coexpressed. Finally, the 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. 2.1b** and prior works[23,33,42,43]). 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 usage can be extended to make functional comparisons across organisms. A key challenge here is that the search results can be polluted by batch effects[11], 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 (**Fig. 2.4** and **Section 2.2.3**) and provides accurate retrieval of functionally related genes and data sets (**Fig. 2.1b** and **Supplementary Figs. 2.1**–**2.3**).

2.3.2 Gene retrieval evaluations.

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 (see Section 2.4 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 2.2**), and it consistently outperformed previous search approaches, including the only query-based human search system, MEM[42]; Gene Recommender[43] (not available for human as a resource); and the correlations on the combined data set (**Fig. 2.1b** and **Fig. 2.2**). 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 (**Fig. 2.2a**) and that the algorithm is robust with respect to query noise (**Supplementary Fig. 2.2**).

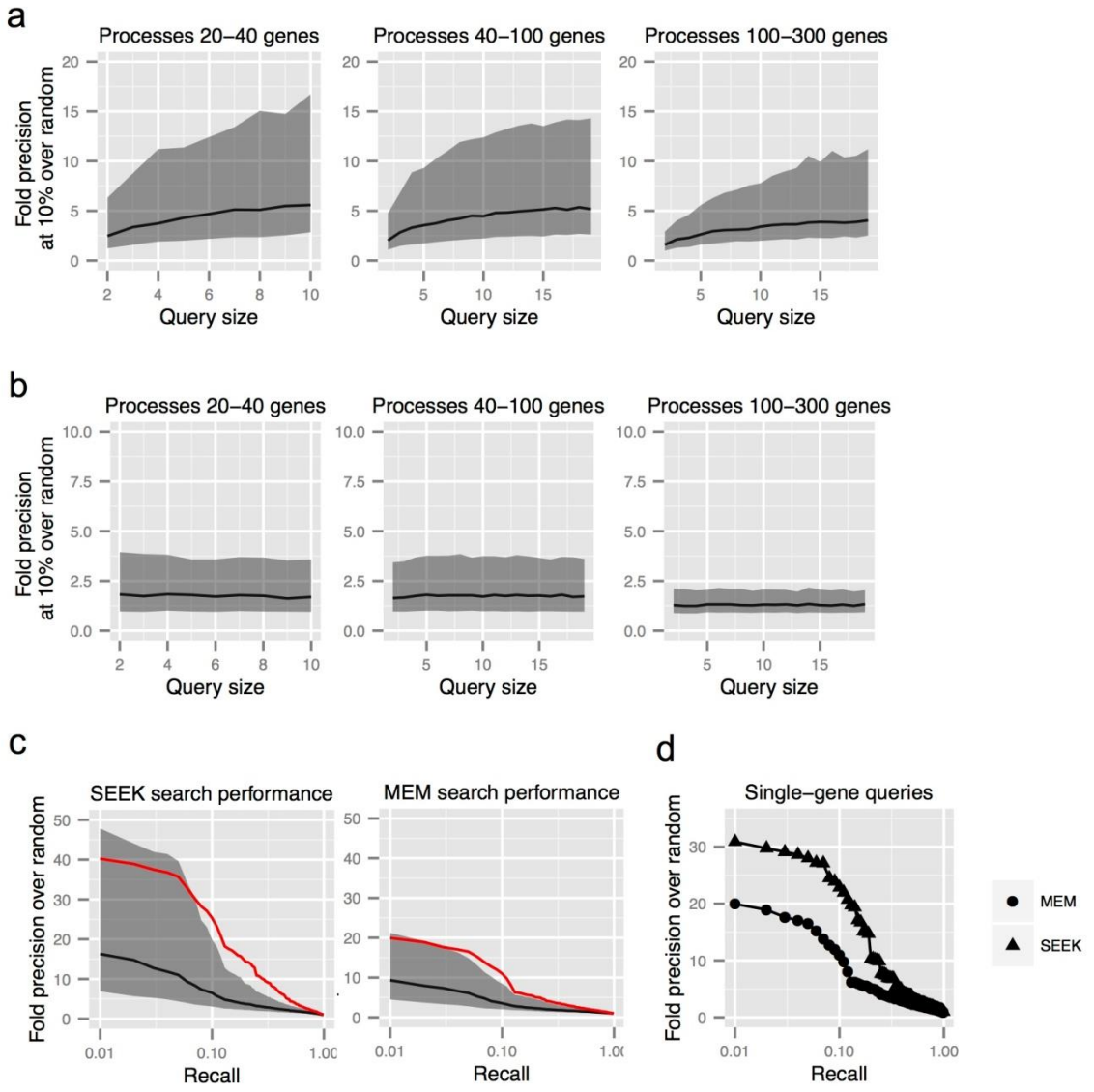


Figure 2.2 | Gene-retrieval performance vs. query size, and comparisons between SEEK and MEM in single- and multiple-gene queries. (**a**) SEEK's performance vs. query size (the number of query genes). The plot shows the median (black line) and the IQR (shaded area). The retrieval performance increases as a function of query size, showing that the improved query context, resulted from including more process-relevant genes in the query, can help boost gene retrieval. (**b**) Gene Recommender's performance vs. query size. (**c**) The performance of SEEK and MEM. The evaluation is the same as used in **Fig. 2** (main text). These plots additionally show the mean (red line), median (black line), and the IQR (shaded area) across 995 processes. (**d**) Single-gene query retrieval performance.

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 (**Fig. 2.3** and **Supplementary Data 2.3**). 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. 2.2**), 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. 2.2**). 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.

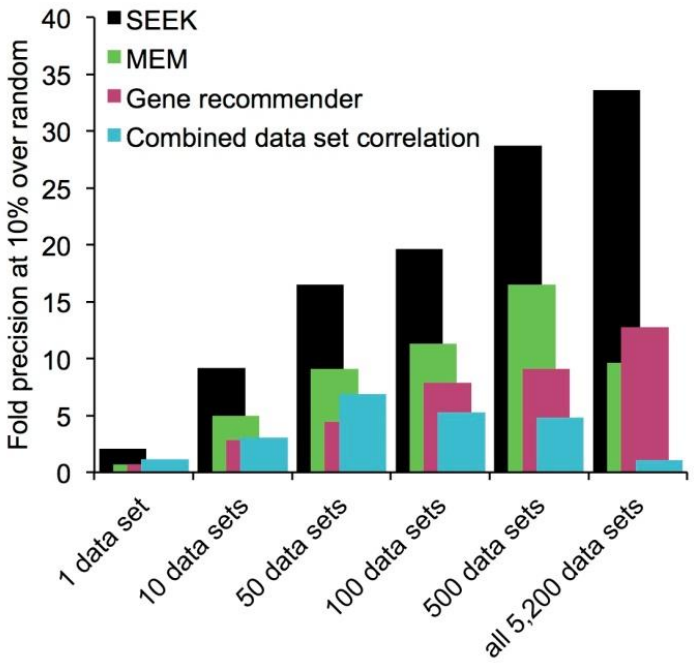


Figure 2.3 | Performance of SEEK and other search systems over increasing numbers of gene expression data sets. A sub set consisting of 121 GO-slim (**Supplementary Data 2.3**) terms were used to evaluate each system's gene retrieval performance on six compendium sizes each built from random subsets of the data sets in its full compendium. The FIOR@10% is measured. All algorithms are applied to the same data compendia, and MEM, Gene Recommender, and combined data set correlation algorithms do not scale to the large human compendium that SEEK is able to effectively utilize.

2.3.3 Batch effect evaluation.

SEEK uses the data set weighting algorithm to systematically address the challenge of the possible batch effects that exist in certain data sets in the compendium. To evaluate SEEK's effectiveness, we identified low quality data sets with severe batch effects in the compendium based on the variation in the samples' expression value distribution within each data set. Specifically, for each data set *d*, with *n* samples, we calculate the standard deviation σ_d

$$\sigma_d = std\left(IQR_{d_1}, IQR_{d_2}, \ldots, IQR_{d_n}\right) \quad \text{Eq2.1}$$

where d_1, …, d_n are the samples in data set *d*, and IQR is the interquartile range for the expression values in a sample d_x. A relatively high σ_d signifies a technical bias or a shift in the median and IQR of the gene expression values in that array, which is generally caused by batch effects. We then examined the 100 datasets with the highest σ_d (and thus most severe batch effects) in the compendium to see where they are ranked in full dataset prioritization (~4,500 data sets) for 121 diverse GO-slim queries (GO-slim[53] provides a curated set of diverse, high-level GO terms that are nonetheless specific enough for experimental evaluation and span the full set of GO biological processes. Each GO-slim query consists of all experimentally annotated genes in that GO-slim term.) There was

indeed a negative enrichment of the 100 data sets in full data prioritizations across 121 GO slim queries (**Fig. 2.4**), indicating that data sets with substantial batch effects are systematically ranked lower than randomly selected data sets, and thus effectively down-weighed in the SEEK search process. In fact, a high proportion (84%) of these 100 low quality data sets have a non-significant data set weight assigned by SEEK (at *P* more than the 0.001 significance cut off) (**Fig. 2.4**, source data), thus demonstrating the effectiveness of the SEEK data set weighting algorithm in automatically handling low-quality data sets.

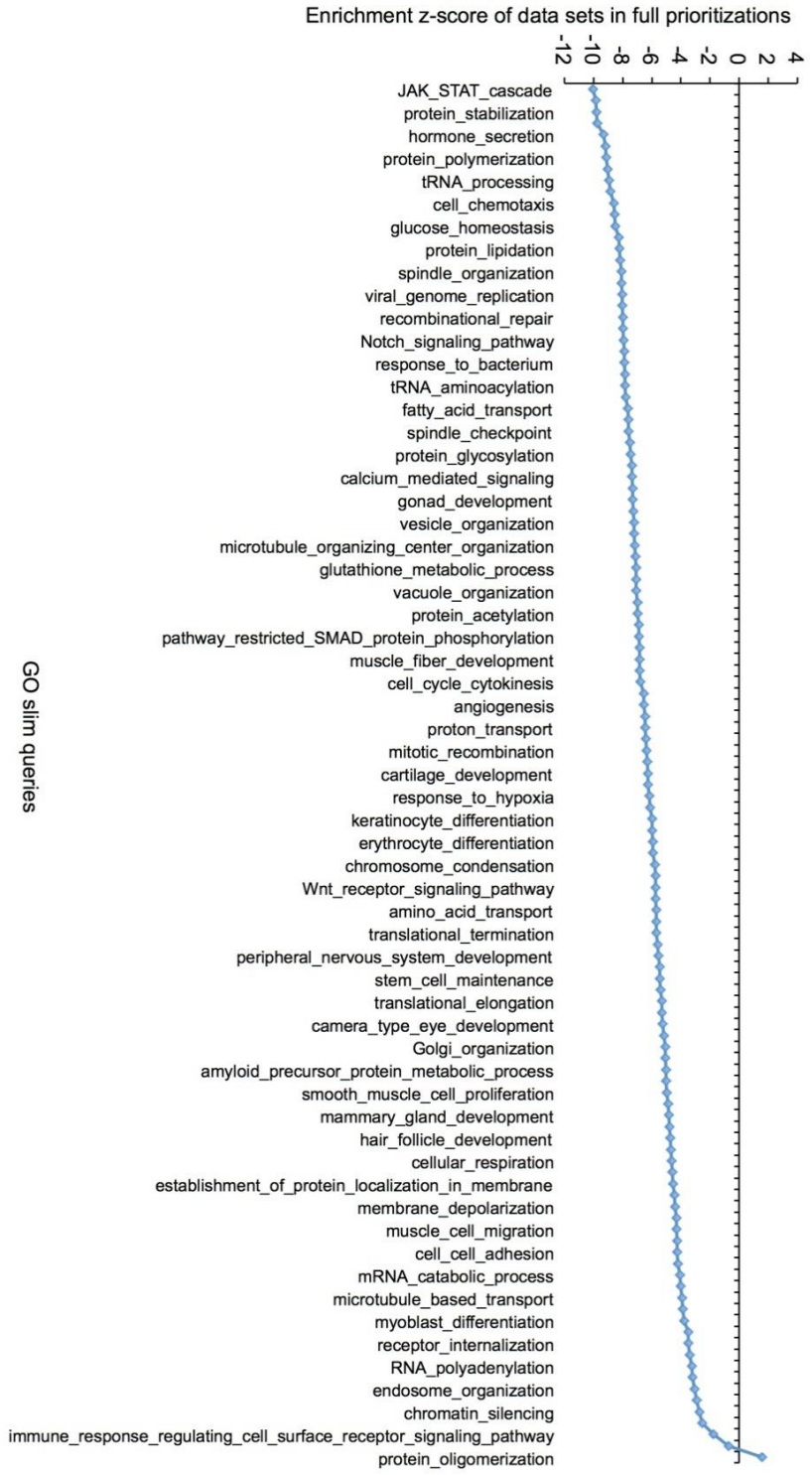


Figure 2.4 | Batch-effect analysis. Enrichment of 100 batch affected data sets (**Section A.1**) in full data prioritizations (4,500 data sets) across 121 GO slim queries. This test was done to check if the 100 data sets (with severe batch effects) have a lower score than randomly selected data sets in the ranking. Score-based PAGE enrichment was used. The

data sets consistently received a lower score than randomly selected data sets (avg. *z*-score = –6.3, $P < 1.39\times10^{-10}$), showing that low quality data sets have a relatively small impact on the prioritizations.

2.3.4 Case study.

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[54]. By examining this query in the context of a large compendium of expression data sets (**Fig. 2.5** and **Supplementary Fig. 2.3**), we observed a wide prevalence of aberrant Hh signaling across many diseased tissues (**Fig. 2.5**). The top-ranked data sets had substantially higher weights, indicating the presence of a strong query-related signal in these data (**Fig. 2.5**), and appeared to be more specific to the Hh query than to random queries (**Supplementary Fig. 2.3**). 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[55,56], (ii) human germ cell tumors, in which Hh pathway mutations have been linked to aberrant Hh activation in human germ cells[57], and (iii) malignant rhabdoid tumors[58,59], in which mutations have been found to lead to Hh signaling activation[59]. 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*[60], the *Cos2* homolog *KIF7* (ref. [61]) (**Fig. 2.6a**) as well as additional genes associated with Hh dysregulation in cancer (**Section A.1**).

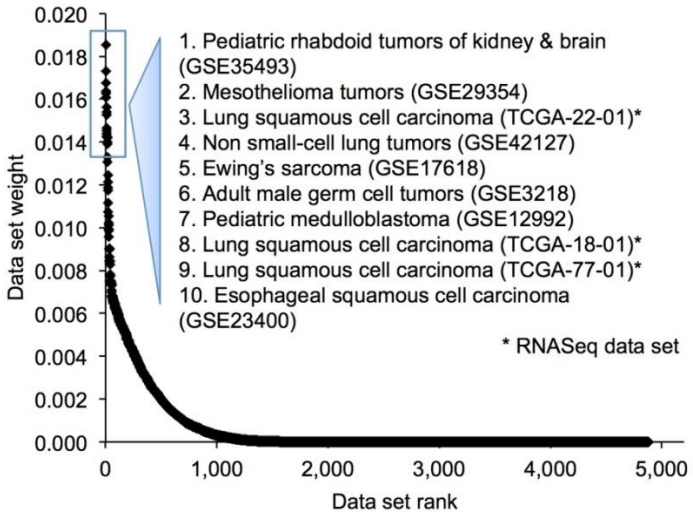


Figure 2.5 | Search results for the Hedgehog (Hh) signaling query GLI1 GLI2 PTCH1: the data sets prioritized for the query. The top 10 data sets among 5,000 data sets prioritized by SEEK are displayed in the insert. These data sets are weighted by the coexpression of the Hh query genes to indicate the abundance of aberrant Hh signaling activations.

2.3.5 Web interface.

The SEEK interface can visualize the aforementioned results—including the top-ranked data sets, genes and coexpression profiles—using flexible and interactive visualizations (**Fig. 2.6a**). 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 (**Section A.2**). Users can also examine the behavior of any gene in a given data set in detail through a condition-specific view (**Fig. 2.1a** step 5), 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. 2.6b**). 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 2.4**). We plan to regularly update SEEK's compendium as new microarray and RNA-seq data sets become publicly available.

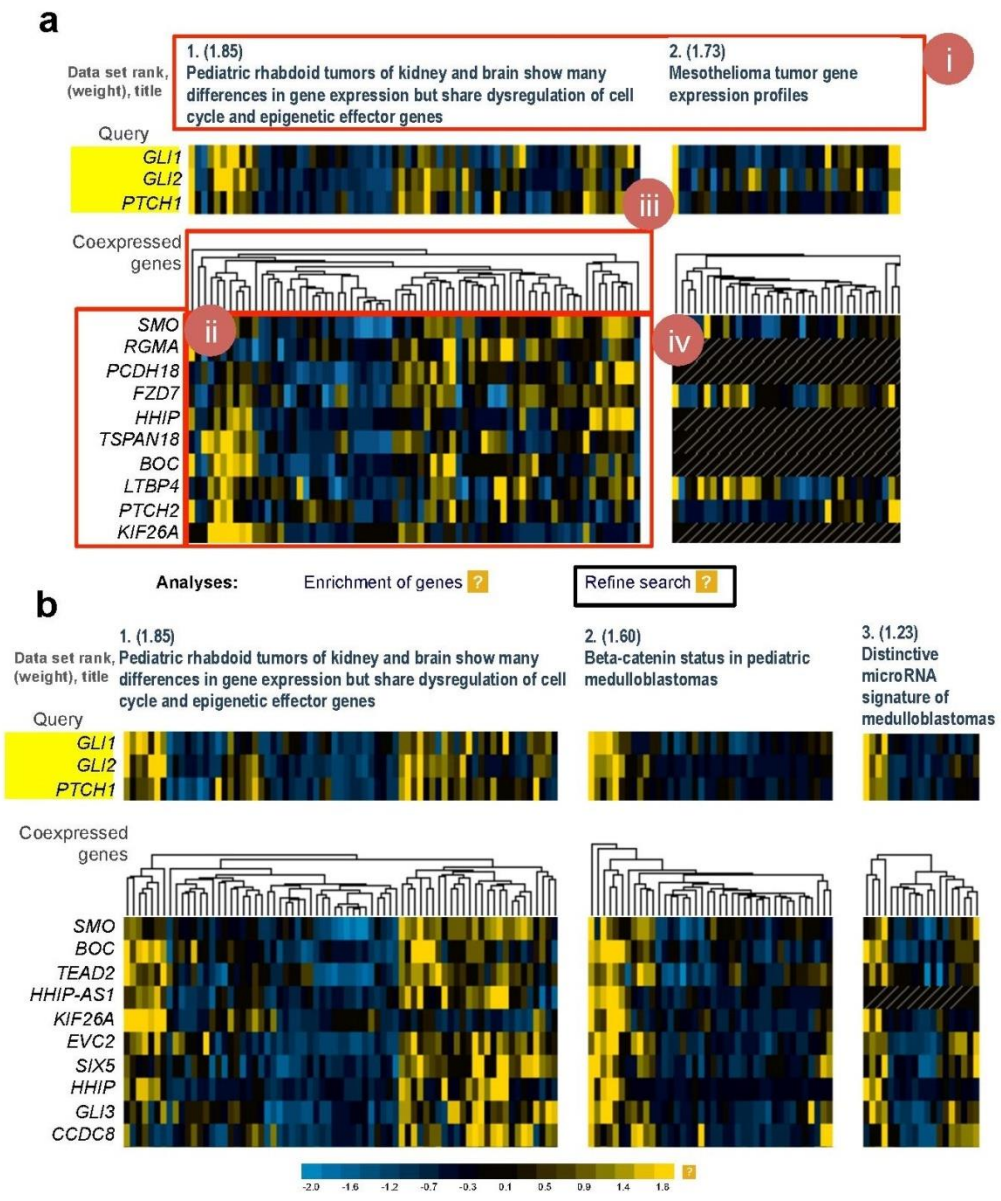


Figure 2.6 | Search results for the Hedgehog (Hh) query (GLI1, GLI2, PTCH1) and search refinement. (**a**) Data sets prioritized and genes retrieved for the query in the main result page, shown in expression view. The top-ranked data sets (i) and the coexpressed gene list (ii) are indicated. Conditions 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**) The "Refine Search" feature allows users to narrow the scope of their search through selection criteria including tissue, cell type or disease categories; platforms; or rank of data sets (**Section A.4**).

2.4 Methods

2.4.1 Data preparation and correlation normalization.

SEEK assembles its human gene expression compendium by obtaining data sets from NCBI's Gene Expression Omnibus (GEO) database[1] and the Cancer Genome Atlas (TCGA)[34]. The compendium consists of data sets from 41 platforms including 32 platforms from Affymetrix, Agilent, and Illumina and 9 RNA sequencing platforms (**Supplementary Data 2.1**). 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 (**Section 2.3.9** and **Supplementary Data 2.5**) to remove the systematic differences among data sets[12]. 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[62] is applied to convert the distribution of correlations to a normal-like distribution:

$$f_d(x,y) = \frac{1}{2}\ln\frac{1+r_d(x,y)}{1-r_d(x,y)} \qquad \text{Eq2.2}$$

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

$$z_d(x,y) = \frac{1}{std(f_d)}[f_d(x,y) - avg(f_d)] \qquad \text{Eq2.3}$$

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

The normalization procedure has been used in previous studies[33,39] 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 (**Fig. 2.7**). Note that SEEK also works well with other correlation measures, such as Spearman and biweight midcorrelation (bicor)[63] (**Fig. 2.8**). 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 (**Fig. 2.8**).

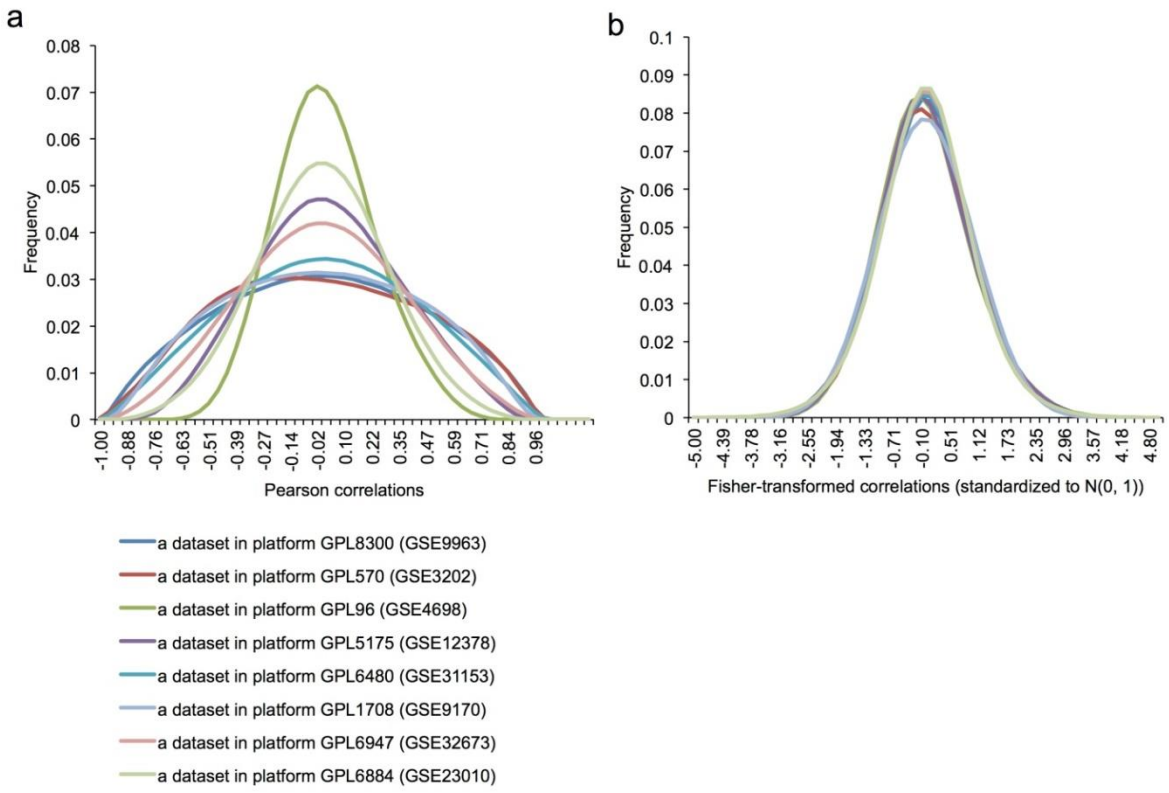


Figure 2.7 | Correlation standardization. (**a**) Prior to standardization, the distribution of Pearson correlation (r) (for all pairs of genes in the data set) is not directly comparable

across platforms. We picked a data set at random from each of 8 major platforms to illustrate this lack of comparability. (**b**) After the normalization of Pearson by Fisher's transform ($\ln[(1 + r) / (1 - r)] / 2$) followed by standardization, all of the selected data sets from these different platforms have been properly standardized to a $N(0,1)$ normal distribution.

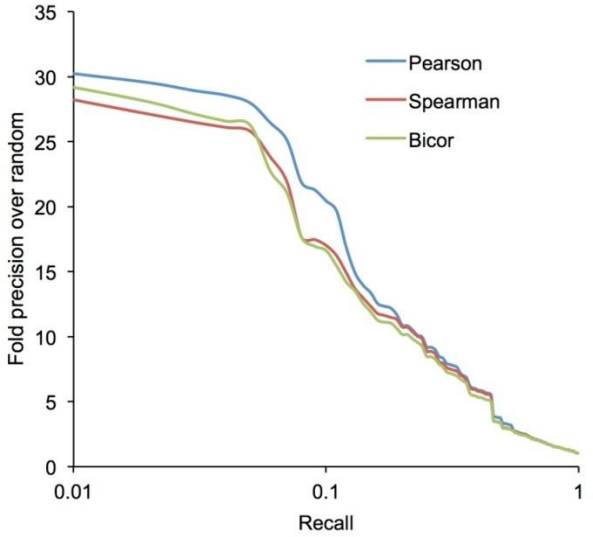


Figure 2.8 | Spearman and bicor correlation measures. Gene retrieval evaluations for GO slim biological process queries, searched using the SEEK algorithm. The correlation measure is varied: Pearson normalized, Spearman, and bicor. The data sets used are a group of 174 breast cancer tumor data sets.

2.4.2 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 hub gene bias correction on each data set *d*. The correction procedure is motivated by the observation that 'hub' genes[51,52] or well-connected genes in the coexpression network represent global, well-coexpressed processes[24] and can contaminate the search results regardless of query composition owing to the effect of unbalanced gene connectivity in a scale-free coexpression network[51,64–66], which can lead to nonspecific results in search or clustering approaches. To avoid the bias created by hub 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:

$$\widetilde{z_d}(g,q) = z_d(g,q) - \frac{1}{|G|}\sum_{x \in G} z_d(g,x) \qquad \text{Eq2.4}$$

where $\tilde{z}$ is the hub gene-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 hub genes 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[33]. 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, hub 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 (**Fig. 2.9**). Then the weight w of the data set is

$$w = \frac{1}{|Q|}\sum_{q\in Q}\left[(1-p)\sum_{t\in Q-q} p^{r(t,R_q)}\right] \qquad \text{Eq2.5}$$

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[67] 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[33], 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.

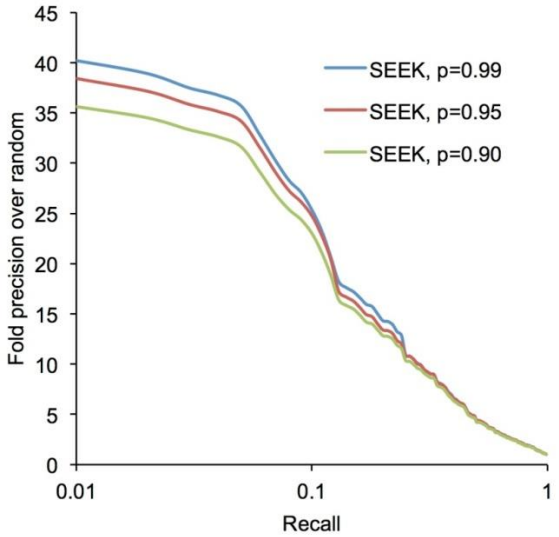


Figure 2.9 | Variation of the parameter p in the weighting formula. The parameter p that is used in **Eq. 2** in Section 2.4 Methods is arrived after testing values in the range from 0.90 to 0.99. At p=0.99, SEEK is most stable in retrieving genes across the 995 GO biological processes.

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:

$$M_{g,d} = \frac{1}{|Q|}\sum_{q \in Q} \widetilde{z_d}(g,q) \qquad \text{Eq2.6}$$

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 **F** is given by

$$F = \mathbf{M}_{G,D} \times \alpha \mathbf{w}^T, \alpha = \frac{1}{\sum_{d \in D} w_d} \qquad \text{Eq2.7}$$

Although previous research had some success with this formulation[33], 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 (**Section 2.3.3**).

The pseudocode for the entire SEEK search algorithm can be found in **Section 2.3.3**. The algorithm is robust to query composition (**Figs. 2.2** and **2.3**) and data set quality, including automatically downweighting data sets with substantial batch effects (**Section 2.2.3** and **Fig. 2.4**). Computer source codes are deposited in 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 hub gene 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.

2.4.3 Search algorithm pseudocode.

The SEEK search algorithm is a general search algorithm that works to integrate coexpressions from diverse data sets across platforms, and tackles the problem of incomplete gene ranking that arises from the diverse gene coverage across data sets. The algorithm is described in the following pseudocode:

Input: query genes (*Q*), genes in genome (*G*), data sets (*D*), correlation z-scores for pairs containing *Q* across data sets ($z_d(g,q), g \in G, q \in Q, d \in D$).
Variables: $M_{g,d}$, matrix of gene scores across *D*; $count_g$, vector of coverage of genes; w_d, vector of data set weight; F_g, vector of final gene scores.
Constants: α, β, θ.
Begin:
Compute $\tilde{z}_d(g,q)$ for each *g*, *q*, and *d*, as described in Eq. 2.4 (see **Section 2.3 Methods**).
//Hubbiness control
Initialize $M_{g,d} = 0$ for all *g*, *d*; $count_g = 0$ for all *g*; $w_d = 0$ for all *d*.
For each data set *d*: //Data set weighting
    Let *V* = set of genes that *d* contains
    If $|V| < \alpha$ or $|V \cap Q| < \beta$: //not enough genes, or query genes present
        continue
    Compute w_d as described in Eq. 2.5 (see **Section 2.3 Methods**).
    $M_{g,d} = \sum_{q \in V \cap Q} \tilde{z}_d(g,q) / |V \cap Q|$, for each gene $g \in V$
    $count_g = count_g + 1$, for each gene $g \in V$
End for
For each *g* in *G*: //Gene scoring
    If $count_g > \theta$: //sufficient data set coverage for *g*

Let U = set of data sets that contain g

$$F_g = \frac{1}{\sum_{d \in U} w_d} \sum_{d \in U} w_d M_{g,d}$$

Else:

$$F_g = -\infty$$

End if

End for

Sort F based on decreasing score, generate gene ranking (R_G)

Sort w based on decreasing weight, generate data set ranking (R_D)

Return R_G and R_D

End

The three thresholds α, β, θ are designed to maximize data utilization while keeping in check the biases introduced by incomplete data. α is the minimum number of genes required to be present in a data set. β is the minimum number of query genes that have to be measured in a data set, and θ is the minimum number of data sets required to contain a gene to include the gene in search ranking. Based on our experience, the following thresholds provide robust performance for a variety of queries and for large compendia with diverse data sets: α = 10,000, β = 2, $\theta = 0.5|D_w|$ where $D_w \subseteq D$ is the set of weighted data sets for the given query.

2.4.4 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.

2.4.5 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[68]. 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.

2.4.6 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[69] and BRENDA[70] 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 2.4**), to provide users with the ability to limit their search to such groups under the "Refine Search" feature of the website.

2.4.7 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 (**Section 2.3.8**). 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 2.2**) 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 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 statements) and NAS (nontraceable author statement). To select the GO slim set (**Supplementary Data 2.3**) 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. [53].

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.

2.4.8 Other search algorithms and implementations

Gene Recommender

Gene Recommender[43] is an algorithm that can retrieve relevant experiments based on the query, and use these experiments to retrieve query coregulated genes. It performs an experiment (or sample)-level weighting rather than a data set-level weighting. First it merges samples from all data sets to form a meta matrix Y_{ij} (i = gene, j = experiment). Given the query genes, the weighting algorithm is based on a number of criteria such as the gene expression of the query genes, and the expression variance of the query in each experiment. The original matrix Y_{ij} of n genes by p experiments (or samples) is transformed to ranks Y_{ij}', where

$$Y'_{i,j} = \frac{R_{i,j} - \frac{p_i + 1}{2}}{\frac{p_i}{2}} \qquad \text{Eq2.8}$$

R_{ij} is the rank of i among Y_{ij} for $j = 1 \ldots p$, and p_i is the number of experiments containing gene i. The experiment scoring is calculated as:

$$Z_j = \sqrt{k_j} \frac{avg(Y_{Q,j})}{\sqrt{V_{Q,j} + \frac{1}{3p^2}}} \qquad \text{Eq2.9}$$

where $avg(Y_{Qj})$ is the average expression (Y_{ij}´) over query genes Q in j, V_{Qj} is the variance of the query in j, k_j is the number of genes in j. This scoring prefers experiments with a tight clustering of the query genes with high expression, low variance. In order to use the experiment scoring to return query-coregulated genes, a threshold ε defines the number of relevant experiments (with top scores), so the final score of gene i is calculated as: S_i = the mean of ($avg(Y_{Qj}) \times Y_{ij}$´) over all relevant experiments j. The parameter ε is set at 0.05 (or 5% of the total experiments).

MEM

The MEM algorithm[42,71] assumes that the query is a single gene q. For each data set j, it first transforms the correlations containing the query gene into ranks, so that each gene has a rank n that represents the n-th correlated gene to the query. Ranks are normalized to [0, 1] by dividing each rank by the maximal rank in each data set. Then, the ranks are transformed so that for each gene g_i, we generate a rank vector $r(q, g_i) = [r_1^i, \ldots, r_m^i]$, where r_j^i is the position of g_i in the query on data set j, and m is the number of data sets. MEM assumes a null hypothesis where in a model rank-list the genes are randomly permuted, and $r(q, g_i)$ contains uniformly distributed ranks. It reorders $r(q, g_i)$ in order to obtain a vector of order statistics, $r_{(1)}^i, \ldots, r_{(m)}^i$ where $r_{(1)}^i$ is the smallest, and $r_{(m)}^i$ is the largest value in $r(q, g_i)$. Assuming null hypothesis, it then calculates the probability from binomial distribution, $b(k)$, that the order statistic $r^*_{(k)}{}^i$ is smaller or equal to $r_{(k)}^i$, where $r^*_{(k)}{}^i < r_{(k)}^i$ is generated by null model:

$$b(k) = \sum_{j=k}^{m} \binom{m}{j} \left(r_{(k)}^i\right)^j \left(1 - r_{(k)}^i\right)^{m-j} \qquad \text{Eq2.10}$$

The score of each gene is then the *p*-value:

$$p(g_i) = \min\left[b(k) \text{ for each } k \text{ in the range } [0, m]\right] \quad \text{Eq2.11}$$

Intuitively, if the rank vector of a gene contains a large number of small ranks (which means that the gene is consistently correlated to the query in large number of data sets), the distribution of $r(q, g_i)$ will be heavily biased towards the small values and different from the uniform distribution.

Combined data set correlation

Combined data set correlation is a simple approach for combining data sets for correlation analysis. Data sets in the compendium were first concatenated into one matrix, forming a combined data set. Genes were next ranked according to the average of Pearson correlations to the query genes in this combined data set. Because different constituent data sets may include different sets of genes, we calculated correlation only for pairs of genes where each gene in the pair is present in at least 50% of the arrays in the combined data set, yielding a reasonable set of 17,689 genes being ranked. Where the array coverage of two genes differs in the combined data set, we chose the entire set of arrays with values present for both genes in the matrix for computing their correlation.

2.4.9 Building compendia: raw data processing

Each microarray platform had a relatively accepted pipeline for processing its data sets. Briefly, for Affymetrix platforms, we normalized each array using Robust Multi-array Average (RMA) [72], which ensures that the distribution of expression values per array is

the same within each data set. We note that SEEK also performs similarly well for data sets processed with other techniques, such as MAS5. For Agilent, there are two types of arrays: single-channel and dual-channel arrays. Dual channel arrays are designed for measuring fold-change between case and control conditions. In dual-channel arrays, individual arrays are normalized by loess normalization (Zahurak *et al*[73]). Next, we calculated the log-2 Cy3 over Cy5 fold change and applied between-array normalization, which is essential in two color array analysis, as it normalizes channel intensities and log-ratios to be comparable across arrays. Single-channel arrays were normalized by within-data set quantile normalizations. The above analysis was done using the Bioconductor R and limma package[74] following the guide in Chapter 6 in the limma manual[75]. For Illumina BeadChip platforms, we limit to the set of data sets that have no missing probe measurements, termed "unnormalized" raw data obtained from the Gene Expression Omnibus. We normalized the arrays using quantile normalization[76] as this is the recommended approach in the study Ritchie *et al*[77]. This use of consistent processing pipeline across all data sets within a given platform helps remove systematic differences between data sets[12].

For data sets from the RNA sequencing platforms, we obtained 5,085 RNASeq samples that were pre-processed level-3 data from TCGA[34]. Discussion of the processing is found in [78,79]. On a high level, for these TCGA samples, we use normalized counts, which are the raw counts divided by the 75th percentile of each column multiplied by 1000 (known as the upper-quartile normalization[80]). TCGA samples have been split into 224 data sets according to unique 'disease type, sample source' pairs. We also extracted 54 RNASeq data sets from GEO that have been processed by submitters of the data sets.

These data sets have been published in their associated studies (**Supplementary Data 2.5**), where the processing of each data set is discussed. We use results summarized in raw counts format, and we further performed upper-quartile normalization on counts data to be consistent with the TCGA samples. Final measurements are normalized by $\log_2$(1+normalized_counts).

The gene expression data sets normalized using the abovementioned procedure are publicly available for download on the SEEK website.

3 MODSEEK: TOWARDS A TARGETED, DATA-DRIVEN VIEW OF MODEL ORGANISM TRANSCRIPTOMES

3.1 Abstract

SEEK's usefulness in searching large collections of human gene expression data has led us to develop the ModSEEK system, which targets gene expression data from multiple organisms. The objective of this multi-gene query-focused system is to encompass the ability of facilitating large-scale coexpression-based retrieval and analyses in 5 commonly studied model organisms. This expansion dramatically increases the biological diversity among datasets, which contain many experimental types that were previously under-represented in the human data. ModSEEK is freely available at http://seek.princeton.edu/ modSeek/.

3.2 Introduction

Modern high-throughput technologies have generated a myriad of expression datasets for a diverse set of organisms. These datasets provide whole-transcriptome view of organisms in various conditions such as knockdown, knockout, overexpression, and other perturbations, and are a great resource for the study of gene functions. Despite their data growth, exploring these datasets for experimental planning and hypothesis generation has remained to be difficult. More recently, data-driven approaches and visualization systems have begun to leverage the whole body of datasets for achieving coexpression analyses in real-time. There are systems that support the coexpression mining of single-gene queries [42,81], mining within Affymetrix platforms [42,81] and RNAseq datasets [82], and mining within

a disease domain of interest (e.g. angiogenesis network PADPIN [83]). We have also developed the system SEEK [84] that is capable of prioritizing human datasets, and of finding gene functions through coregulated genes using multi-gene queries and thousands of datasets from a number of platforms. SEEK further demonstrates that increasing the size of the compendia can lead to an increase in the accuracy of retrieving functionally related genes, and that its search algorithm is capable of overcoming the heterogeneity of the large compendia [84]. To date, this integrative search and visualization system has been available for human. However, in model organisms, achieving the tasks of prioritizing expressions of genes, datasets, and conditions in a query-dependent manner and mining coexpression in full scale are still hurdles for many experimental biologists. To address those challenges and provide integrated search across organisms, we developed a system called ModSEEK. ModSEEK is the first multi-organism coexpression analysis system for integrated cross-platform search of coexpressed genes, relative to a query of interest. The system extends SEEK's data compendium from only human previously to five commonly studied model organisms and fully benefits from the robustness of SEEK in retrieving coexpressed genes. Additionally, ModSEEK offers full dataset prioritizations, and fast statistical testing of coexpression association between candidate genes.

3.3 Methods

3.3.1 Source data and preparation

Our expression compendia for five commonly studied model organisms consist of microarray and RNASeq datasets obtained from Gene Expression Omnibus [1]. The number of datasets is provided in **Tables 3.1** – **3.2**, and a listing of datasets is provided on the ModSEEK website. These datasets were normalized to gene-level expression

measurements, and were largely processed by following the consistent data processing protocol mentioned in SEEK [84]. Pearson correlations were calculated and Fisher-transformation was applied to obtain a gene-by-gene correlation matrix for each dataset. The final correlation values within a dataset follow a standard normal distribution and the distribution is comparable across datasets. The correlation matrices and the query are the input of the search algorithm that is described in the next section.

Table 3.1 | ModSEEK repository of expression datasets

	No. of samples	No. of datasets*	No. of platforms	No. of genes	No. of correlation values** (billions)
M. musculus	54,299	2,438	10	22,116	597
D. melanogaster	5,101	351	3	13,521	32
C. elegans	2,826	225	7	18,010	36
D. rerio	1,372	95	3	19,602	18
S. cerevisiae	4,566	321	3	6,976	8

*includes both microarray and RNAseq. Sequencing datasets in each organism are 244 (mouse), 51 (fly), 19 (worm), 21 (yeast).
**number of gene-gene correlations totaled for all datasets in the compendium. These correlations are pre-computed and stored into database.

Table 3.2 | ModSEEK types of datasets

	Knockout, transgenic datasets	Knockdown datasets	Cancer datasets
M. musculus	547	61	110
D. melanogaster	40	30	-
C. elegans	21	15	-
D. rerio	20	9	-
S. cerevisiae	13	-	-

3.3.2 Search algorithm

The ModSEEK single-organism search algorithm uses gene-hubbiness correction and robust dataset weighting as previously published[84]. Briefly, for each query *Q* defined by the user, the algorithm calculates a dataset weight that reflects the degree of query coexpression in the dataset [33,84]. This dataset weighting step adopts cross-validations, and a scoring approach that is inspired by rank-biased precision [67]. Each dataset's weight is:

$$w = \frac{1}{|Q|}\sum_{q\in Q}\sum_{q'\in Q-q}(1-p)p^{rank(cor(q',q),R_q)} \qquad \text{Eq3.12}$$

where p is a parameter from rank-biased precision that determines the contribution of rank to the validation score (p is fixed at 0.99), R_q is the ranked-list of correlations between every gene in the genome and q, $rank(cor(q', q), R_q)$ is the rank of correlation for the pair (q', q) in R_q, and cor is the hubbiness-corrected z-scored correlation. Essentially, this scoring function repeatedly examines how well a single query gene can retrieve the rest of the query in the dataset and evaluates it against the background of all correlations which are associated with that single gene. With the weight determined for every dataset, the next step, the gene retrieval step, merges the coexpression scores together from many datasets to produce a single score for each gene

$$s_g = \frac{1}{\sum_{d\in D} w_d}\sum_{d\in D}\left(\frac{w_d}{|Q|}\sum_{q\in Q} cor_d(g,q)\right) \qquad \text{Eq3.13}$$

where s_g is the coexpression score of g to the query, w_d is the weight from earlier Eq3.12. Note that the correlation $cor_d(g, q)$ is hubbiness corrected.

3.3.3 Query coexpression P-value estimation

Previously in SEEK, the query coexpression *P*-value is estimated empirically based on randomly selected queries of matched size. However, the empirical *P*-values may fluctuate due to error and the accuracy of *P*-values depends on the number of random queries for building the null distribution. Recently, an estimation procedure based on generalized pareto modeling allows one to estimate accurate *P*-values with fewer permutations[85]. In a similar spirit, we adopted generalized pareto distribution (GPD) to model the distribution of coexpression score observed from null queries.

Given a query, the goal is to provide a *P*-value for query coexpression score (QCS) for each dataset to represent how likely that coexpression is to arise by random chance. The first step is to build a null distribution of QCS per dataset per query size, as we wish to have a distribution sensitive to query size. For this, we simulated 5000 random queries for each possible query size (from 2 to 100 genes) and calculated QCS for each random query using the dataset weight formulation in Eq3.12. Next, we recognize that only random queries with the highest QCS values are most meaningful for estimation (corresponding to the rarest events), so we extracted only the right 5% portion of the distribution, which represents the highest extreme values, for GPD modeling. Next, we smoothed them using GPD function. This GPD smoothing is essential to estimating more accurate *P*-values with a limited number of permutations [85]. In the past, GPD has been used to model extreme values in weather, failure detection, insurance, and financial applications.

Specifically, to model the right 5% tail of distribution of permutation values from random queries using GPD, we first obtained the excess (or exceedance) $M_{0.05}$:

$$M_{0.05} = \{\log(n) - \log(n_{0.05}), \forall n \in N, n > n_{0.05}\} \quad \text{Eq3.14}$$

where $n_{0.05}$ is the top 5-th percentile value of the null distribution, and N is the set of values in the null distributions. Because any function that preserves monotonicity can be applied on N, we chose log-function in order to keep N within a narrow range. Next, we tried a number of ways to estimate the GPD parameters, *shape* and *scale*, such as the gPdtest R library [86]. Here, we assume *shape*<0 based on the fit of the null distribution data. This presents a difficult problem because when the *shape* is smaller than 0, GPD is light-tailed and subject to a finite range [85], preventing us from estimating *p*-values for large QCS scores not observed in the random trials. To resolve this issue, we need to choose a bound for QCS score when *shape*<0. This bound can be derived in that there is a maximum theoretical limit to QCS, *m*, that is dependent on query size and can be derived exactly (see derivation procedure below). We also note the relation *m*=-*scale*/*shape* (the maximum bound of GPD for the case *shape*<0) (see below). These constraints have been built into the estimation procedure accordingly, similar to what Villasenor-Alva *et al* have done.

Shape and scale parameter exact formula derivation: The expected value of generalized pareto distribution is given by $m = \mu + \gamma/(1 - \alpha)$ where μ, γ, and α are the location, shape, and scale respectively. Here, μ=0, and an estimate of m is computed from null data: $m=1/n\Sigma T$, where T is the exceedance values. We note that the maximum achievable coexpression scores for a query of size q is: $r= \Sigma (1 - p)p^{rank}$, for $rank$ = 0, 1, 2, …, q– 2, and p=0.99. Converting into log-scale, the corresponding exceedance for r is $T_{max}=\log(r) - \log(n_{0.05})$,where $n_{0.05}$ is the 95th percentile value of the null distribution r. When γ<0, the range of the GPD modeled values are restricted to $0<T<-\alpha/\gamma$. The

maximum GPD-modeled value is T_{max}, which sets the bound of the model. Therefore, T_{max}= - α/γ. By rearranging m, we get $\gamma = m(1 - \alpha)$. Finally, we arrive at the parameter estimates $\gamma = m/(m - T_{max})$and $\alpha = -\gamma T_{max}$.

After the *shape* and *scale* parameters have been estimated, we computed the *P*-value given a QCS score *x* using the cumulative distribution function of the GPD:

$$P = 0.05\left(1 + \frac{s(\log(x) - \log(n_{0.05}))}{a}\right)^{-\frac{1}{s}} \quad \text{Eq3.15}$$

where *s* is the shape, *a* is the scale parameter. We assume $x \geq n_{0.05}$. If $x < n_{0.05}$, which means that $P > 0.05$, we use the original distribution *N* to estimate *P*-value that is accurate to every 0.01.

Why is GPD suitable? The choice of GPD is suitable because: 1) in a null query, QCS is generated by a process of summing the exponentially weighted scores p^{rank} for randomly sampled *rank*, 2) this sum is more extreme than the distribution of the maximum of randomly drawn samples (see Eq3.16 below), which has a known extreme value distribution to which GPD can be applied to model the distribution.

$$\sum_{rank \in R} p^{rank} > \max(P), P = \{p^{rank}, rank \in R\}, R \sim Uniform \quad \text{Eq3.16}$$

where *R* is a set of randomly selected ranks drawn from a uniform distribution (from rank 1 to |*G*|, the size of genome). If |*R*| is small, then $\sum_{rank \in R} p^{rank} \approx \max(P)$, which is close to extreme value distribution. If |*R*| is large, then $\sum_{rank \in R} p^{rank}$ will be close to a normal distribution.

3.3.4 Evaluation of GPD fit to null coexpression distribution.

We constructed quantile-quantile goodness of fit plot between GPD-estimated and observed exceedance values (see Eq3.14). R^2 is calculated.

3.3.5 Large-scale gene-retrieval evaluation.

To perform a large-scale functional evaluation, we obtained experimental gene-sets for 274 KEGG pathways and 609 Gene Ontology biological processes from all organisms. Queries of length ranging from 2 to 20 genes were constructed by a random selection of genes from each term and the remaining genes were used for held-out evaluation. We plotted a precision-recall (PR) curve for each query based on its retrieval result and then averaged all the queries' PR curves at each recall point in the interval [0.01, 0.99] (step size of 0.01) to build a system's PR curve. To calculate fold precision over random, we generated a random retrieval ranking list by shuffling the genes' rank positions from a true ranking list and then computed the ratio of the true ranking's precision over that of the random ranking.

3.3.6 MeSH enrichment

All datasets have been systematically annotated with keywords in the Medical Subject Heading (MeSH) controlled vocabulary [87] (our focus is primarily on the disease, anatomy, and experimental branches). To do this, we applied a text-mining algorithm on each dataset's title, description, as well as its containing samples' characteristic field. Just as genes may be enriched for GO or KEGG terms, a prioritization of datasets may show possible enrichment of tissue or disease MeSH concepts in the top ranked datasets. To calculate MeSH enrichment significance, we used a hypergeometric distribution:

$$P(X \geq k) = \sum_{i=k \ldots K} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \quad \text{Eq3.17}$$

where *N* is the number of datasets with annotations, *K* is the term size, *n* is the depth in the dataset list, and *k* is the number of overlapped datasets. To correct for multiple hypothesis testing, we ranked all terms by their *P* value and applied the Benjamini-Hochberg[88] FDR procedure to derive FDR-controlled *P* value.

3.4 Results and discussion

3.4.1 Dataset composition

ModSEEK is composed of a large number of gene expression datasets for 5 extensively studied model organisms – *S. cerevisiae*, *D. melanogaster*, *C. elegans*, *M. musculus*, and *D. rerio*. These organisms are chosen due to the wide availability of datasets and the organisms' popularity. Together, the datasets in the compendium represent the collective knowledge of over 2,273 publications. They cover a diverse range of topics including development, apoptosis, neuroscience, and a large variety of experimental conditions, such as mutations, RNA interference, knockouts, knockdowns, etc (**Tables 3.1 – 3.2**). The large compendium provides biologists with a good platform to examine their genes of interest in diverse experimental perturbations and tissue contexts. The multi-organism compendium contains MeSH-annotated datasets categorized into several main experimental types (**Fig. 3.1**). The extracted experimental vocabulary consists of a large number of knockout, transgenic studies (**Fig. 3.1**), cell culture studies, and developmental stage related conditions.

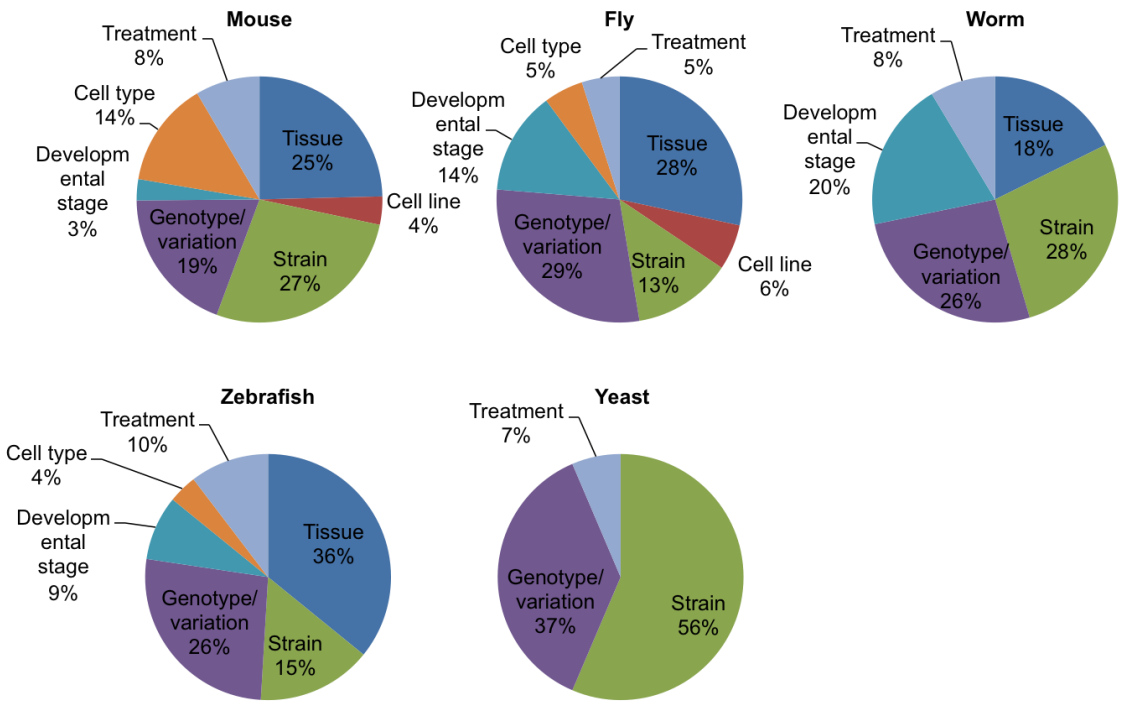


Figure 3.1 | Proportion of datasets with the different types of characteristics.

3.4.2 ModSEEK description

First, datasets are consistently processed and correlations are calculated, normalized for each dataset to ensure that differences and biases of individual datasets are accounted for (see **Section 3.3 Methods**). Then, ModSEEK applies our previously developed and query-dependent dataset weighting algorithm to discover relevant datasets, on the criteria of the query coexpression strength in each dataset (see **Section 3.3 Methods**). This dataset weighting filters away poor quality datasets that are unlikely to exhibit query coexpression, and at the same time achieves query context-specificity in the retrieval of coexpressed genes. After that, this algorithm generates an integrated, context-dependent

list of coexpressed genes, from which the expression profiles are visualized in the web result page.

3.4.3 Evaluations

We evaluated ModSEEK's gene retrieval ability for a broad range of GO biological processes. ModSEEK consistently outperforms five other alternative search approaches (variance, meta-dataset, SPELL, Gene Recommender, MEM) even when we applied them to the same compendium as ModSEEK (**Fig. 3.2**). Note that none of these approaches is available across all the ModSEEK included organisms. The variance approach gauges at which datasets exhibit the largest expression variance of the query genes, as highly varied genes are likely to be true signals in the dataset [89]. The approach's weakness, however, is in its bias towards datasets with high variance baseline. High variance could arise by virtue of genes being highly expressed (not necessarily related to the query genes), because of the possible mutual dependence between variance and average expression [90]. ModSEEK outperforms SPELL in yeast and on a per-GO-term basis, where ModSEEK holds fold-improvement of 1.2X in area under the precision recall curve (AUPRC) (**Fig. 3.3**). For meta-dataset correlation, this approach merges all datasets in the compendium to form a meta-matrix, and calculates correlation on it. It is evident that ModSEEK's performance lead extends across organisms. The mean fold-over-random precision values at 10% recall across organisms are: 38-fold for fly, 58-fold for yeast, 33-fold for worm, 34-fold for mouse, and 14-fold for zebrafish.

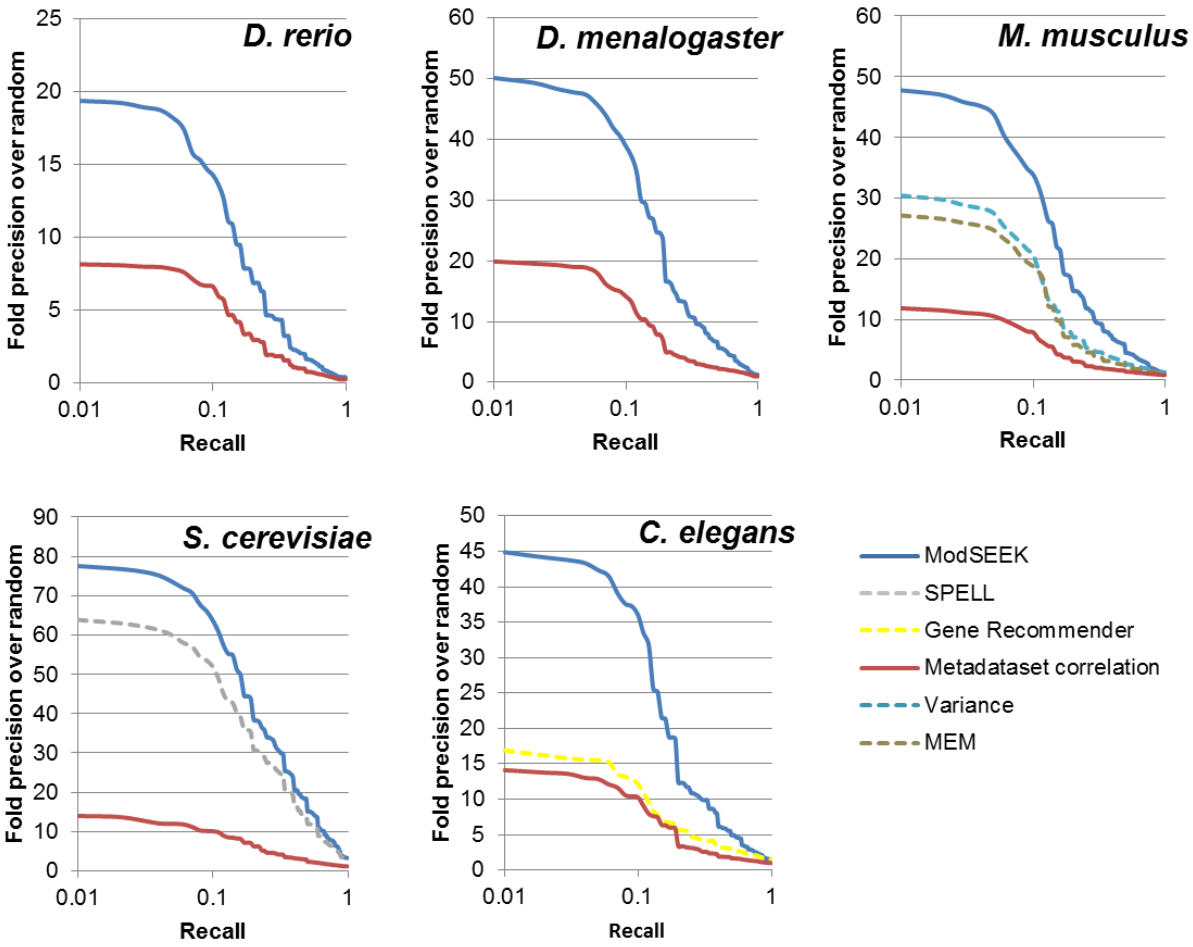


Figure 3.2 | Functional evaluation comparison between ModSEEK and other systems. This comparison evaluates each system's ability to retrieve functionally related genes from GO biological process terms in each model organism. Where there is an existing search method for the organism, there is a line to represent it in the plot. If there is no existing method, the standard baseline method meta-dataset correlation is plotted. Performance in fold-precision-over-random is first calculated from averaging performances of queries per GO term, then it is averaged across terms in the organism.

Examining the dataset weight plot with increasing rank (**Supplementary Fig. 3.1**) reveals two common types of queries in the study of biological processes. The first one is the ubiquitous type, such as proteasome, in which the query is coregulated in a large majority of datasets (e.g., 234 out of 400 datasets in **Supplementary Fig. 3.1**). The

second type (**Supplementary Fig. 3.1**) is the specific type where the coregulation is expected to occur in only a subset of datasets that possibly originate from a specific context. ModSEEK is robust to both types of queries and also robust to discovering genes from 275 collected KEGG pathways (in addition to GO). In our KEGG evaluation results, the number of pathways with retrieval average fold-precision over random (at 10% recall) greater than 10-fold are respectively: 118 / 253 (human), 141 / 254 (mouse), 69 / 96 (fly), 61 / 88 (worm), and 62 / 76 (yeast), where the denominator denotes the total number of pathways in each organism. Queries from proteasome, ribosome, and DNA replications tend to be ubiquitous, where these are coexpressed (with $P<0.05$) in 146, 159, 124 datasets respectively in fly, representing approximately 50% of the fly's compendium size. Amazingly, specific-type queries from pathways, such as endocytosis, arachidonic acid metabolism, sphingolipid metabolism, hedgehog signaling pathway, can be retrieved equally well with the search algorithm. Specifically, ModSEEK achieved precisions of 10-fold, 18-fold, 12-fold, 15-fold respectively for these pathways, despite the fact that these have only low representation in the fly compendium and they are coexpressed in only 33, 26, 28, and 30 datasets. Thus, ModSEEK is robust to both ubiquitous and specific queries.

3.4.4 Dataset prioritization and coexpression testing

An important usage example of ModSEEK is finding tissues, conditions that are associated with a gene set. This task can be accomplished by ModSEEK's dataset prioritization function. Although SEEK is also capable of performing coexpression testing, what is new in ModSEEK is that it builds random coexpression model that is specific to each "dataset, query-size" pair based on permuted queries, and applies

generalized pareto distribution (GPD) modeling to provide an accurate fit of null coexpression score distribution (see **Section 3.3 Methods**, and see **Supplementary Fig. 3.2** for goodness-of-fit assessment of an example dataset). This modeling at the same time permits very quick coexpression testing. Dataset prioritization has thus benefited significantly from this procedure.

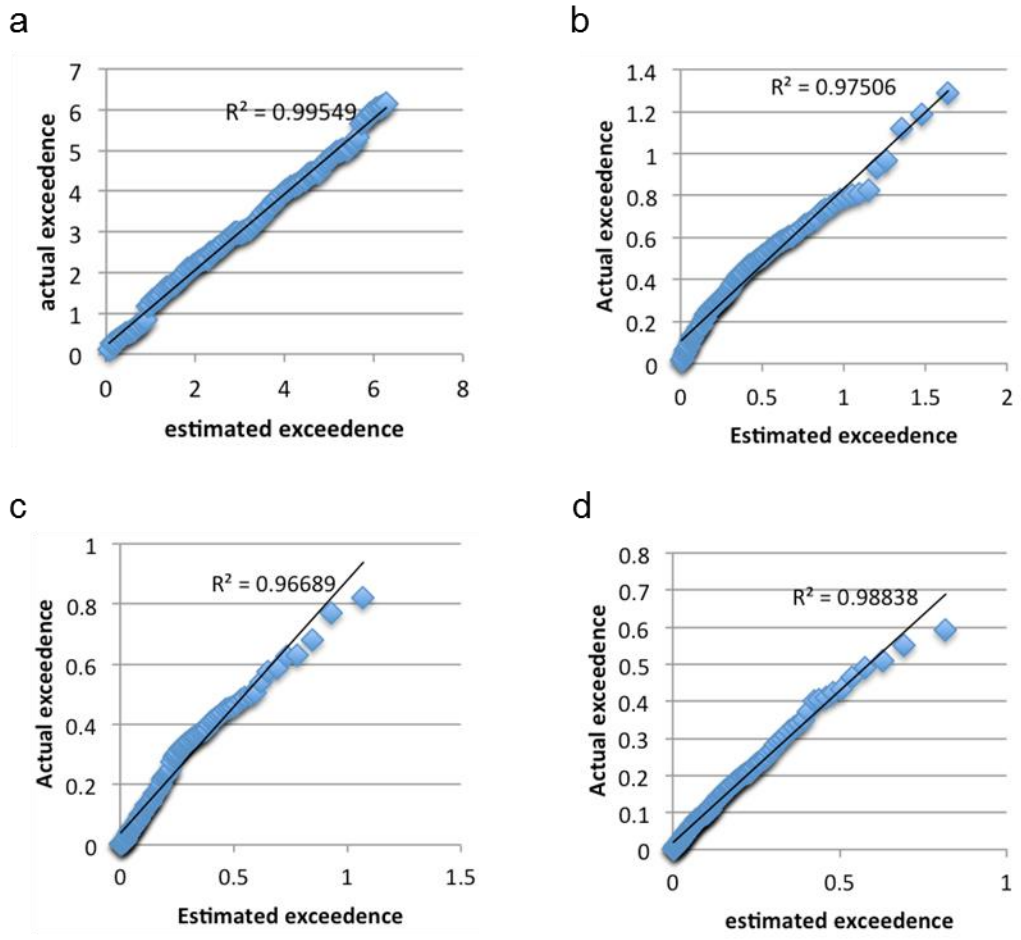


Figure 3.4 | Quantile-quantile goodness of fit plot for GPD fitting of null query coexpression distribution. The plot displays the estimated and observed exceedance values for an example mouse dataset GSE5876, that is used for modeling permuted queries of size 2 (**a**), 5 (**b**), 10 (**c**), and 20 (**d**). Tail exceedance values are defined as the set $T > 0$, $T = \{\log(QCS) - \log(QCS_{0.05})\}$, where $QCS_{0.05}$ is the 95th percentile of null coexpression score distribution. Estimated values are derived from GPD fitting of coexpression score distribution.

3.5 Conclusions

We described ModSEEK, a system that for the first time enabled multi-organism coexpression search for five most commonly studied model organisms - mouse, fruit fly, zebrafish, worm, yeast. The system allowed us to explore the entirety of the expression compendium within a model organism, by providing a full prioritization of all available expression datasets and a full ranking of coexpressed genes to the query.

ModSEEK is robust in diverse processes. As expression datasets continue to grow in number, large-scale weighted integration of coexpressions will become increasingly appreciated due to its unbiased gene retrieval, on-the-fly dataset weighting, and its genome-wide assessment. We believe that ModSEEK will be well-suited for assigning functions to unannotated genes in model organisms, especially in worm, where currently other experiments are scarce for hypothesizing the function of many uncharacterized genes. ModSEEK can also be useful for revealing tissue/cell-type or other context-dependent roles of existing genes.

ModSEEK houses several important community-standard tissue/localization/cell line expression datasets such as the GNF Mouse Atlas, Mouse Brain Atlas, FlyAtlas, which are very useful for checking expression values of single and multiple genes of interest. We will continue to regularly maintain and update ModSEEK's compendia as more datasets are collected in the public repository. We believe that ModSEEK will undoubtedly play an important role in facilitating gene function discovery in the model organisms.

4 CROSS-ORGANISM GENE RETRIEVAL

4.1 Abstract

The previous chapter has demonstrated the power of ModSEEK in single-organism gene retrieval studies. Nonetheless, the true strength of the ModSEEK system lies in its ability to harness information from other model organisms' gene expression data not only for coexpression comparisons but also for boosting of gene retrieval performances from single-organism studies. In addition, bringing the cross-organism capability to ModSEEK would permit the graphic visualization of coexpression patterns between members of the orthologous group in a pair of organisms, which can explicitly illuminate key members of the query coexpression context.

4.2 Introduction

Our knowledge about biological processes is unequally distributed: some biological processes are well studied in one organism, but not in others. This phenomenon is clearly manifested in the number of gene annotations per biological process. For example, cell cycle genes are better studied in yeast mutant strains, thus they receive more annotations; and some signal transduction processes such as phototransduction have been traditionally studied in fruit flies. Annotations of understudied processes can be notably limited in an organism due to the scarcity of experiments in the organism or the possibility that the experimentation cannot be done altogether. This limitation poses great challenges to gene function prediction. Park *et al*[91] addressed this problem with a cross-organism gene function prediction algorithm called function knowledge transfer whereby gene

annotations are transferred between organisms via function-similar orthologs in a machine-learning framework. In this work, we propose an alternative unsupervised strategy to tackle this problem. Our approach is based on whole organism data compendium and coexpression, and our goal is query-sensitive cross-organism gene retrieval through search algorithms. As ModSEEK is based on coexpression, which is relatively unbiased to training data and prediction algorithm, there is a clear need to identify which coexpressed genes are simultaneously coexpressed to the query in a pair of organisms. This chapter highlights and addresses this need.

4.3 Methods

4.3.1 Definitions

We are given a pair of organisms *S* and *T*, the query gene set *Q* in *S*, and the corresponding orthoquery *Q'* in *T*. Each gene g_X in *S* receives a coexpression score to the query *Q*, and similarly gene g'_X in *T* has a coexpression score to *Q'*. An orthogroup (**Fig. 4.1a**), defined across two organisms, groups together sequence similar paralogs (duplicated genes within the same organism) and orthologs (sequence-similar genes across organisms) based on sequence alignment score (bidirectional best hit), species tree phylogeny, and other evidences. For our purpose we use externally constructed orthogroup definitions provided by OrthoMCL[92] and InParanoid[93]. The goal is to identify co-similar orthogroups that are coexpressed to the query, whereby at least one member of the orthogroup from *S* is coexpressed with *Q*, and at least one member from *T* is coexpressed with *Q'* (**Fig. 4.1b**).

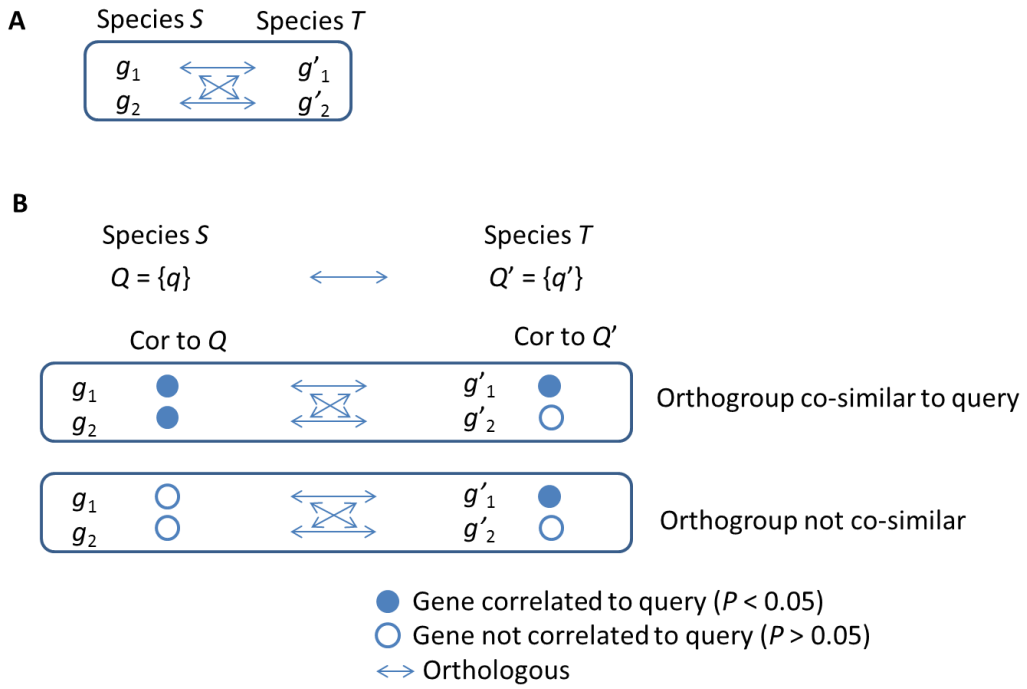


Figure 4.1 | ModSEEK combines orthology and coexpression evidences to identify orthogroups with co-similar orthologs. (**a**). Orthogroup definition. (**b**). Orthogroup co-similar to the query in two organisms. Each member g_1, g_2, g'_1, g'_2 has a coexpression score to the respective organism's query Q or Q'. Then, the orthogroup is checked if there is at least one member per organism that is significantly coexpressed to the organism's query (Q or Q'). If this is satisfied, it is considered a functionally co-similar orthogroup to the query (first orthogroup illustrated). The second orthogroup illustrated does not have species S containing at least one coexpressed member, so it is not co-similar.

The above formulation offers much-needed flexibility in the definition of functionally similar orthologs as it recognizes that the expression of orthologs can vary with the query context. Accordingly it uses the query context to identify function-similar orthologs, i.e. specific pairs of orthologs that exhibit similar coexpression to the orthoqueries. Finding co-similar orthogroups given two coexpression gene rankings is straightforward and can be done relatively quickly. Typically, users first define a *P*-value

threshold with which to define significant correlation linkages within an orthogroup. In SEEK and ModSEEK, this *P*-value reveals whether a gene is specifically correlated to the query and not to other random query genes. This definition of *P*-value carries specificity and is much more stringent than permutation-defined *P*-value that asks a different and easier question – the question of whether the correlation is significantly different from that of a randomly shuffled expression matrix. In ModSEEK's web interface, the co-similar orthogroups are defined with the default *P*-value threshold 0.05.

4.3.2 Usage scenario

To illustrate how these orthogroups are used, we can envision the following real-world scenario (**Fig. 4.2**). A user first presents the query in an organism say *H. sapiens*, and selects what he wishes to compare and search in *D. melanogaster*. Next, within-organism coexpressed genes are retrieved independently of each other and the results of two searches are grouped into orthogroups. Subsequently they are classified into groups that show whether or not they are co-similar orthogroups to the query. Those co-similar ones are selected for visual presentation to the user, while illuminating key members of pathways and processes (**Fig. 4.2**).

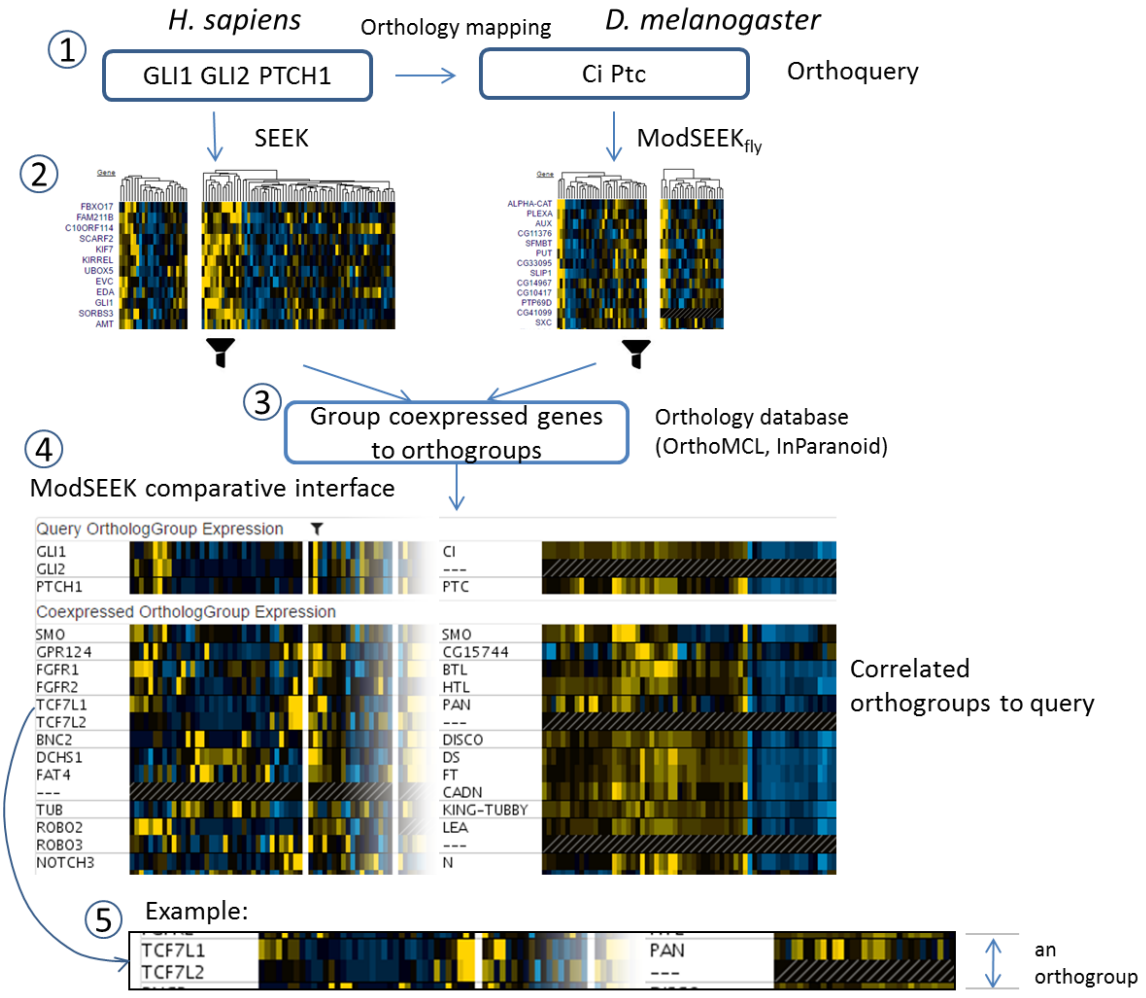


Figure 4.2 | Cross-organism search process. For illustration purpose, a human-fly comparison is used. (**1**) A human query is entered. ModSEEK automatically maps it to its orthologs in fly. (**2**) Each query is searched independently to get human and fly coexpressed genes. These are further filtered by correlation *P*-value. (**3**) The filtered coexpressed genes are combined to form orthogroups, pre-defined by OrthoMCL and InParanoid. (**4**) The results, orthogroups co-similar to the query, are displayed in a comparative web interface. See **Section 4.3.1** for definition of expression co-similarity. (**5**) An example of orthogroup. Web interface denotes each orthogroup by a pair of grey horizontal lines spanning across the entire page.

4.3.3 Evaluation procedure

We have performed both a small-scale case study using short hedgehog gene query, and large-scale systematic gene-retrieval evaluations using 5-gene and 10-gene queries. We detail the evaluation procedure for the latter case. This evaluation requires us to first construct a combined whole-genome ranking or the "consensus" correlation ranking of genes to the query among two organisms. Specifically, the following procedure was used to construct combined retrieval ranking:

Let *S* and *T* be two organisms. Using *S*-*T* orthology mapping, we map genes *g* and *g'* from *S* and *T* respectively to orthogroups O_{ST}. Let $P_S(g, Q)$ be correlation *P*-value of gene *g* to *Q* in organism *S*, and similarly let $P_T(g', Q')$ be correlation *P*-value of gene *g'* to *Q'*, where *Q* and *Q'* are orthoqueries. An orthogroup $O_X \in O_{ST}$ consists of genes g_X and g'_X from *S* and *T*.

For each gene g_X in organism *S*, such that g_X is part of some orthogroup $O_X \in O_{ST}$, do

$P_T(O_X, Q')$= min($P_T(g'_X, Q')$ for each g'_X in O_X) (i.e. g'_X are members in O_X from organism *T*)

X^2 = -2.0 (ln($P_S(g_X, Q)$) + ln($P_{T,rescaled}(O_X, Q')$))

Combined $P(g_X, Q)$ = *P*-value of X^2 statistic with 2*k* degrees of freedom (*k*=2)

Done

Specifically, $P_T(O_X, Q')$ is the orthogroup O_X's correlation *P*-value to *Q'* from organism *T* side and $P_{T,rescaled}(O_X, Q')$ is its rescaled *P*-value. Here, rescaling the *P*-values corrects the $P_T(O_X, Q')$ according to extreme value distribution of uniform random variables. This adjustment is especially necessary as the min($P_T(g'_X, Q')$) in the 2nd line produce a non-uniform *P*-value distribution when null hypothesis is true. With the adjustment, the

distribution of rescaled *P*-values can be expected to be uniform under null hypothesis. Then the Fisher's combined probability method (last two lines) would be applicable. The Fisher's combined probability method was used to merge correlation *P*-values of two coexpression rankings into a X^2 statistic. The combined ranking is thus used for precision-recall evaluation to check if cross-organism retrieval improves performance at various recall levels.

4.4 Results

4.4.1 Illustration example

As an example of this comparative approach, we first look at a simple human query *GLI1*, *GLI2*, *PTCH1* – transcription factors and a receptor of the conserved hedgehog signaling pathway. The corresponding orthoquery in fly is *ci* and *ptc*. Upon ranking the orthogroups according to the correlation p-value to the query, ModSEEK has identified many orthogroups sharing coexpression with the orthoquery in human and fly. These include the critical developmental genes *SMO* (*smo* in fly), *NOTCH3/1/4* (*n* in fly), *SOX2* (*soxn* in fly), *FZD8* (*fz2* in fly), *TCF7L1/2* (*pan* in fly), all of which are ranked highly and participate in various orthogroup arrangements (**Fig. 4.3**). As hedgehog signaling is a developmental pathway, the retrieval of these genes appearsrelevant.

4.4.2 Evaluations

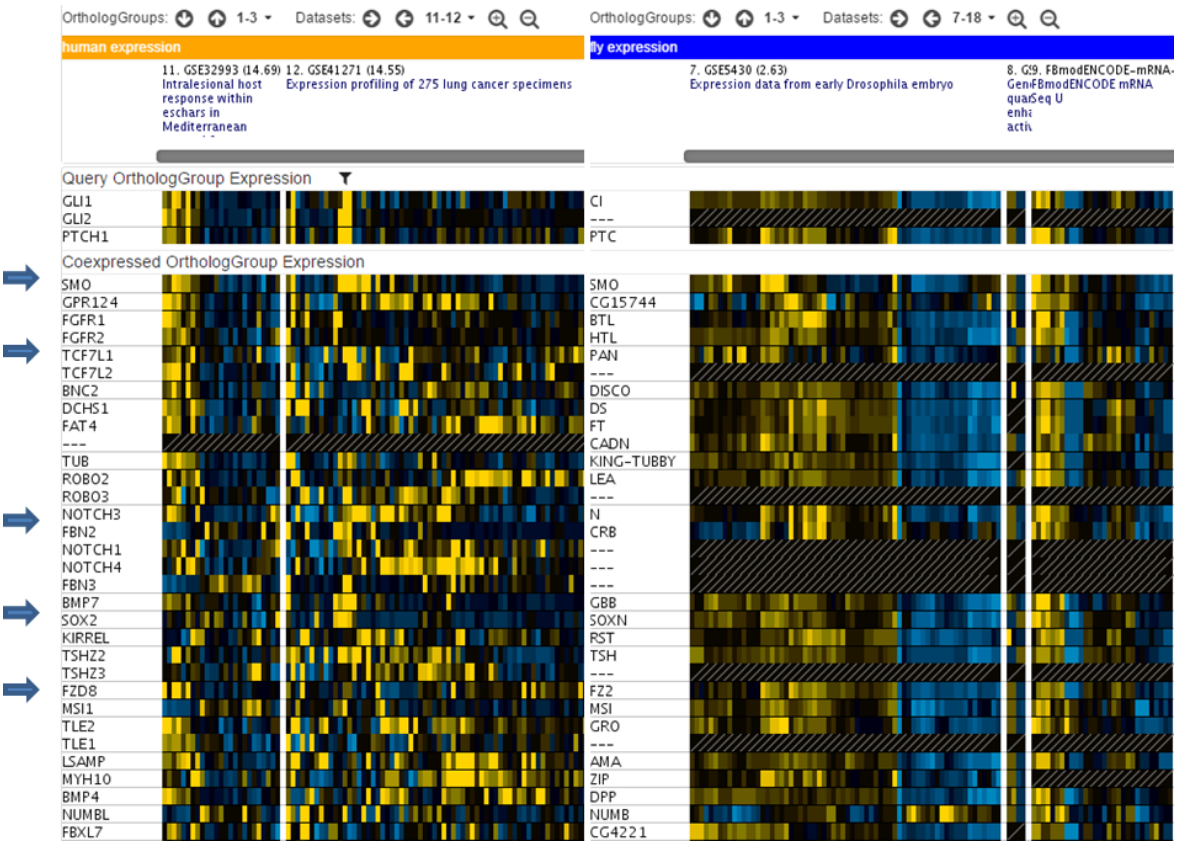


Figure 4.3 | Example hedgehog query in the actual search interface. Arrows denote key developmental genes retrieved to the hedgehog query.

ModSEEK not only works well for the hedgehog signaling pathway. In a large-scale systematic evaluation, we gathered KEGG[94] annotated pathway and metabolism gene-sets. We focused on these because signaling pathways and metabolic processes are two of the most conserved categories of biological process across diverse organisms. To evaluate the improvement brought about by cross-organism retrieval, we calculated precision-recall curve for single-organism unaided retrieval and for combined retrieval which utilizes both organisms' coexpression rankings. In combined retrieval, search rankings from respective organism's orthoquery are merged into a meta-ranking using Fisher's combined probability method (see Methods). Subsamples of 5-gene or 10-gene

are taken from each term as query to retrieve the remaining genes in the term. The results of searching these subsampled queries indicate that combined retrieval outperforms single-organism retrieval, by as much as 15% in human when aided by mouse, and 18% in human when aided by fly. Those combined retrieval cases leveraging model organism mouse or fly indeed perform better than human-only gene retrieval. This clearly demonstrates the potential for performance gain despite the fact that human data compendium is the largest of all organisms which already consists of thousands of datasets.

Fly and worm KEGG retrieval have each benefited significantly from using human orthoqueries as aid, which is a clear indication that a large data collection like human may improve gene function annotations of model organisms with much smaller data holding. In some cases, performance gain is not always bi-directional. For example mouse metabolism terms appear to drop slightly in performance when it is combined with human orthoqueries. One reason to account for this is that the mouse compendium may contain a greater experimental diversity of datasets than human, and those datasets encourage more accurate retrieval of diverse terms.

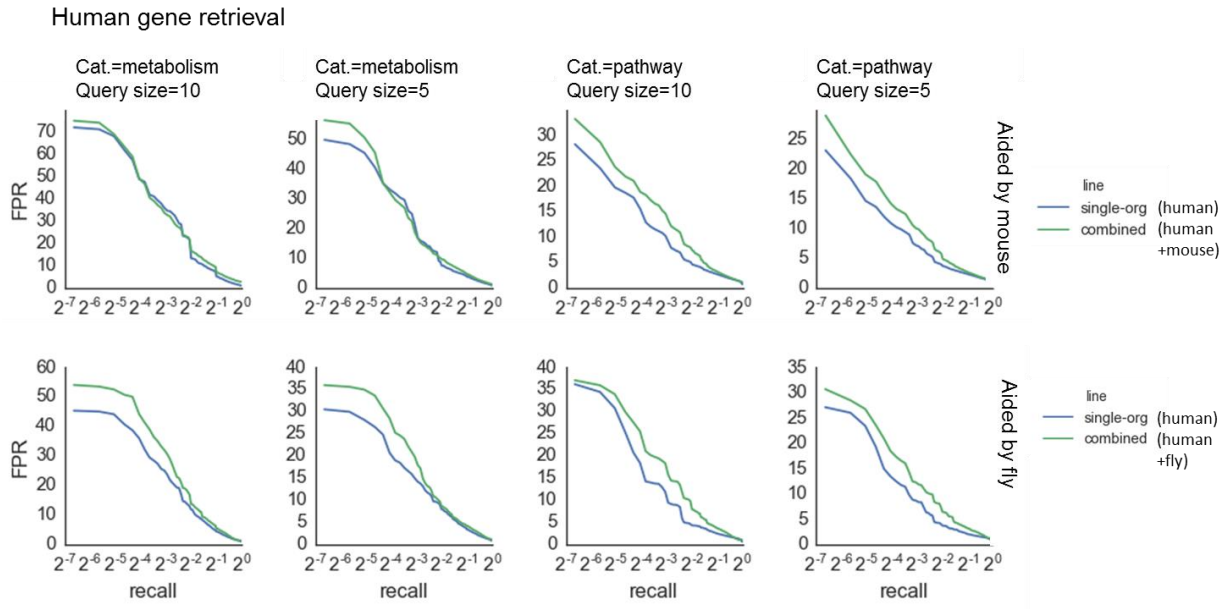


Figure 4.4 | Leveraging model organism orthoquery and search ranking improves the gene retrieval performance of human queries. Single-org: single organism human without any aid. Combined: combined human and mouse or human and fly orthogroup ranking. Evaluation gold standard used for the construction of PR curve was human in both cases.

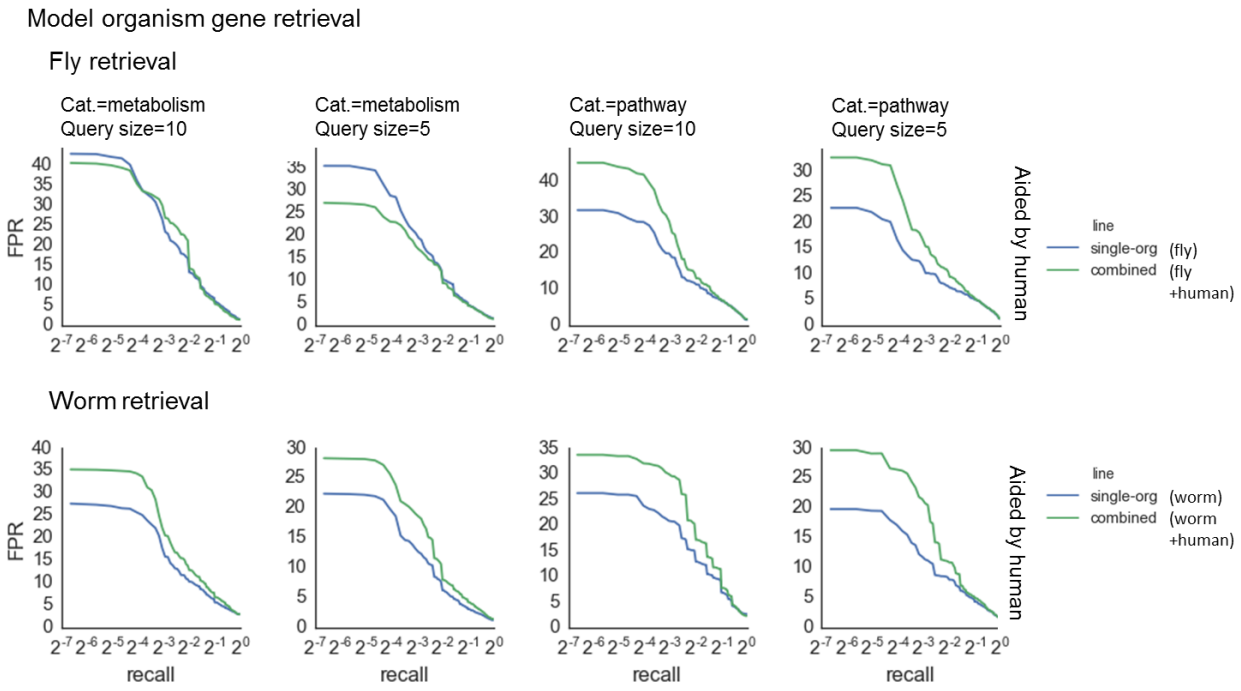


Figure 4.5 | Leveraging human orthoquery in the search process also improves the performance of model organism gene retrieval. Single-org: single organism without any aid. Combined: combined fly and human or worm and human. When human is employed as aid, evaluation gold standard uses fly and worm KEGG annotations respectively.

ModSEEK makes cross-organism expression search and functional analysis straightforward for the users. The comparative search interface is flexible and user-friendly and is supplemented with flexible options such as the ability to specify correlation *P*-value threshold for orthogroup selection, easy switch between multiple zoom modes in the expression viewer, and genome-wide coexpression score with *P*-value for custom filtering and analysis. Each search session generates a unique session ID that is stored and fully traceable by the user should there be a need to revisit the results. The

comparative search interface is available at http://seek.princeton.edu/modSeek/viewer_index.jsp.

4.5 Conclusion

Just as our universe of compendium information is growing, so are the datasets for a diverse set of organisms. The data compendium composition of each organism is not unbiased, as it is a unique mix of research topics that are of interest to model organism biologists, of experimental setups that are widely adopted, and of many other preferences. ModSEEK comparative search function leverages these unique differences to allow biologists to transfer experimental knowledge between organisms in a coexpression-based, unsupervised manner. ModSEEK is the first unsupervised system to our knowledge that combines comparative genomics orthology data, multi-organism functional genomics coexpression scores, and visualizations for truly user-driven, data-driven exploratory discoveries of gene function. We expect that by utilizing cross organism data ModSEEK should enable a more balanced, more accurate retrieval of query coregulated genes supported by conservation and functional genomics evidences.

5 IDENTIFICATION OF BREAST CANCER SUBTYPE-SPECIFIC REGULATORS AND TARGETS INFLUENCED BY GENETIC AND EPIGENETIC ALTERATIONS

5.1 Introduction

Tumorigenesis in breast cancer is thought to be the result of a combination of somatic genetic events including copy number aberrations (CNA), point mutations, and epigenetic alterations such as DNA methylation. In contrast to normal tissue development, somatic mutations can accumulate at various points of the differentiation process, making normal cells possess properties of stem cells that turn them into cancer cells. Despite the extensive generation of molecular data, how these mutations and aberrations specifically affect transcription factors and their targets is not well understood.

Breast cancer is a heterogeneous disease comprised of several molecular subtypes: luminal A, luminal B, Her2, basal, and normal-like, that are clinically important[6,95–98]. Recently a new and refined classification was proposed based on both mRNA expression and CNA[99]. Aberrations in breast cancer are manifested in a subtype-specific manner, and distinct biological processes are uniquely perturbed in these subtypes[100]. Previous efforts have identified mutational events at an unprecedented resolution and scale, linking those to the breast cancer subtypes[97,99]. The full picture of what TF networks need to be perturbed in order to lead to the development of different breast cancer subtypes, how

they evolve downstream of germ-line and somatic genetic events and what subtype-specific regulators are affected in order to give rise to the subtype-specific coexpression of genes is not yet understood. Here we address these questions, focusing specifically on basal and luminal breast cancer subtypes, which are the most distinct and best molecularly characterized subtypes, with distinct clinical outcomes and treatment regimens.

Transcription factor (TF) binding to proximal and/or distal regulatory elements is a critical mechanism regulating gene expression. Various TF regulatory networks have been constructed from ChIP-seq and ChIP-ChIP data, including a general TF network[101] based on a large number of cell types from the ENCODE project, and for breast – a nuclear receptor TF network focused on MCF7 cells[102]. Other studies focused on the identification of master regulators of differentially expressed and coexpressed genes[103,104]. All these studies are valuable in elucidating the regulatory structure of human cells, but suffer from a lack of specificity towards breast cancer subtypes. Computational approaches have focused on identifying candidate TFs through motifs overrepresented in the promoter regions of a set of breast cancer related marker genes, including works by us and others[105,106]. The downside of these *in silico* analyses was that transcriptional regulation was assumed to occur merely in promoter regions and distal enhancer elements were largely ignored. Yet it has become clear that important regulators in breast cancer such as *GATA3*, *ESR1* and *FOXA1* act through interacting regions occurring mostly at distal enhancer elements[107–110].

The expression and overexpression of genes in each subtype is mediated in a large part by the molecular processes that are turned on and the transcription factors that control the genes involved in these processes. The goal of this work is to identify breast cancer subtype-specific coexpressed genes and understand the transcriptional regulation that underlies this coordination. Our approach is to identify co-expressed genes in each breast cancer subtype and then find transcription factors that may regulate these co-expressed genes, as well as molecular lesions that disrupt this regulation (somatic copy number aberrations and altered methylation status).

Toward this goal, we integrate breast cancer subtype specific transcriptomic (expression) with cistromic (ChIP-seq) datasets to infer regulators and targets underlying breast cancer subtypes. We used over 130 breast cancer gene expression/epigenomics datasets from our compendium[111] and ENCODE[112] ChIP-seq data. Through integrated analysis of ENCODE data and transcriptional behavior of genes associated with each breast cancer subtype (identified by our context-relevant search algorithm based on breast cancer expression data[111]), we identified the most suitable cell line models from the ENCODE database: MCF7, T47D, previously well-known models for luminal A, and the identified here A549 and H1-hESC as novel suitable surrogates for the basal subtype. We further identified common regulatory regions in the coexpressed genes using epigenomic ChIP-seq experiments from these relevant cell lines. We show that these regulatory regions are enriched for motifs corresponding to TFs identified by SEEK as coexpressed with subtype specific seed genes, suggesting the regulatory role of the coexpressed TFs in a subtype specific manner. Overall, the integrated network that we construct from ChIP-seq TFs and motif-derived TFs (identified from motif analysis and subtype specific

coexpression), and refined gene target lists help us to better understand the regulatory events underlying the development of the breast cancer subtypes. Examination of aberration data revealed a tendency for coexpressed TFs in luminal A cancers to be subjected to DNA hypomethylation, whereas basal subtype TFs were often associated with somatic CNAs corresponding to regulation of epithelial-mesenchymal transition.

5.2 Methods

5.2.1 SEEK coexpressed gene search.

A short seed list of subtype-specific genes[113,114] (**Supplementary Data 5.1**) serve as input genes (or *query* in search terminology) for coexpression analysis. We have four input seed lists corresponding to the four breast cancer subtypes: luminal A, luminal B, basal-like, Her2-enriched. Each subtype gene list was queried in SEEK and the top returned coexpressed genes ($P<0.05$) were retained for subsequent analyses (**Supplementary Data 5.3**). Approximately 130 breast tumor RNA sequencing and microarray datasets were used for this coexpression analysis (**Supplementary Data 5.2**). The retrieved genes are termed coexpressed genes and are thus specific to each breast cancer subtype.

5.2.2 ChIP-seq data processing.

We used the collection of ChIP-seq datasets processed by the ENCODE Analysis Working Group (AWG)[115]. The AWG collection uses the SPP peak caller coupled with the irreproducibility discovery rate (IDR) procedure to discover significant and consistent peaks between two replicates of each ChIP-seq experiment within ENCODE[115]. We need to derive a gene-based score representing the amount of binding per gene. For this, we

first divide the peak signal score (the number of tags) by the 75-percentile peak score of the whole experiment[80], multiplied by base score 500, to adjust for sequencing depth. Next the peaks' chromosomal locations are aligned to all gene regions +/- 50kb TSS using BEDOPS[116] and human hg19 UCSC genome build. To account for the case when a peak falls within multiple genes' region, we calculate each gene score as sum of normalized contribution of peak scores: $g = \sum_{f \in P(g)} p_f / n_f$ where g is the gene score, f in $P(g)$ is the set of peaks in the vicinity of g, $p(f)$ is the peak score of f, $n(f)$ is the number of genes f overlaps.

5.2.3 Finding subtype-specific TFs from ENCODE data.

ChIP-seq derived TFs and motif-derived TFs.

We derived two categories of TFs relevant to cancer subtype-specific coexpressed genes. The first category is ChIP-seq-derived TFs. The relevant TFs are inferred by assessing if the list of ChIP'd target genes in a given ENCODE experiment is significantly similar to the list of cancer-subtype-specific co-expressed genes found by SEEK. To find the ChIP-seq derived TFs, we used the above procedure (Section 5.2.2) and obtained DNA-binding score per gene for each ENCODE ChIP-seq experiment C_i. To ask if a Chip-seq experiment C_i is relevant to subtype A, the ranked list of genes in C_i is compared to coexpressed genes of subtype A for significance testing using minimal hypergeometric overlap statistic (GORILLA[117]). GORILLA finds enrichment between a rank-list and the coexpressed genes without a need to specify depth of rank-list to compare with the gene-set. A relevant ChIP-seq experiment must satisfy P < 1e-5, and if so the ChIP'd TF becomes a relevant TF.

The second category of TFs is inferred from the presence of binding motifs in the regulatory region of cancer subtype-specific coexpressed genes ("motif-derived") with additional filtering applied. To find the motif-derived TFs in each subtype (**Supplementary Fig. 5.1**), we first get the regulatory sequences within 50kb of TSS for all coexpressed genes in a subtype. Regulatory sequences include transcription factor binding sites from ChIP-seq experiments and open-chromatin regions from DNase hypersensitivity experiments (**Supplementary Data 5.4**). We then performed motif-enrichment analysis on these sequences using *in vitro* motif database Weirauch et al[118]. Pscan-ChIP algorithm[119] was applied to find motifs enriched such that the probability of finding motif within the ChIP-seq region is greater than finding it within the flanking regions according to Welch's *t*-test (**Supplementary Fig. 5.1**). We further built on top of this algorithm since enriched motifs may not be regulating specifically coexpressed genes (but also other genes), and yet specificity is our interest, so we repeatedly draw random groups of genes of matched size with coexpressed genes, and performed the same motif discovery on 50kb regulatory regions of random genes (**Supplementary Fig. 5.1**). Qualified motifs must therefore be ranked in 95-percentile among the number of random trials. As a final stage of filtering, we checked whether TFs corresponding to these motifs are coexpressed with the subtype seed genes. (**Supplementary Fig. 5.1**). For example, if V$Gata family is an enriched motif, the TFs matching it (GATA1-6) are each checked for coexpression with luminal A subtype. GATA3 is finally selected as regulator due to it being the only member coexpressed.

5.2.4 Differentially enriched ChIP-seq TFs.

ChIP-seq derived TFs are categorized as differentially enriched depending on $\Delta(-\log(P_{val}))$ between the luminal A/B columns in Figure 5.1a and the basal column in Figure 5.1a. These TFs' ChIP-seq experiments were used to prioritize subset of coexpressed genes that are true transcriptional targets.

5.2.5 Breast cancer methylation, CNA aberration.

DNA methylation, copy number aberration (CNA) datasets were gathered from TCGA[97] sequencing and array-based sources available at ICGC[120]. These ICGC-provided data were supplied in processed forms. For our purpose, we excluded blood-derived normal samples, and metastatic samples from the TCGA list. Focusing on primary tumor samples, we derived tumor subtype-related dysregulation gene-sets using the procedure described below. We also constructed a separate normal breast tissue sample set from TCGA for comparison. To account for patient-to-patient variation within each subtype, we further require significance threshold to be satisfied to ensure that the level of dysregulation is consistent within subtype. To do this, we performed 1-sample or 2-sample *t*-tests (depending on the mechanism) between subtype and normal group (procedure described below).

Testing Association between TFs and DNA methylation or TFs and copy number aberrations. *Preparation of datasets*: Nucleotide-resolution methylation frequency (%) is binarized using 50% threshold, and aggregated to gene-level values using additive summary. We calculate subtype-specific DNA methylation frequency by averaging patients within subtype. To derive hyper- and hypo-methylation, for every gene we

calculate the difference between each cancer subtype's DNA methylation frequency and that of healthy normal group consisting of normal breast tissue samples. The resulting values per subtype are upper-quartile normalized[80] and subject to significance testing (2-sample *t*-tests with unequal variance, with respect to normal samples) to detect genes with substantial hyper- or hypo-methylation. For CNA, original data consists of copy number for various chromosomal segments detected per patient. Each copy number ranges from -1 to +1 (in log scale) which is relative to normal copy number. To derive gene-based copy number, let *cna* be the copy number of segment *f*; *len* be the number of genes contained in *f*; $F_G(g)$ be the set of gained fragments containing gene *g*; $F_L(g)$ be the set of lost fragments containing g. We calculate: $gain(g) = \sum_{f \in F_G(g)} cna(f)/len(f)$and $loss(g) = \sum_{f \in F_L(g)} cna(f)/len(f)$. As before, copy number values were aggregated among patients in the same subtype, followed by upper-quartile normalization, and are further subject to significance testing (1-sample *t*-test) to obtain gene-based CNA gain and CNA loss values for each cancer subtype. The gene-based copy number values derived this way are found reasonably close to the reported values found in the Metabric study[99].

5.2.6 Testing of association between TFs and dysregulation.

This uses GORILLA[117] which finds enrichment between a rank-list and a gene-set of interest to determine if the gene-set is enriched at the top of the rank-list. For our context, the rank-list is the whole-genome list of genes sorted by absolute dysregulation values. Test is independently performed for each mechanism of dysregulation. The gene-set of interest is the coexpressed genes of each cancer subtype. Tests that involve TFs (coexpressed and ChIP-seq derived TFs) used a random background of ~1000 TFs (in

human genome). Otherwise, tests involving coexpressed genes used the whole-genome background (~17000 genes). Multiple hypothesis testing procedure, by Benjamini-Hochberg[88], was applied to comparisons within each mode of dysregulation.

5.2.7 Dysregulation heatmap construction.

Values in the heatmap represent previously computed cancer subtype-specific upper-quartile normalized dysregulation values. Choosing normalized values is especially appropriate for this visualization purpose, as it permits cancer subtypes to be directly compared. A blue or red dot in the heatmap requires that *t*-test $P<0.01$ in 1-sample *t*-test (CNA) and $P<0.01$ in 2-sample *t*-test (DNAmeth) with respect to the normal group. If significant, the intensity of a dot in the heatmap is proportional to the effect size (dysregulation magnitude). Otherwise, a black dot (no signal) is displayed for the specific entry.

5.3 Results

In order to understand how transcriptional decisions influence the development of breast cancer cells, we reverse engineer the process, starting with coexpressed genes upregulated in specific breast cancer subtypes, which reflect distinct disease manifestation and then discover the basis for the coordinated transcriptional regulation. We used SEEK[111] to accurately identify coexpressed genes from a large compendium of over 130 breast cancer datasets publicly available in Gene Expression Omnibus[1]. First, a seed list of breast cancer subtype-specific genes was used to query SEEK and the top significant coexpressed genes were retained and used in subsequent analyses (**Fig. 5.1**). Next, two categories of transcription factors were inferred for each subtype: one inferred

by experimental ChIP-seq data from the ENCODE consortium (further referred in the text as ChIP-seq derived TFs) (**Fig. 5.1, #1**). These are TFs that are experimentally shown to bind to the regulatory region of a significantly large number of the expanded coexpressed genes as evident from the ChIP-seq data. The other category TFs, is the motif derived TFs (**Fig. 5.1 #2**, **Supplementary Fig. 5.1**), that satisfy the following: 1) they are themselves coexpressed with subtype-specific seed genes; 2) their binding motifs are significantly enriched among the regulatory regions of the coexpressed genes in a given cancer subtype.

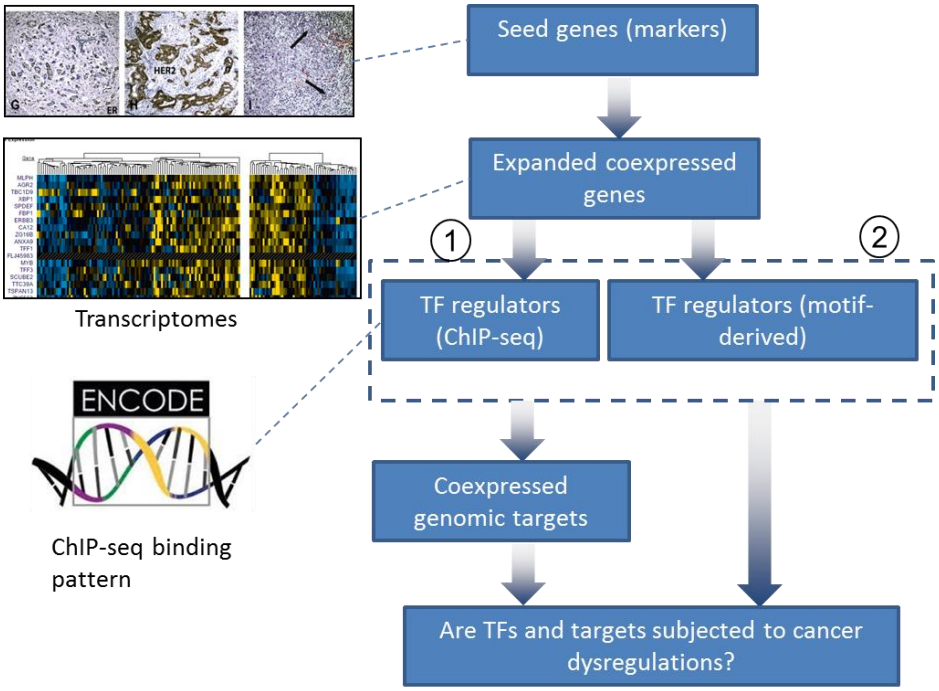


Figure 5.1 | Schematic of the workflow. As a first step, SEEK coexpression analysis enlarges the subtype specific seed genes to a larger signature gene set. Then, the next step is to find TFs that may regulate these coexpressed genes. Two sources of data, namely ENCODE ChIP-seq experiments (marked with #1) and motif-based analysis of cis-regulatory sequences of coexpressed genes (marked with #2), help reveal distinct regulators of subtype coexpressed genes. Afterward, to validate the TF regulators, we ask

whether they are more often than random subjected to breast cancer subtype specific dysregulations (copy number aberrations and DNA methylations).

5.3.1 Identification of TFs relevant to cancer subtypes

Identification of ChIP-seq-derived TFs and motif-derived coexpressed TFs

Our goal was to identify TFs that were known to bind to regulatory regions of coexpressed genes for each cancer subtype. To do this, we looked for ENCODE ChIP-seq experiments where the set of targets bound by the tested TF significantly overlapped with the set of coexpressed genes for a given cancer subtype. To find these ChIP-seq TFs, we first preprocessed ENCODE ChIP-seq datasets by mapping ChIP-seq fragments to genomic locations of individual genes based on a window size of 50kb+/- from the gene transcription start site. Then, read counts of these regulatory regions were summed to report a single value per gene. This assumes a local regulation model, whereby the regulatory regions close to the TSS are assumed to be important for the expression of all the genes nearby (50kb distance). The choice of 50kb allows for a range of informative distances to be established linking enhancer elements to the TSS region, and has been used previously[102,109]. Afterward, the coexpressed genes for each cancer subtype were compared to gene-level ChIP-seq read counts and the GORILLA hypergeometric test was applied to determine significant overlap (See Methods). In other words, the identified ChIP'd TF experiments must have significantly high TF-binding reads at the coexpressed genes compared to non-coexpressed genes. When significant, we inferred that the TF in the ChIP-seq experiment may regulate the coexpressed genes for that cancer subtype.

Because ChIP-seq TFs are biased towards the knowledge of what TFs scientists have chosen as interesting to antibody and profile, the scope of TFs from ENCODE is notably limited. To complement ChIP-seq data and to expand the scope of inferable TFs, we further studied the coexpressed TFs which were found in the SEEK output as themselves coexpressed with the subtype marker genes. An additional requirement is that there must be a motif enrichment for the respective coexpressed TF in the regulatory region of the subtype coexpressed genes. This precludes the possibility that coexpressed TFs are merely targets rather than drivers of coexpression transcriptional programs.

5.3.2 Subtype-specificity of ChIP-seq TFs

Luminal A TFs

We expected that the lists of inferred cancer subtype-specific ChIP-seq TFs would include known subtype-specific TFs. We indeed found three well-known luminal A-associated TFs on the list of inferred luminal A TFs: Esr1, Gata3, and Foxa1 (**Fig. 5.2**). Furthermore, the experiments identifying these TFs were done in MCF7 and T47D cell lines, well-accepted models of luminal A biology, indicating that the set of TF targets are subtype-specific as well (**Fig. 5.2**). Accordingly, basal-like coexpressed genes were not significantly enriched among the targets of the luminal A ChIP-seq TFs (difference in $-\log_2(P \text{ value})$ is 7, 9, and 16 respectively for Gata3, Esr1, and Foxa1 (**Fig. 5.2**). Other ChIP-seq TFs with strong target enrichment with luminal A genes, but not with basal-like coexpressed genes, were Znf217, Nr2f2, Myc, Foxm1, Max, and Tead4, all from MCF7 cell lines (**Fig. 5.2**). Altogether, over 50 of the 59 top Chip-seq experiments with $-\log(P \text{ value})$ enrichment greater than 5 come from MCF7, corroborating with the fact that MCF7 is the most appropriate model for the physiology of luminal A subtype.

Basal TFs

The specific set of ChIP-seq TFs most actively binding basal-like genes include Cfos, Stat3, Myc ($-\log_2(P)$ ranging from 20 to 40), Gr, Fosl2, Tcf12, Atf3 (of A549, $-\log_2(P)$ is 15 to 19), and Tead4, Chd1, Jund, Rbbp5, Ctbp2, and several others of H1hesc ($-\log(P) > 10$) (**Fig. 5.3**). The above suggested basal regulators are subtype specific given that they show weak ($-\log(P)<5$) or no enrichment for targeting luminal A coexpressed genes. As we observed above, we were particularly intrigued by the connection of A549 and H1hesc cell lines to basal as no other cell lines were previously suggested as good models of basal-like subtype than the conventional MCF10aes. Yet, our unbiased analysis shows that the epithelial and messenchymal stem cell lines A549 and H1hesc can serve as useful models, matching the known epithelial and mesenchymal stem cell like characteristics of the basal subtype[121]. Overall, overlapping analysis between coexpressed genes and genes ranked by ChIP-seq reads successfully prioritized relevant cell line models, and ChIP'd TFs to be candidate regulators of basal and luminal A coexpressed genes.

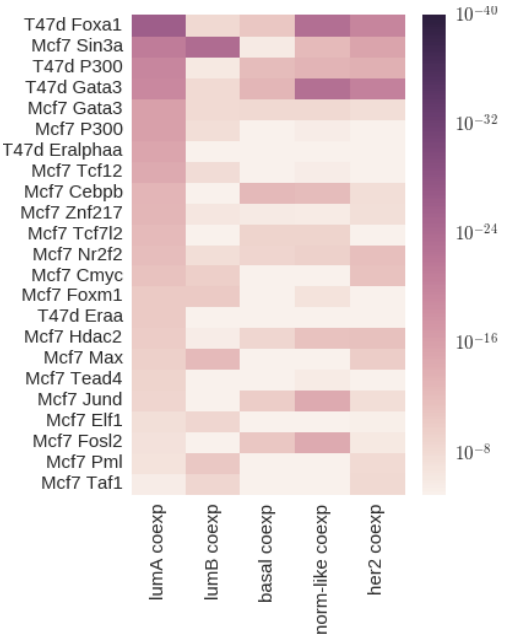


Figure 5.2 | Top ChIP-seq experiments ranked highest in terms of luminal A coexpressed genes. Each entry shows the –log(*P*) of gene-set overlap between coexpressed genes of a subtype and target genes of a ChIP-seq experiment. A significant *P*-value indicates the ChIP'd TF binds regulatory sequences of a large number of coexpressed genes, and is thus a candidate regulator of subtype. The right bar is *P*-value legend.

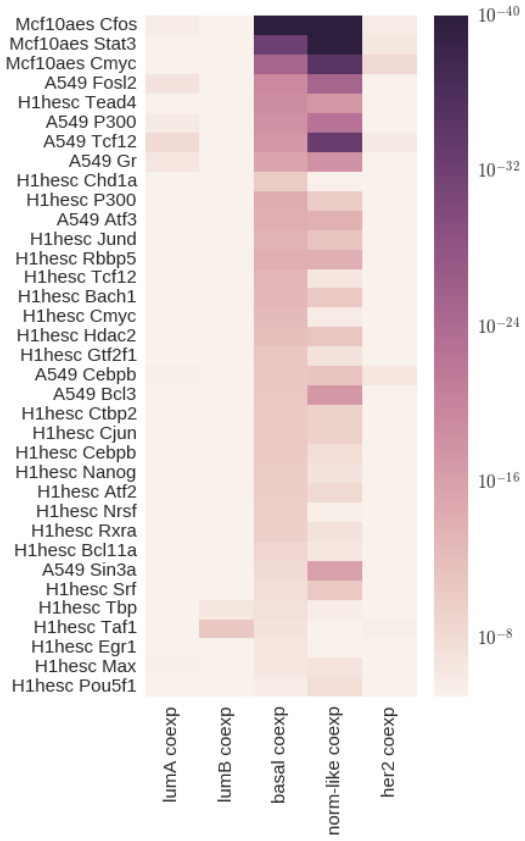


Figure 5.3 | Top ChIP-seq experiments ranked highest for basal-like coexpressed genes. This is a similar graph as Fig. 5.2, except this shows the basal-like ChIP'd TFs.

Table 5.1 | Luminal A ChIP-seq derived TFs

ChIP-seq derived TFs	
GATA3	10p15
TCF7L2	10q25
FOXM1	12p13
TEAD4	12p13
ELF1	13q13
FOXA1	14q12
MAX	14q23
TCF12	15q21

PML	15q22
SIN3A	15q22
NR2F2	15q26
JUND	19p13
CEBPB	20q13
ZNF217	20q13
FOSL2	2p23
HDAC2	6q21
ESR1	6q24
MYC	8q24

Table 5.2 | Basal-like ChIP-seq derived TFs

ChIP-seq derived TFs	
CTBP2	10q26
NANOG	12p13
TEAD4	12p13
MAX	14q23
FOS	14q24
TCF12	15q21
STAT3	17q21
GTF2F1	19p13
JUND	19p13
BCL3	19q13
USF1	1q22
ATF3	1q32
RBBP5	1q32
CEBPB	20q13
BACH1	21q22
BCL11A	2p16
FOSL2	2p23
ATF2	2q32
REST	4q12
CHD1	5q15
EGR1	5q23
NR3C1	5q31
TAF7	5q31
SRF	6p
POU5F1	6p21
HDAC2	6q21
TBP	6q27
MYC	8q24
RXRA	9q34
TAF1	xq13

5.3.3 Coexpressed targets of ChIP-seq TFs: literature-based validation

Luminal A targets identify ChIA-PET supported ER-regulated regions and basal targets are partitioned into epithelial and stem cell phenotypes, further confirming EMT in basal subtype

We next identified the subset of coexpressed genes most frequently targeted by the inferred regulators (ChIP-seq targets). Using ChIP-seq TFs that are differentially enriched between the luminal and basal subtypes, we prioritize and identify subset of coexpressed genes that are likely targets. We identified over 141 luminal A targets and 137 basal targets. Luminal A targets, expectedly, include *FOXA1*, *ESR1*, as well as *EVL*, *PREX1*, *KCTD3*, *VAV3*, *MYOF*, *PKIB*, *PBX1*, *SIAH2* and others. The self- and co-regulatory functions of ESR1, *FOXA1* and *GATA3* have been well described[122]. The luminal A target list includes *KRT8*, *SIAH2*, *TFF1* which each contain ER-alpha binding sites in distal regulatory regions. They were experimentally verified by ChIA-PET experiments[109,123] to form chromatin loops to activate gene expression in MCF7 cells. *SIAH2* is part of a single-gene chromatin loop, while *KRT8* is the anchor gene part of a larger keratin interaction loci[109]. Thus, some of our targets are known to activate gene expression through mediating chromatin interactions in a Luminal A-relevant cell line. Because target genes are required to be coexpressed (and upregulated) in luminal A, our list of coexpressed targets can potentially inform those chromatin interactions that likely result in active or overactive expression.

In the basal subtype, frequent targets were *NFIB* (nuclear factor 1B), and cell surface genes *EDN1* (endothelin 1), *SVIL*, *FAT1*, *ANXA1* (annexin A1), *KANK1* (KN motif and ankyrin repeat domains 1), *EGFR*, *RND3* (Rho family GTPase 3), *NCOA7*

(nuclear receptor coactivator 7), *ROR1*, *VGLL4* (vestigial like family member 4), of which several are TFs or co-factors themselves. Specifically, this list of basal targets was derived from experiments in two cell types: A549 (epithelial phenotype) and H1hesc (mesenchymal phenotype). Amongst the basal target genes we found markers of both epithelial and mesenchymal phenotype: 1) *ERRFI1*, *ANXA1*, *EDN1*, *MID1* (using epithelial A549 cell line); 2) *BCL11A*, *LPHN2*, *ROR1*, *ZNF532*, *GCNT2*, *PODXL*, *SPRY2*, *EPHB3* (using stem cell H1hesc) (**Supplementary Fig. 5.2**). A hallmark of the basal breast cancer subtype is the expression of EMT genes.

5.3.4 Validation of coexpressed targets in siRNA and knockdown experiments

To further validate targets, we looked for public gene expression datasets that investigate the effect of TF knockout or knockdown in relevant breast cancer cell lines. If the TF has been perturbed, then we expect gene expression changes should be greater for the target genes than for non-target genes (the random control in this analysis). For this purpose, we obtained TF-perturbation experiments for four TFs from our basal and luminal A subtype lists: *BCL11A* (basal), *CTBP2* (basal), and *REST* (basal), *FOXA1* (luminal A). Fold change expression between knockdown and a control condition without knockdown was measured. For each of the luminal A and basal subtypes coexpressed genes we found that coexpressed targets have greater absolute fold changes between the two conditions than a random control set of genes (**Fig. 5.4**). A slightly greater proportion of targets are enriched with high absolute *FCH*, where $FCH > mean + std$, if the target genes are TFs than non-TF targets. These results further reaffirm the importance of our inferred TFs and target lists.

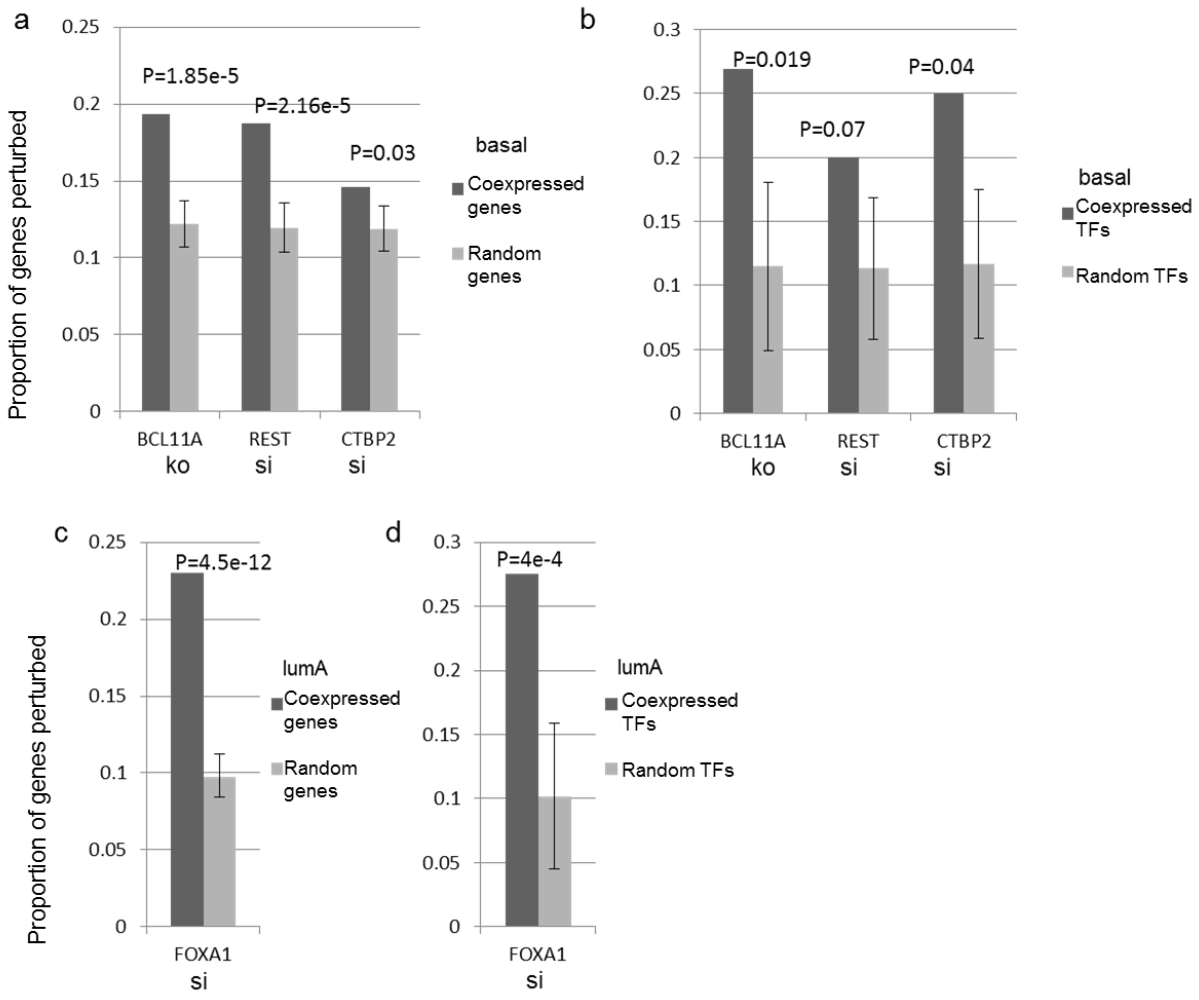


Figure 5.4 | Proportion of coexpressed genomic targets and TFs having substantial FCH after TF knockout or siRNA knockdown. Substantial fold change (FCH) is defined as abs(FCH) > mean + std, where FCH is fold-change relative to a control condition without perturbation. (**a**–**b**) Basal coexpressed genes and TFs are evaluated for FCH for TFs inferred to be important in basal-subtype: BCL11A (knock out), REST (siRNA), CTBP2 (siRNA).(**c**–**d**) Luminal A coexpressed genes and TFs are tested for FOXA1 perturbation (siRNA). Experimental data are obtained from public sources (GSE25315, GSE36529, GSE63389, and GSE63608).

5.3.5 Further expanding the subtype-relevant TFs: motif-derived TFs from coexpressed genes.

The TFs inferred from ChIP-seq experiments were chosen based on the similarity between the target gene regulatory regions they bind and the subtype-specific coexpressed genes. The TFs themselves may not be coexpressed in the cancer with the subtype-specific genes we have identified with SEEK. We observed that there are TFs in the list of co-expressed genes and decided to test whether these coexpressed TFs could in fact regulate the coexpressed genes also. Examples supporting this case includes ESR1, GATA3, and FOXA1 which all are both coexpressed to, and regulate luminal A's coexpressed genes.

Motif-derived TFs from coexpressed groups identify luminal steroid hormone regulators and basal regulators

To confirm that each subtype's coexpressed TFs could regulate the coexpressed genes, we used motif-analysis (**Supplementary Fig 5.1**). We identified 16 luminal A motif-derived coexpressed TF regulators (2 of which are also nonspecifically targeting non-subtype genes, i.e., ubiquitous), and 23 basal regulators (4 of which are nonspecific) (**Tables 5.3–5.4** motif-derived column). Among the list of motif-verified coexpressed luminal A TFs we found Xbp1 (X-box binding protein), Pgr (progesterone receptor), Ar (androgen receptor), in addition to Foxa1, Esr1, Gata3 for which ChIP-seq data were available. Among the basal-like motif-derived regulators (**Table 5.4**) we found Bcl11a, Id4, En1, Sox9, for which literature evidence supports their roles in triple-negative breast cancer or part of the stem cell differentiation programs[124,125]. The highly relevant basal Foxq1 is a driver of the TGF-beta signaling pathway, participates in crosstalk with Wnt

signaling pathway, and influences EMT[126]. The results confirm the presence of motifs for steroid hormone receptors, including Esr, Pgr, and Ar within 50kb open chromatin region of luminal A coexpressed genes.

Table 5.3 | Luminal A motif-derived TFs

Motif-derived TFs	
GATA3	10p15
PGR	11q22
TBX3	12q24
FOXA1	14q12
ZBTB42	14q32
IRX5	16q11
SREBF1	17p11
CREB3L4	1q21
PBX1	1q23
XBP1	22q12
FOXP1	3p14
SPDEF	6p21
ESR1	6q24
AR	xq12

Table 5.4 | Basal motif-derived TFs

Motif derived TFs	
ELF5	11p13
ETV6	12p13
SOX8	16p13
SOX9	17q23
RUNX3	1p36
CEBPB	20q13
ETS2	21q22
MAFF	22q12
SOX10	22q13
BCL11A	2p16
EN1	2q13
ETV5	3q28
ID4	6p22
FOXC1	6p25
FOXQ1	6p25
CREB3L2	7q34
NFIB	9p24
NFIL3	9q22

5.3.6 Associations of TFs with dysregulations

Groups of ChIP-seq and coexpressed TFs are distinctly associated with subtype-specific breast cancer aberrations including DNA methylation, copy number aberrations

Development of cancer involves dysregulation at multiple levels, including changes to the DNA, such as somatic copy number aberrations (CNA) and changes in promoter methylation. We therefore assessed whether the sets of cancer subtype specific TFs described here, were more often than expected by chance subjected to genetic (CNA) and/or epigenetic (DNAmeth) aberrations, i.e. significantly more often than random sets of TFs in breast cancer subtypes. To address this, we used data from 700 breast tumors with subtype-specific DNA methylation (DNAmeth), and copy number aberrations (CNA) data at TCGA. We summarized aberrations on a gene-level to facilitate comparisons with the coexpressed genes, and tested the relevance of each type of subtype-specific aberration (CNA, DNAmeth) on the subtype-specific regulators. As we are concerned with whether or not they are subject to dysregulations, absolute CNA or DNAmeth frequencies were measured.

Table 5.5 | Associations (Copy Number Aberrations and DNA Methylations).

		Gene sets tested		
		LumA ChIP-seq TFs (*Q*-val)	**LumA motif-derived TFs (*Q*-val)**	**LumA Targets* (*Q*-val)**
TCGA lumA population	CNA	3.07E-01	5.79E-02	2.34E-02
Curtis *et al* lumA population	CNA	2.85E-01	3.18E-03	8.51E-02
TCGA lumA population	DNAmeth	4.64E-01	4.02E-02	1.88E-04
Fleischer *et al* lumA population	DNAmeth	8.55E-02	1.19E-03	1.07E-04

	Gene sets tested

		Basal ChIP-seq TFs (*Q*-val)	Basal motif-derived TFs (*Q*-val)	Basal Targets* (*Q*-val)
TCGA basal population	CNA	2.21E-03	1.41E-01	1.38E-03
Curtis *et al* basal population	CNA	2.07E-02	1.40E-03	1.81E-04
TCGA basal population	DNAmeth	9.10E-01	1.32E-01	4.58E-01
Fleischer *et al* basal population	DNAmeth	8.17E-01	5.43E-02	9.40E-02

*ChIP-seq derived targets within coexpressed group

Our results (**Table 5.5**) indicate that among luminal A population, luminal A coexpressed motif-derived TFs as a whole group are significantly dysregulated by CNA (Q<0.06 in TCGA and Q<0.00318 in Curtis et al[99]) and by DNAmeth (Q<0.04 in TCGA and Q<0.00119 in Fleischer et al[127]) (note that Curtis *et al* and Fleischer *et al* are external cohorts provided in addition to TCGA). In contrast, no specific associations could be made between the luminal A ChIP-seq TFs and CNA or DNAmeth in luminal A population (Q<0.4 – 0.6). In basal population, basal ChIP-seq TFs are dysregulated by CNA (Q<0.00462 in TCGA and Q<0.019 in Curtis et al), but not by DNAmeth (Q<0.87 and Q<0.79). The results reflect distinct preference towards a specific type of dysregulations depending on subtype, and type of TFs examined. Note that because motif-derived TFs are coexpressed with the subtype seed genes, they are found more relevant and more generally informative than ChIP-seq TFs in carrying subtype dysregulations. They are not reliant on the availability of ChIP-seq experiments. It suggests that disruption of motif-derived TFs may be highly relevant to tumor development.

Individual TFs in the ChIP-seq and coexpressed sets are illustrated to have unique clustering pattern based on patterns of dysregulations across the four tumor types (**Figs. 5.5-5.6**). Distinct CNA targeting of basal ChIP-seq TFs in basal subpopulation is noted,

but not noted in other subtypes and are not noted in coexpressed groups, suggesting specificity.

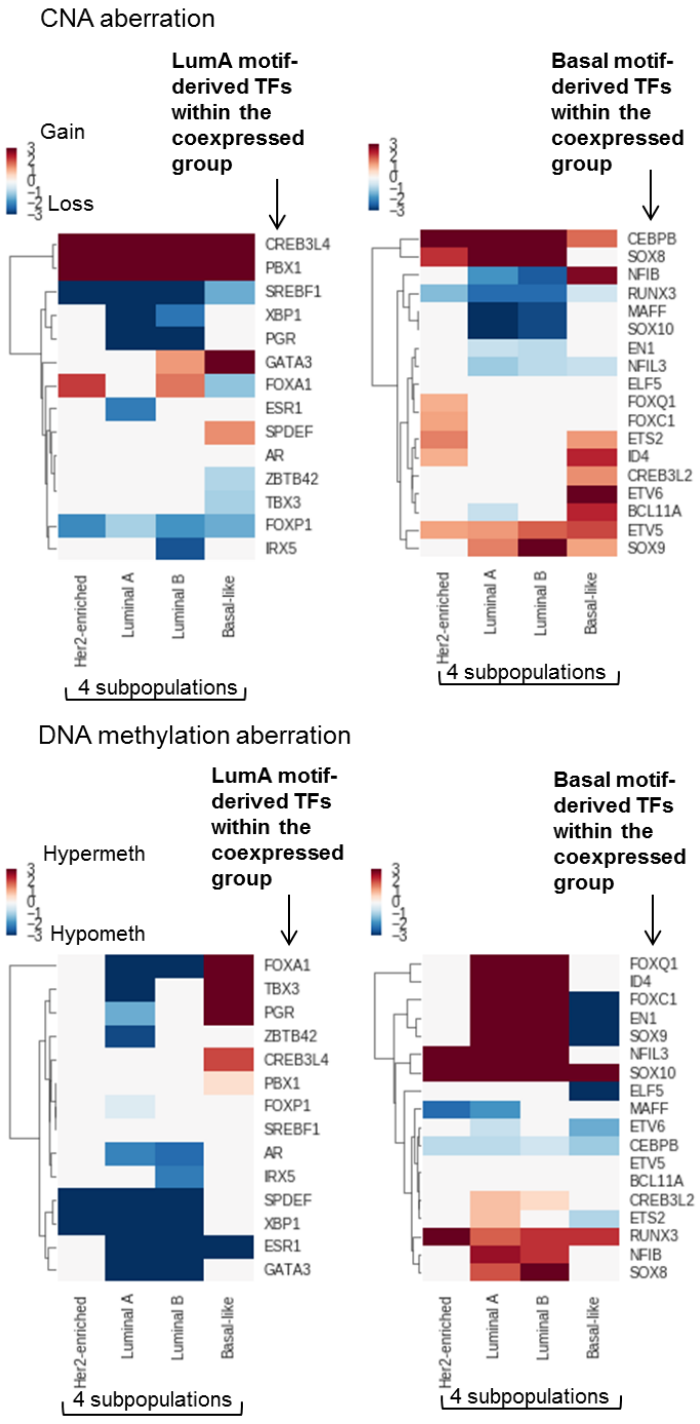


Figure 5.5 | CNA and DNAmeth maps on motif-derived TFs within the coexpressed groups. TCGA breast cancer cohort. DNAmeth, both hyper and hypomethylation is

widely prevalent across luminal A/B subpopulations at the luminal A specific TFs (bottom). This is consistent with the reported association in Table 5.5 top. A significant number of basal motif-derived TFs have CNA in the basal subpopulation. Differential pattern is obvious between the basal and luminal A subtypes.

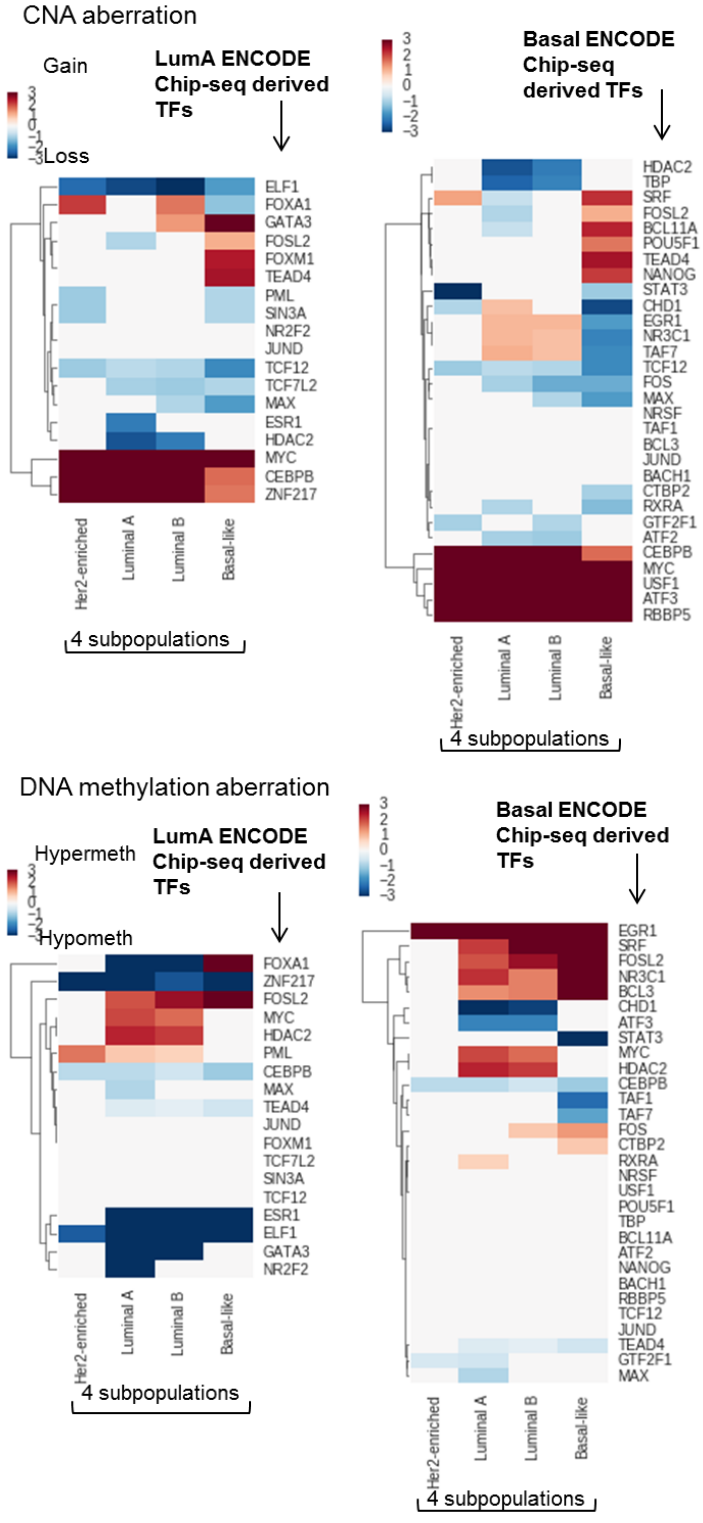


Figure 5.6 | CNA and DNAmeth maps on ENCODE ChIP-seq derived TFs. TCGA breast cancer cohort. A number of TFs (MYC, CEBPB, USF1, ATF3, RBBP5) are universally targeted across all subtypes. However, a large number of basal ENCODE Chip-seq derived TFs (top right) are distinctly associated with CNA gain (e.g. BCL11A) and CNA loss (e.g. FOS, STAT3), supporting the overall association reported in Table 5.5 bottom. ENCODE Chip-seq TFs are devoid of DNAmeth (bottom right).

Patterns of DNAmeth dysregulation at the coexpressed TF level are particularly subtype specific (**Fig. 5.5** bottom). For example, we located a subset of TFs containing stem cell differentiation factors *SOX9*, *EN1*, *GRHL1*, *FOXC1*, *ETS2*, *ETV6* which are hypomethylated in basal subpopulation and hypermethylated in luminal A subpopulation. Such factors in hyper/hypomethylation states possibly suggest that stem-like properties are effectively suppressed in the non-basal subtypes through DNA methylation. On the other hand, luminal A coexpressed TFs are characterized by having hypomethylations in luminal A/B subpopulations (evidence of hypomethylation marks is noted at *GATA3*, *BHLHE40*, *ZBTB42*, *SPDEF*, *TOX3* in luminal A patients and not in basal) (**Fig. 5.5**). Thus, differential DNA methylation at TFs plays a critical role in maintaining luminal progenitor states and initiating cancer stem cell states in the basal subtype.

5.4 Discussion

Understanding transcriptional regulatory processes in breast cancer subtypes is a prerequisite step to understand subtype specific susceptibility and to develop therapy strategies that target individual subtypes. Our work enables us to gain insights into the molecular factors of each individual subtype using data-driven analysis. We combined

existing cistromic and a large compendium of transcriptomic datasetsin an integrative framework to better understand the roles of transcription factors behind tumorigenesis.

Previously, transcriptional regulators and targets have been inferred using differentially expressed genes of interest, with cistromic ChIP-seq data, and with promoter information. However, inferring relevant TFs by using cistromic data is not sufficient because the scope of ChIP'd TFs in ENCODE consortium is rather limited to general and well known TFs such as EP300, HDAC2, etc. Other TFs which have not been ChIP'd and play critical roles in the development of cell types and tissue types have been missed. So far for breast cancer, only a few tissue-specific TFs, such as ESR1, FOXA1, GATA3 have corresponding ChIP-seq experiments generated by ENCODE. In this work, we used the large-scale integrative system SEEK coupled with motif analysis, to discover perhaps an understudied class of subtype-specific TF regulators, which are coexpressed with the subtype biomarkers, and have their motifs enriched among the subtype's coexpressed genes. We illustrate the value of such approach in analyzing luminal A and basal subtypes of breast cancer.

By analyzing the two categories of TFs, i.e. ChIP-seq TFs from the ENCODE project, mostly noncoexpressed, and motif-derived TFs within the coexpressed groups, we are able to reveal the rewiring of transcription networks at different levels. For example, we found in this work that DNA hypomethylation tend to be associated with coexpressed TFs upregulated in the luminal A population, but does not involve a cell state change. In basal population, however, we found that CNA affects more often (than in other subtypes) general set of ENCODE ChIP-seq TFs that regulate stem cell and

epithelial phenotype switch, suggesting that basal cell state is much more dynamic and alterations occur at the more general ubiquitous TF level. The important contribution of CNA to basal like breast cancer tumors has also been reported previously, where basal CNA identify genes involved in genomic instability[128].

Forces of epigenetic and genetic alterations have severe impacts on transcriptional networks in breast cancer. It is for this reason that we studied CNA and DNAmeth on transcription factors in subtype specific breast cancer. Not only do we observe MYC is implicated in perturbed transcription factor network as revealed by ChIP-seq experiments, MYC also has CNA gain in subtypes of breast cancer. It is likely that more TFs which regulate the expression of the coexpressed genes, have been also subjected to CNA and DNAmeth dysregulations. Our study investigates the tight inter-play between three elements of regulation (CNA, DNAmeth, TF binding) for the first time in breast cancer, and will contribute to a better understanding of how subtype-specific breast cancer arises.

6 CONCLUSIONS AND FUTURE WORK

In this thesis, I have designed search systems to address the critical need for unsupervised targeted analysis of the massive gene expression data collections. These systems have wide ranging applications, such as gene function prediction, prioritization of datasets, and inference of genes for further motif-based analysis and master regulator inference.

Specifically, I first developed the human-only search system, SEEK, that supports the targeted integration of over 5000 gene expression datasets covering 48 different platforms. The number of datasets integrated is the largest to date. I showed that this system is general enough to support the investigation of diverse areas of biology using large-scale gene function prediction evaluations from 995 gene ontology biological processes. A key reason for SEEK's accuracy is the query-based weighting of datasets, that can automatically detect relevant datasets from the compendium for retrieving coexpressed genes. This novel rank-based, cross-validation-based weighting algorithm shows great discriminatory power for the most query-relevant datasets. As the dataset weighting specifically exploits multigene queries, multigene queries have allowed highly expressive query context to be constructed, thus enabling accurate search results.

In the next chapter, I have extended the functionality from human-only to 5 other commonly studied model organisms: mouse, fly, worm, yeast, and zebrafish. Systematic evaluations have shown that the SEEK algorithm work equally well to retrieve model organism-specific biological process genes, given member genes as query. I developed a fast coexpression testing procedure based on generalized pareto distribution modeling of coexpression score. The computed coexpression score summarizes the full and partial

coexpression relationships between the query genes, and is robustly tested against randomly selected genes of matched size. This procedure has been adopted for large-scale dataset prioritization given a query of interest.

In the following chapter, I have developed the ModSEEK comparative search system, a system that permits researchers to combine orthology data with coexpression data to identify orthogroups with functionally co-similar orthologs relative to a query reference. The system has been evaluated systematically in gene retrieval studies, where I showed that cross-organism retrieval leveraging the transfer of annotations along functionally similar orthologs enabled better retrieval performance compared to unaided single-organism-based gene retrieval.

Finally, I demonstrate an important use of SEEK in a breast-cancer focused case study. The coexpressed genes retrieved by SEEK have become subtype-specific gene signatures, an important starting point for ensuing investigations. Specifically, using coexpressed genes I was able to trace subtype-specific regulators and motifs governing individual subtypes' regulation. Motif analysis and ChIP-seq data were integrated to bring about a full picture of subtype-specific transcriptional landscape. An important class of TF regulators that are found coexpressed to subtype seed genes, has been uncovered and has been shown to carry subtype-specificity not only in expression but also in cancer dysregulations such as copy number aberrations and DNA methylations.

The SEEK search algorithm can be applied to more organisms. The 5 model organisms shown in this thesis provide promising examples that it would likely work for other organisms with large data compendia. The multigene query ability has opened up

unexpected opportunities for new analysis. For example, large multigene queries constructed from differentially expressed gene sets can enable researchers to identify datasets with coordinated up- or down-regulation in the query genes. This analysis would be based on SEEK's query-based dataset weighting score. SEEK provides an appealing alternative and a data-driven extension to services such as MsigDB.

In future, there will be an increasing need to integrate even larger amounts of expression data, so scalable algorithms that can handle several times the existing data size should be developed to anticipate future data growth. As well, methods that can intelligently handle compendium update and mechanism of keeping track data provenance will be an important area of research to support the goals of open and reproducible workflow. The era of big data is full of exciting opportunities and challenges. Rather than being fearful of big data, I believe that we should embrace it for what it is capable of achieving, and its transformative potential to enable novel insights and shift research paradigm.

SUPPLEMENTARY NOTES

A.1 Hedgehog (Hh) query – detailed analysis of the retrieved genes

Below we describe additional details of the top retrieved genes for the Hh pathway example described in the manuscript. The known Hh pathway members *SMO* (rank 1), *HHIP* (rank 6), *BOC* (rank 7), and *PTCH2* (rank 9) are all among the top 10 SEEK-retrieved genes, and *KIF7* is ranked 22 – all in the first view immediately available to the biologist running SEEK. Other Hh-associated genes are also retrieved with top ranks. Multiple studies show that the TGF-beta pathway genes *RGMA* (rank 2), *LTBP4* (rank 8) are significantly co-induced with *GLI1* and *GLI2* in recurrent tumors[129,130]. The ortholog of protocadherin 18 (*PCDH18*, rank 3) interacts with *DAB1*, which functions in concert with the Hh pathway to control retina development[131]. *FZD7* (rank 4) is an important receptor in the Wnt pathway that extensively cross-talks with the Hh pathway [132]. The Notch signaling protein *HEYL* (rank 15) regulates *HES1*, which directly modulates *Gli1* expression and Hh signaling[133,134]. *HHIP-AS1* (rank 20) encodes the antisense RNA of the Hh interacting protein *HHIP*, which is a vertebrate-specific inhibitor of Hh signaling[135]. Many others genes among the top 25 retrieved – *KIF26A* (rank 10), *CRMP1* (rank 11), *CCDC8* (rank 13), *SLC26A10* (rank 14), *RUNX1T1* (rank 17), *MRAP2* (rank 18), *GPR124* (rank 19), and *PCYT1B* (rank 21) – have literature evidence for either regulatory interactions (direct or indirect), or pathway-level cross-talk with members of the Hh signaling pathway.

A.2 Web interface details

SEEK has been implemented as an interactive, easy-to-use website that allows biologists to perform queries, view expression patterns of the retrieved coexpressed genes, and perform visualization-based analyses. The goal of the SEEK web interface is to offer a Google-like engine for expression and coexpression retrieval, enabling biomedical researchers to fully utilize the thousands of expression data sets for accomplishing their analyses with a focused yet flexible and interactive web-based system. The web interface offers three flexible modes of visualizing users' results: **expression view**, **coexpression view**, and **condition-specific view**.

Expression view is the first view that the user sees upon completion of their search. **Fig. 6a** (main text) shows an example. The top 100 coexpressed genes are shown for the query *GLI1*, *GLI2*, and *PTCH1* (the user can easily see other lower-ranked genes of interest). The data sets are displayed in order of relevance, allowing the user to focus on those most related to their area of interest based on query coexpressions. In this view, expression levels for each gene are displayed, and a score is provided for each gene that conveys its level of normalized, hubbiness-corrected, and weighted coexpression to the query. A weight is provided for each data set, which offers a measure of the coexpression between the query genes in that data set as an indication of data set relevance. Each page juxtaposes multiple data sets' expression matrices to allow quick comparison and navigation. Within each data set's expression matrix, SEEK hierarchically clusters the conditions in the data set according to expression of the retrieved genes that are shown to the user. This clustering provides a quick visualization for identifying up- and down-regulation pertinent to the query genes.

Condition-specific view (**Supplementary Fig. 2.4**) is activated by clicking on the expression pattern of a gene in a particular data set. This view allows users to associate coexpressed genes with the meta-information (or measured outcome) attached to the data set, such as disease state, cell type, cell line, drug treatment, and patient characteristics. Users can choose among the data set's available attributes, and re-cluster the selected data set based on an attribute of interest. For example, by selecting the attribute "anatomical sites" for a Hedgehog related data set, and viewing the Hedgehog genes in the context of anatomical sites, they can observe that Hedgehog signaling is abundant in testis and pancreas, but not in lymph node tissues (**Supplementary Fig. 2.4**). Thus, potential associations to various measured outcomes can be readily uncovered post-search through the condition-specific view.

SEEK's **coexpression view** (**Supplementary Fig. 2.5**) provides a "bird's-eye" view of the coexpression landscape across up to 50 data sets at a time. Users can readily identify the data sets that are most relevant for the query, based on the coexpression of each retrieved gene to the query visualized as single columns. Users can readily assess the contribution of each data set. This view also serves to visually analyze the query coherence (**Supplementary Fig. 2.5,** top heat-map), helping users in constructing a coherent query gene set, which in turn guides SEEK in producing more relevant results.

Downstream analyses – Refine Search

An important feature of SEEK is providing the user with flexible search refinement options (**Supplementary Fig. 2.6**). Although the SEEK algorithm enables robust search over the whole expression compendium, there are cases when users intend to restrict the

search domain to a subset of data sets, for instance, when they desire a tissue- or disease-oriented coexpression analysis, or when the user encounters a situation when her query is too small or heterogeneous, and the intended context is not readily identifiable from the query alone. The **Refine Search** function provides users with several ways of refining the search analysis. Users may narrow their results down by:

1) Limiting to a tissue or disease of interest. There are currently hundreds of selectable tissues, cell-types, and diseases defined by UMLS and BRENDA keywords.
2) Limiting the search to only cancer or non-cancer data sets. The cancer data compendium includes primary tumors, metastasized tumors, and cancer cell lines. The non-cancer compendium includes diverse non-cancer samples, including stem cells, muscle and adipose cells, neurodegenerative, immune and infectious disease samples, epithelial and endothelial cell types, and blood cell types in non-cancerous diseases.
3) Limiting to multi-tissue profiling data sets only. This group of 13 data sets is useful for checking the expression of gene(s) across normal tissues, cell lines, cell types, and diseased tissues from various organs.
4) Limiting to primary tumor data sets only. Users can select the 224 TCGA RNASeq data sets as well as around 200 data sets from independent research studies that profile single-tissue tumors in each data set.

SEEK provides users with an easy-to-use and easily searchable data set-type selector (**Supplementary Fig. 2.6**). After a category has been selected, SEEK will

perform the data set prioritization and coexpressed gene search within the chosen category of data sets only.

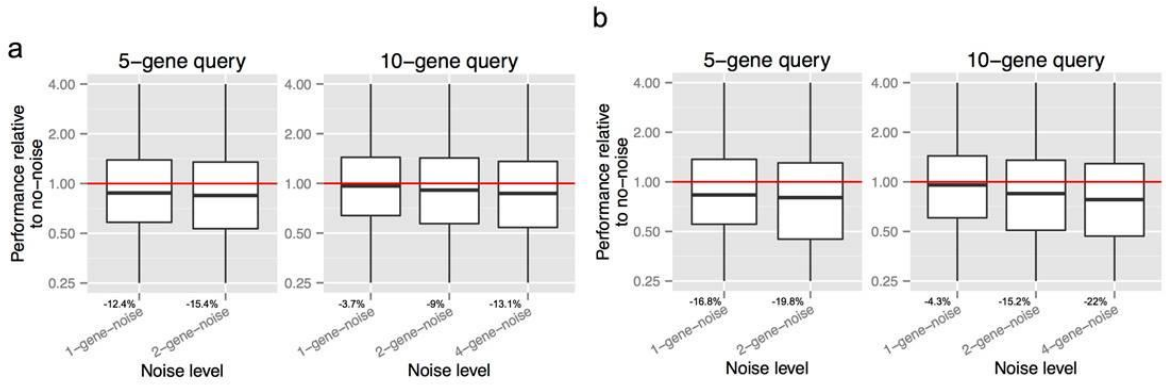
A.3 ModSEEK hedgehog ligand tissue contexts

In our evaluation, ModSEEK's automatic prioritization of murine datasets has accurately rediscovered the divergent roles of the hedgehog ligands *Shh*, *Dhh*, and *Ihh*. These results are consistent with multiple previous independent studies of individual ligands:

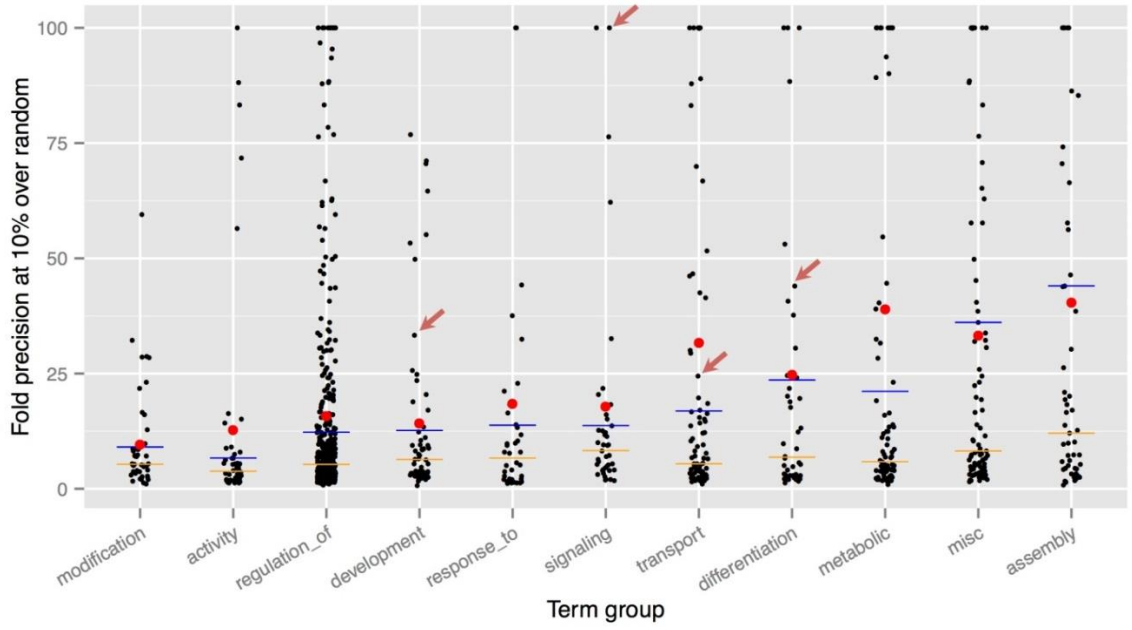
Ligand	Tissue contexts revealed by ModSEEK's data-driven dataset prioritizations	Supporting evidences
Shh	Skin, hair follicles, epidermis, hypothalamus, brain	Hedgehog signalling in skin development and cancer (Athar *et al* [136]) Shh expression is required for embryonic hair follicle but not mammary gland development (Michno *et al* [137])
Dhh	Gonads, testis, ovary	Distinct roles for Steroidogenic factor 1 and Desert hedgehog pathways in fetal and adult leydig cell development (Park *et al* [138])
Ihh	Chondrocytes, intestines	Indian hedgehog signaling regulates proliferation and differentiation of chondrocytes and is essential for bone formation. (St-Jacques *et al* [139]) Indian hedgehog regulates intestinal stem cell fate through epithelial-mesenchymal interactions during development (Kosinski *et al* [140])

The link between SHH-mediated hedgehog signaling and basal cell carcinoma (BCC) is well supported by the study Daya-Grosjean *et al* [141]. In our results, we have inferred squamous cell carcinoma which includes BCC.

SUPPLEMENTARY FIGURES

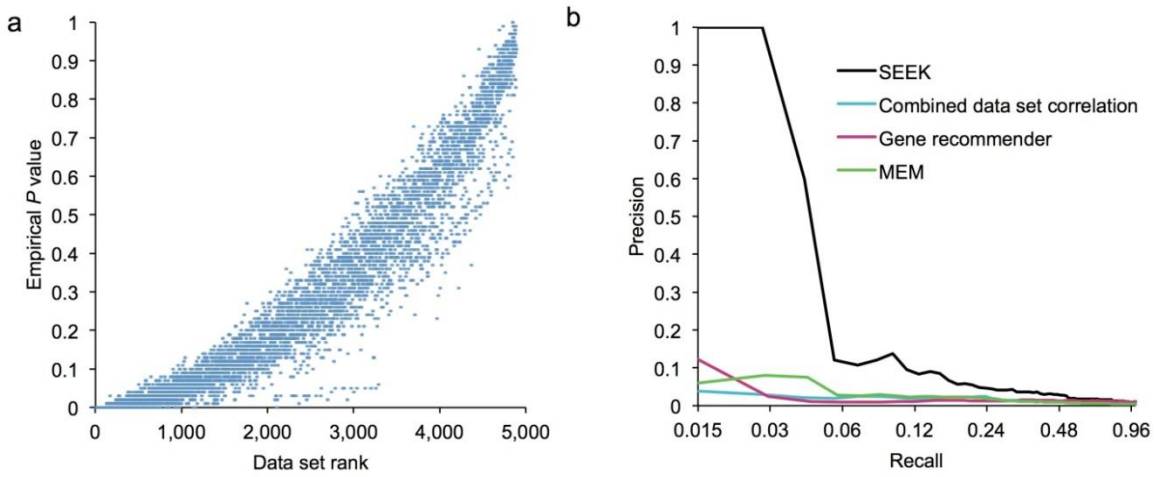


Supplementary Figure 2.1 | Robustness of SEEK and Gene Recommender to noisy query genes. (**a**) SEEK's retrieval robustness in the presence of random gene noise in the query. Red line (at 1.0) denotes the no-noise queries' performance level. Relative performance, defined by the fraction in fold improvement of precision over random at 10% recall (FIOR@10%) between noisy and no-noise queries, is plotted (see Section 2.4 **Methods**). The percentage numbers below the box plot shows the median per-query performance drop. (**b**) Gene Recommender's gene retrieval robustness in the presence of random gene noise in the query.

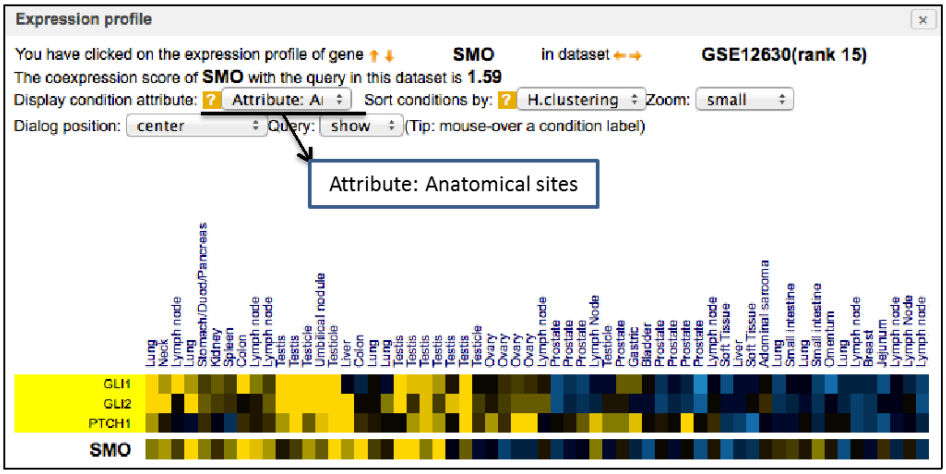


Super-groups' definition:
Modification: phosphorylation, methylation, glycosylation, acetylation, alkylation
Activity (i.e., enzymatic activity terms): cAMP, GTPase, ligase, kinase activities
Misc (i.e., miscellaneous terms): transcription, translation, mitosis, meiosis, cell migration, cell motility, cell respiration, cell adhesion, and DNA replication

Supplementary Figure 2.2 | SEEK's performance across process groups. Each black dot represents a process. Different statistics were used to summarize performance (FIOR@10%) per group: red dot (mean), blue line (75^{th} percentile), orange line (median). Memberships of the biological processes to the 11 term groups are determined by text-mining the process title, except for the 3 super-groups (see figure for their definitions). Red arrows indicate examples of top-performing processes: "erythrocyte differentiation" (44-fold), "lysosomal transport" (25-fold), "glutamate signaling" (104-fold), and "digestive system development" (33-fold).

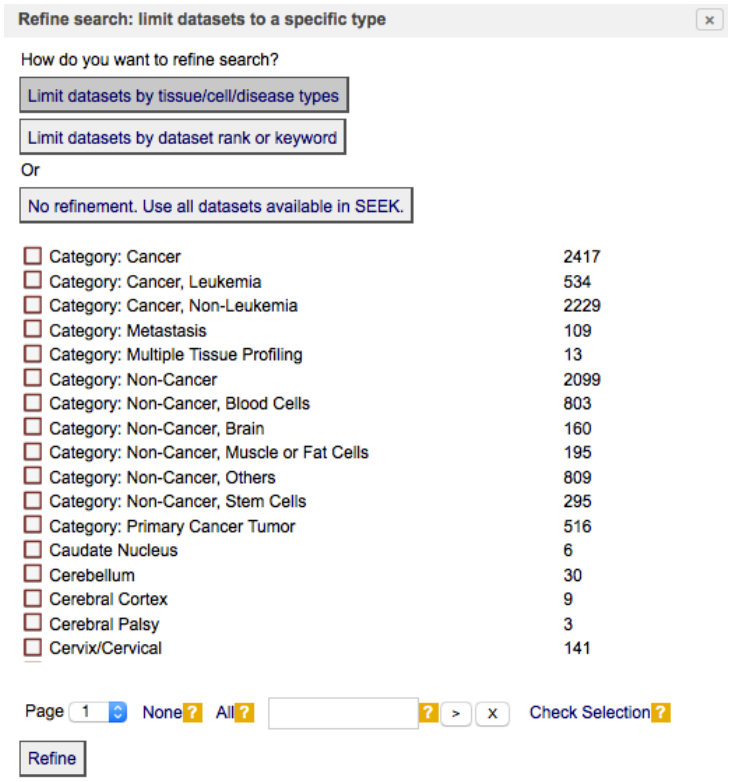


Supplementary Figure 2.3 | Search results for the Hedgehog (Hh) signaling query GLI1 GLI2 PTCH1: data set weight significance and gene-retrieval validation. (a) Top-ranked data sets are specifically highly weighted to the Hedgehog (Hh) query. Data set weight significance is calculated by a comparison with 100 random queries. Empirical P value for each data set *d* represents the fraction of 100 random queries where the score (or weight) of the data set *d* prioritized by a random query is higher than *d*'s score in the Hh query. (b) Gene retrieval validation, which serves as an indication of the relevance of the search results (i.e., coexpressed genes) to the Hh context. The gold standard consists of 71 Hh genes assembled from KEGG and GO.

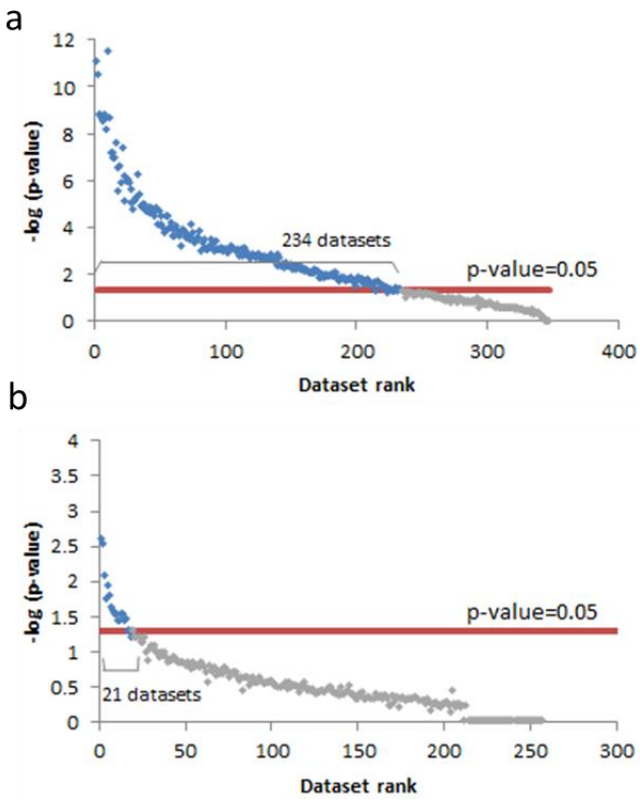


Supplementary Figure. 2.4 | Condition-specific view. This zoom-in view is generated by clicking on the row corresponding to *SMO* and GSE12630 data set in the result page of the *GLI1*, *GLI2*, *PTCH1* query.

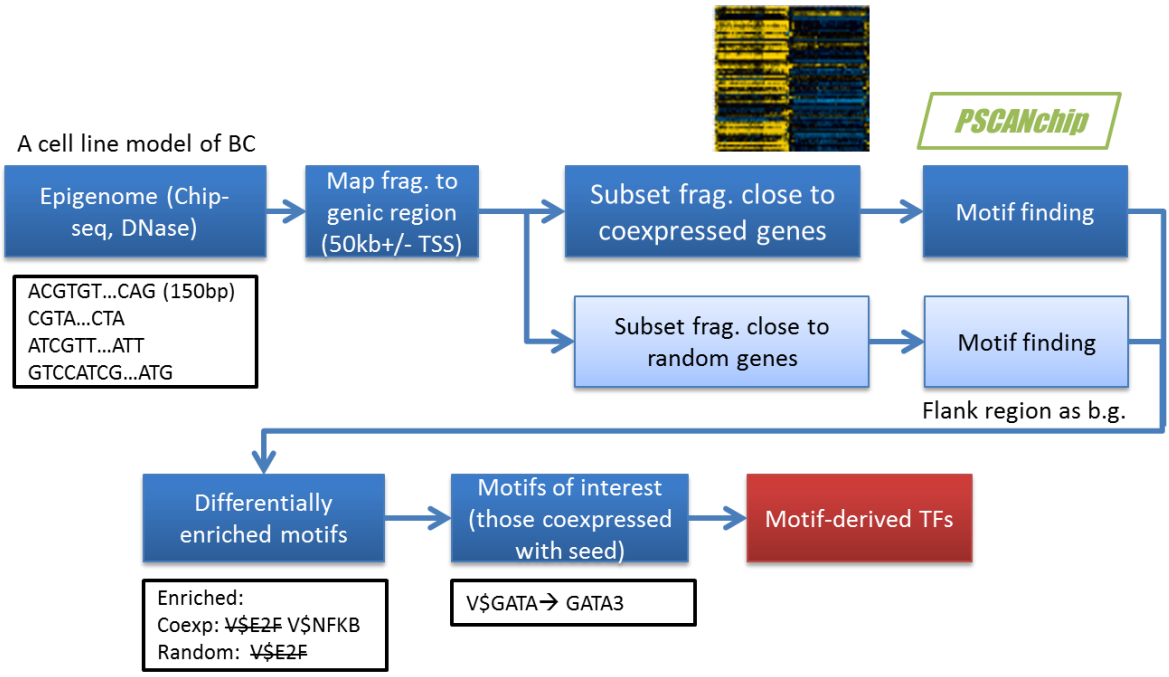
Supplementary Figure 2.5 | Coexpression view. Top heat map: query coherence, measured by the degree to which each query gene correlates with the rest of the query across the top 50 data sets. Each column represents a data set. Any "outlier" genes can thus be identified and subsequently removed from the query.



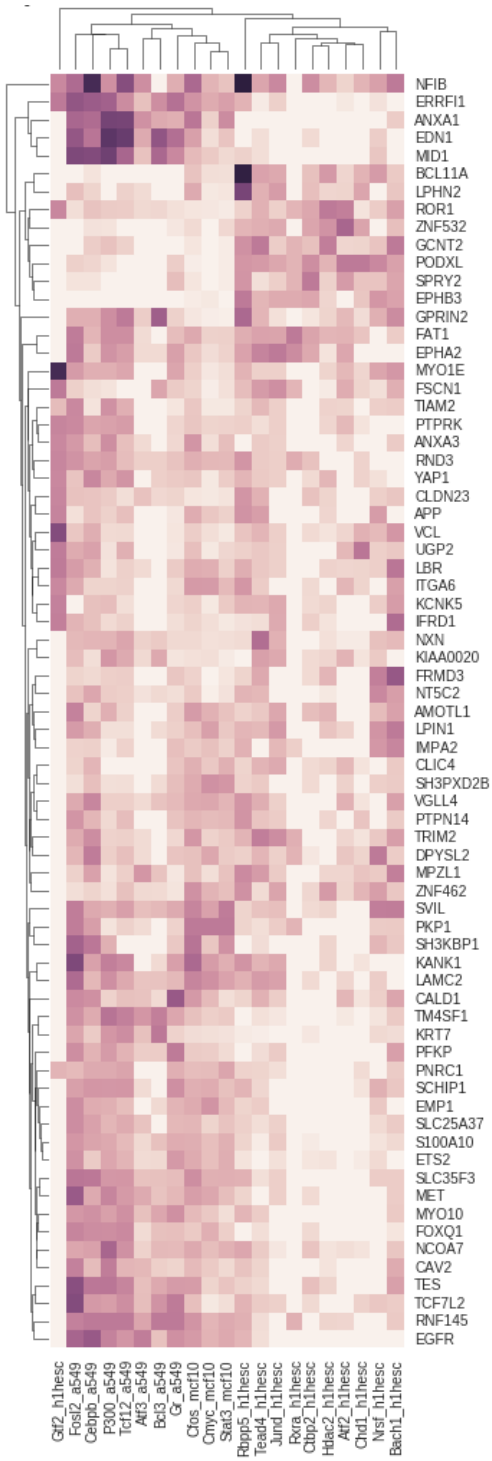
Supplementary Figure 2.6 | Available options within the Refine Search window. The second column lists the number of data sets in each data set category.



Supplementary Figure 3.1 | Two types of query. (**a**) A ubiquitous-type query (*PROS29 RPN12 PROSBETA2 PROS26.4 CG30382*). (**b**) A unique-type query in fly endocytosis process, *HTL VPS28 SKTL HSP70BC GAP69C*. Plots show the -log of *p* value of dataset weight as a function of dataset rank.



Supplementary Figure 5.1 | Generation of motif-derived TFs.



Supplementary Figure 5.2 | Basal coexpressed targets reveal epithelial and stem cell lineages. **Rows**: top basal coexpressed genes that are targets of basal-specific TF regulators. **Columns**: TF regulators, ChIP'd from ENCODE with cell line indicated. **Entry** in heatmap: binding abundance of a TF at the upstream region of a coexpressed gene. The heatmap indicates that the binding pattern at the coexpressed genes can be

largely divided into epithelial cell specific (A549) (e.g. NFIB, ERRFI1, ANXA1, EDN1, MD1), stem cell specific (H1hesc) (e.g. BCL11A, LPHN2, FOR1, etc), and both (A549 + H1hesc) (e.g. GRPN2, FAT1, EPH2). This is evident from the clustering of Chip experiments.

SUPPLEMENTARY DATA

Supplementary Data 2.1 – 2.5.

All supplementary data can be found at the Nature Methods website or at the PubMed Central website:

http://www.nature.com/nmeth/journal/v12/n3/full/nmeth.3249.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4768301/

Supplementary Data 5.1. Seed genes.

Luminal A	Ref[113]	ENPP5 TCEA3 NPNT SNX13 STEAP2 MGRN1 CYP2A6 KIAA0182 ECHDC2 FMO5 SELENBP1 MUC1 CRAT CFB RARRES3 CYB5A GALNT10 HSD17B4 APPL2 PTP4A2 ASAH1 ALCAM MSX2 SLC40A1 SVEP1 SNED1 PLAT FMOD ADRA2A ECE1 BCAM SHC2 ACBD4 GSTM3 CAMK2N1 RALGPS1 PTPRN2 BLVRA AGTR1 NPY1R TLE3 PHF15 MED13L CCND1 QDPR SIAH2 COX6C SCNN1A TFF3 MCCC2 FBP1 ANXA9 REEP5 LRBA HEXIM1 BECN1 TCEAL1 RERG SLC39A6 RABEP1 ESR1 ACADSB VAV3 NAT1 SCUBE2 GATA3 FOXA1 XBP1
Luminal B	Ref [114]	CDC6 CCNB1 UBE2T NUF2 BLVRA SLC39A6 ESR1 CXXC5
Her2-enriched	Ref[113]	TBPL1 TLK1 FLOT2 SMARCE1 MED24 STARD3 GRB7 ERBB2 S100P CEACAM6
Basal-like	Ref[113]	ZNF532 B3GNT5 CDK6 KDSR NCL SLC5A6 CHI3L2 SLPI CXCL1 VGLL1 DSC2 FOXC1 MFGE8 ACTG2 GABRP TRIM29 KRT5 KRT17 CX3CL1 CDH3 SGCE FZD7 VCL EXT2
Normal-like	Ref[113]	KRT13 RAPGEF3 RAB11FIP5 ACSS2 GNB2L1 TFAP2C GSTA4 CA2 AQP3 AKR1C1 ACSL1 LTF PIK3R1 ABLIM1 PTPRM PAM

Supplementary Data 5.2. Breast cancer compendium.

GSE10270.GPL570	GSE10810.GPL570	GSE11395.GPL1352	GSE12306.GPL96
GSE10281.GPL570	GSE11001.GPL570	GSE11965.GPL96	GSE12622.GPL1708
GSE10780.GPL570	GSE11121.GPL96	GSE12093.GPL96	GSE12665.GPL6480
GSE10797.GPL571	GSE11394.GPL1352	GSE12276.GPL570	GSE12763.GPL570

GSE12814.GPL570	GSE22035.GPL570	GSE29832.GPL570	GSE43365.GPL570
GSE13274.GPL570	GSE22093.GPL96	GSE2990.GPL96	GSE43502.GPL570
GSE13671.GPL570	GSE22495.GPL6947	GSE30010.GPL570	GSE44408.GPL571
GSE1456.GPL96	GSE22544.GPL570	GSE30480.GPL6480	GSE45255.GPL96
GSE15363.GPL1708	GSE22580.GPL570	GSE30682.GPL6884	GSE45581.GPL6480
GSE1561.GPL96	GSE22597.GPL96	GSE31192.GPL570	GSE45584.GPL6480
GSE15852.GPL96	GSE22652.GPL6947	GSE31259.GPL6947	GSE4779.GPL1352
GSE16058.GPL3921	GSE22840.GPL570	GSE31429.GPL6947	GSE4823.GPL1708
GSE16391.GPL570	GSE23061.GPL3921	GSE31429_1.GPL6947	GSE4917.GPL96
GSE16446.GPL570	GSE23177.GPL570	GSE31519.GPL96	GSE4922.GPL96
GSE16873.GPL96	GSE23399.GPL570	GSE32072.GPL96	GSE6596.GPL96
GSE17072.GPL6884	GSE23500.GPL6947	GSE32518.GPL96	GSE6772.GPL96
GSE17215.GPL3921	GSE23593.GPL570	GSE32531.GPL6480	GSE9014.GPL1708
GSE17539.GPL570	GSE23988.GPL96	GSE32532.GPL6480	GSE9195.GPL570
GSE17700.GPL570	GSE23994.GPL570	GSE32646.GPL570	GSE9574.GPL96
GSE17700.GPL96	GSE24185.GPL96	GSE33692.GPL5175	GSE9649.GPL570
GSE17705.GPL96	GSE24202.GPL3921	GSE34487.GPL4133	GSE9662.GPL96
GSE17907.GPL570	GSE2429.GPL96	GSE3494.GPL96	GSE9691.GPL3921
GSE18728.GPL570	GSE24450.GPL6947	GSE35031.GPL6244	GSE9936.GPL96
GSE18864.GPL570	GSE24468.GPL570	GSE35118.GPL6244	TCGA-A1-01.RNASEQ
GSE18931.GPL570	GSE25011.GPL96	GSE36766.GPL570	TCGA-A2-01.RNASEQ
GSE19383.GPL570	GSE25173.GPL571	GSE36772.GPL96	TCGA-A7-01.RNASEQ
GSE19536.GPL6480	GSE25407.GPL570	GSE36773.GPL96	TCGA-A8-01.RNASEQ
GSE19615.GPL570	GSE25835.GPL3921	GSE37126.GPL6480	TCGA-AC-01.RNASEQ
GSE19697.GPL570	GSE26082.GPL6480	GSE3744.GPL570	TCGA-AN-01.RNASEQ
GSE19783.GPL6480	GSE26349.GPL571	GSE37485.GPL3921	TCGA-AO-01.RNASEQ
GSE20086.GPL570	GSE26457.GPL570	GSE37543.GPL6480	TCGA-AQ-01.RNASEQ
GSE20181.GPL96	GSE26910.GPL570	GSE37946.GPL96	TCGA-AR-01.RNASEQ
GSE20266.GPL570	GSE27220.GPL570	GSE38506.GPL570	TCGA-B6-01.RNASEQ
GSE20271.GPL96	GSE28583.GPL570	GSE39976.GPL571	TCGA-BH-01.RNASEQ
GSE20437.GPL96	GSE28694.GPL570	GSE41036.GPL571	TCGA-C8-01.RNASEQ
GSE20711.GPL570	GSE28796.GPL570	GSE41198.GPL1352	TCGA-D8-01.RNASEQ
GSE21422.GPL570	GSE28821.GPL570	GSE41227.GPL1352	TCGA-E2-01.RNASEQ
GSE21653.GPL570	GSE28826.GPL570	GSE41986.GPL6480	TCGA-E9-01.RNASEQ
GSE21947.GPL96	GSE29431.GPL570	GSE42568.GPL570	TCGA-EW-01.RNASEQ
GSE21974.GPL6480	GSE29561.GPL96	GSE42822.GPL96	TCGA-GM-01.RNASEQ

Supplementary Data 5.3. SEEK expanded coexpressed gene lists.

Luminal A SEEK analysis	http://seek.princeton.edu/viewer33.jsp?sessionID=1469824943666&sort_sample_by_expr=true Coexpressed genes (includes P-values for all genes): http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469824943666&type=gene_score&keyword=all_sorted&show_query=true Prioritized datasets: http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469824943666&type=dataset_weight&keyword=all_sorted
Luminal B SEEK analysis	http://seek.princeton.edu/viewer33.jsp?sessionID=1469825046409&sort_sample_by_expr=true Coexpressed genes (includes P-values for all genes): http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=14698250464

	09&type=gene_score&keyword=all_sorted&show_query=true Prioritized datasets: http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469825046409&type=dataset_weight&keyword=all_sorted
Basal-like SEEK analysis	http://seek.princeton.edu/viewer33.jsp?sessionID=1469825090287&sort_sample_by_expr=true Coexpressed genes (includes P-values for all genes): http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469825090287&type=gene_score&keyword=all_sorted&show_query=true Prioritized datasets: http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469825090287&type=dataset_weight&keyword=all_sorted
Normal-like SEEK analysis	http://seek.princeton.edu/viewer33.jsp?sessionID=1469825139414&sort_sample_by_expr=true Coexpressed genes (includes P-values for all genes): http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469825139414&type=gene_score&keyword=all_sorted&show_query=true Prioritized datasets: http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469825139414&type=dataset_weight&keyword=all_sorted
Her2-enriched SEEK analysis	http://seek.princeton.edu/viewer33.jsp?sessionID=1469824901640&sort_sample_by_expr=true Coexpressed genes (includes P-values for all genes): http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469824901640&type=gene_score&keyword=all_sorted&show_query=true Prioritized datasets: http://seek.princeton.edu/servlet/GetScoreServlet?sessionID=1469824901640&type=dataset_weight&keyword=all_sorted

Supplementary Data 5.4. ENCODE DNase and Chip-seq experiments used for motif-finding analysis

Luminal A	MCF7 DNase Hypersensitivity (ENCODE)
	T-47D DNase Hypersensitivity (ENCODE)
	Eralpha T-47D Chip-seq (ENCODE)
	Foxa1 T-47D Chip-seq (ENCODE)
	Gata3 T-48D Chip-seq (ENCODE)
Basal-like	A549 DNase Hypersensitivity (ENCODE)
	H1hesc DNase Hypersensitivity (ENCODE)
	Cfos Mcf10aes Chip-seq (ENCODE)
	Stat3 Mcf10aes Chip-seq (ENCODE)

REFERENCES

1. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30,** 207–10 (2002).

2. Brazma, A. *et al.* ArrayExpress - A public repository for microarray gene expression data at the EBI. *Nucleic Acids Research* **31,** 68–71 (2003).

3. Brazma, A., Hingamp, P. & Quackenbush, J. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29,** 365–71 (2001).

4. Brown, P. O. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23,** 41–46 (1999).

5. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

6. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 8418–23 (2003).

7. Scherzer, C. R. *et al.* Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc.Natl.Acad.Sci.U.S.A* **104,** 955–960 (2007).

8. Gawantka, V. *et al.* Gene expression screening in Xenopus identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mech. Dev.* **77,** 95–141 (1998).

9. Zhang, W. *et al.* The functional landscape of mouse gene expression. *J. Biol.* **3,** 21 (2004).

10. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci.* **98,** 13790–13795 (2001).

11. Leek, J., Scharpf, R. & Bravo, H. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).

12. Ramasamy, A., Mondry, A., Holmes, C. C. & Altman, D. G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* **5,** e184 (2008).

13. Tseng, G. C., Ghosh, D. & Feingold, E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* **40,** 3785–99 (2012).

14. Rhodes, D. R. *et al.* Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.

Proc. Natl. Acad. Sci. U. S. A. **101,** 9309–14 (2004).

15. Rhodes, D. R. *et al.* Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia* **9,** 166–180 (2007).

16. DeConde, R. P. *et al.* Combining results of microarray experiments: a rank aggregation approach. *Stat. Appl. Genet. Mol. Biol.* **5,** Article15 (2006).

17. Hong, F. *et al.* RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* **22,** 2825–2827 (2006).

18. Choi, J. K., Yu, U., Kim, S. & Yoo, O. J. Combining multiple microarray studies and modeling interstudy variation. in *Bioinformatics* **19,** (2003).

19. Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M. & Nobel, A. B. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics* **24,** 1154–1160 (2008).

20. Mistry, M. & Pavlidis, P. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics* **9,** 327 (2008).

21. McGinnis, S. & Madden, T. L. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32,** (2004).

22. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. **9,** 3273–3297 (1998).

23. Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. & Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **14,** 1085–94 (2004).

24. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249–55 (2003).

25. Tian, W. *et al.* Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. *Genome Biol.* **9 Suppl 1,** S7 (2008).

26. Liao, B.-Y. & Zhang, J. Evolutionary Conservation of Expression Profiles Between Human and Mouse Orthologous Genes. *Mol Biol Evol* **23,** 530–540 (2006).

27. Movahedi, S., Van de Peer, Y. & Vandepoele, K. Comparative Network Analysis Reveals That Tissue Specificity and Gene Function Are Important Factors Influencing the Mode of Expression Evolution in Arabidopsis and Rice. *PLANT Physiol.* **156,** 1316–1330 (2011).

28. Bergmann, S., Ihmels, J. & Barkai, N. Similarities and Differences in Genome-Wide Expression Data of Six Organisms. *PLoS Biol* **2,** e9 (2003).

29. Netotea, S., Sundell, D., Street, N. R. & Hvidsten, T. R. ComPlEx: conservation and divergence of co-expression networks in A. thaliana, Populus and O. sativa. *BMC Genomics* **15,** 106 (2014).

30. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34,** 166–76 (2003).

31. Madeira, S. C. & Oliveira, A. L. Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. Comput. Biol. Bioinforma.* **1,** 24–45 (2004).

32. Alter, O., Brown, P. O. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc.Natl.Acad.Sci U.S.A* **97,** 10101–10106 (2000).

33. Hibbs, M. a *et al.* Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics* **23,** 2692–9 (2007).

34. Cancer, T. & Atlas, G. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–8 (2008).

35. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14863–8 (1998).

36. Howe, E. *et al.* MeV: MultiExperiment Viewer. *Biomed. Informatics Cancer Res.* 267–277 (2010).

37. Saeed, A. I. *et al.* [9] TM4 Microarray Software Suite. *Methods Enzymol.* **411,** 134–193 (2006).

38. Tanay, A., Sharan, R., Kupiec, M. & Shamir, R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 2981–6 (2004).

39. Huttenhower, C. *et al.* Exploring the human genome with functional maps. *Genome Res.* **19,** 1093–106 (2009).

40. Wong, A. K. *et al.* IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res.* **40,** W484–90 (2012).

41. Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C. & Morris, Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.* **9 Suppl 1,** S4 (2008).

42. Adler, P. *et al.* Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. *Genome Biol.* **10,** R139 (2009).

43. Owen, A. B., Stuart, J., Mach, K., Villeneuve, A. M. & Kim, S. A Gene Recommender Algorithm to Identify Coexpressed Genes in C. elegans. *Genome Res.* **13,** 1828–1837 (2003).

44. Kretzler, M. & Cohen, C. Integrative biology of renal disease: towards a holistic understanding of the kidney's function and failure. *Semin. Nephrol.* **30,** 439–442 (2010).

45. Chen, R., Mallelwar, R., Thosar, A., Venkatasubrahmanyam, S. & Butte, A. J. GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed. *BMC Bioinformatics* **9,** 548 (2008).

46. Zinman, G. E., Naiman, S., Kanfi, Y., Cohen, H. & Bar-Joseph, Z. ExpressionBlast: mining large, unstructured expression databases. *Nat. Methods* **10,** 925–926 (2013).

47. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33,** e175 (2005).

48. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24,** 1151–1161 (2006).

49. Lukk, M. *et al.* A global map of human gene expression. *Nat. Biotechnol.* **28,** 322–4 (2010).

50. Schmid, P. R., Palmer, N. P., Kohane, I. S. & Berger, B. Making sense out of massive data by going beyond differential expression. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 5594–9 (2012).

51. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5,** 101–13 (2004).

52. Han, J.-D. J. *et al.* Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430,** (2004).

53. Myers, C., Barrett, D. & Hibbs, M. Finding function: evaluation methods for functional genomic data. *BMC Genomics* **7,** 187 (2006).

54. Kimura, H., Stephen, D., Joyner, A. & Curran, T. Gli1 is important for medulloblastoma formation in Ptc1+/- mice. *Oncogene* **24,** 4026–36 (2005).

55. Oliver, T. G. *et al.* Transcriptional profiling of the Sonic hedgehog response: a critical role for N-myc in proliferation of neuronal precursors. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 7331–6 (2003).

56. Berman, D., Karhadkar, S. & Hallahan, A. Medulloblastoma growth inhibition by hedgehog pathway blockade. *Science (80-.).* **297,** 1559–1561 (2002).

57. Carpenter, D. *et al.* Characterization of two patched receptors for the vertebrate.

Proc. Natl. Acad. Sci. U. S. A. **95,** 13630–13634 (1998).

58. Oue, T., Yoneda, A. & Uehara, S. Increased expression of the hedgehog signaling pathway in pediatric solid malignancies. *J. Pediatr. Surg.* **45,** 387–92 (2010).

59. Jagani, Z., Mora-Blanco, E. & Sansam, C. Loss of the tumor suppressor Snf5 leads to aberrant activation of the Hedgehog-Gli pathway. *Nat. Med.* **16,** 1429–33 (2010).

60. Cohen, M. The hedgehog signaling network. *Am. J. Med. Genet. Part A* **123A,** 5–28 (2003).

61. Cheung, H. O.-L. *et al.* The kinesin protein Kif7 is a critical regulator of Gli transcription factors in mammalian hedgehog signaling. *Sci. Signal.* **2,** ra29 (2009).

62. Fisher, R. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10,** 507–521 (1915).

63. Song, L., Langfelder, P. & Horvath, S. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* **13,** 328 (2012).

64. Horvath, S. & Dong, J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* **4,** e1000117 (2008).

65. Ruan, J., Dean, A. K. & Zhang, W. A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Syst. Biol.* **4,** 8 (2010).

66. Xulvi-Brunet, R. & Li, H. Co-expression networks: graph properties and topological comparisons. *Bioinformatics* **26,** 205–14 (2010).

67. Moffat, A. & Zobel, J. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.* **27,** 1–27 (2008).

68. Huttenhower, C., Schroeder, M., Chikina, M. D. & Troyanskaya, O. G. The Sleipnir library for computational functional genomics. *Bioinformatics* **24,** 1559–61 (2008).

69. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32,** D267–70 (2004).

70. Gremse, M., Chang, A. & Schomburg, I. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39,** D507–13 (2011).

71. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28,** 573–80 (2012).

72. Irizarry, R. a *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–64 (2003).

73. Zahurak, M. *et al.* Pre-processing Agilent microarray data. *BMC Bioinformatics* **8,** 142 (2007).

74. Smyth, G. Limma: linear models for microarray data. in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds. Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.) 397–420 (Springer, 2005).

75. Smyth, G. K. *et al.* limma: Linear Model for Micrarray Data User's Guide. at <http://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf>

76. Bolstad, B. M., Irizarry, R. a, Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19,** 185–93 (2003).

77. Ritchie, M. E. *et al.* A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **23,** 2700–7 (2007).

78. National Cancer Institute. RNASeq Version 2. at <https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2>

79. Mose, L. & Parker, J. V2_MapSpliceRSEM: UNC V2 RNA-Seq Workflow - MapSplice genome alignment and RSEM estimation of GAF 2.1. at <https://confluence.broadinstitute.org/download/attachments/29790363/DESCRIPTION.txt>

80. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).

81. Obayashi, T. & Kinoshita, K. COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res.* **39,** D1016–D1022 (2011).

82. Van Dam, S., Craig, T. & De Magalhães, J. P. GeneFriends: A human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.* **43,** D1124–D1132 (2015).

83. Chu, L.-H., Vijay, C. G., Annex, B. H., Bader, J. S. & Popel, A. S. PADPIN: Protein-Protein Interaction Networks of Angiogenesis, Arteriogenesis, and Inflammation in Peripheral Arterial Disease. *Physiol. Genomics* physiolgenomics.00125.2014 (2015). doi:10.1152/physiolgenomics.00125.2014

84. Zhu, Q. *et al.* Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat. Methods* **12,** 211–4 (2015).

85. Knijnenburg, T. A., Wessels, L. F. A., Reinders, M. J. T. & Shmulevich, I. Fewer permutations, more accurate P-values. in *Bioinformatics* **25,** (2009).

86. Villaseñor-Alva, J. A. & González-Estrada, E. A bootstrap goodness of fit test for

the generalized Pareto distribution. *Comput. Stat. Data Anal.* **53,** 3835–3841 (2009).

87. Lipscomb, C. Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.* (2000). at <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC35238/>

88. Benjamini, J. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).

89. Adler, P., Peterson, H., Agius, P., Reimand, J. & Vilo, J. Ranking genes by their co-expression to subsets of pathway members. *Ann. N. Y. Acad. Sci.* **1158,** 1–13 (2009).

90. Alemu, E. Y., Carl, J. W., Bravo, H. C. & Hannenhalli, S. Determinants of expression variability. *Nucleic Acids Res.* **42,** 3503–3514 (2014).

91. Park, C. Y. *et al.* Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput. Biol.* **9,** e1002957 (2013).

92. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).

93. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33,** (2005).

94. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36,** D480–D484 (2008).

95. Perou, C., Sørlie, T. & Eisen, M. Molecular portraits of human breast tumours. *Nature* **406,** 747–52 (2000).

96. Hu, Z. *et al.* The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* **7,** 96 (2006).

97. Cancer, T. & Atlas, G. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

98. Sørlie, T. & Perou, C. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* **98,** 10869–74 (2001).

99. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–52 (2012).

100. Nik-Zainal, S., P, V. L., DC, W., LB, A. & CD, G. The life history of 21 breast cancers. *Cell* **149,** 994 (2012).

101. Gerstein, M. M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489,** 91–100 (2012).

102. Kittler, R. *et al.* A Comprehensive Nuclear Receptor Network for Breast Cancer

Cells. *Cell Rep.* **3,** 538–551 (2013).

103. Janky, R. *et al.* iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Comput. Biol.* **10,** (2014).

104. Altman, R. B., Dunker, A. K. & Hunter, L. Bioprospector: Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes. *Pacific Symp. Biocomput.* **138,** 127–138 (2001).

105. Joshi, H., Nord, S. H., Frigessi, A., Børresen-Dale, A.-L. & Kristensen, V. N. Overrepresentation of transcription factor families in the genesets underlying breast cancer subtypes. *BMC Genomics* **13,** 199 (2012).

106. Tongbai, R. *et al.* Transcriptional Networks Inferred from Molecular Signatures of Breast Cancer. *Am. J. Pathol.* **172,** 495–509 (2008).

107. Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Lee Kraus, W. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* **23,** 1210–1223 (2013).

108. Li, W. *et al.* Condensin I and II Complexes License Full Estrogen Receptor ??-Dependent Enhancer Activation. *Mol. Cell* **59,** 188–202 (2015).

109. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).

110. Theodorou, V., Stark, R., Menon, S. & Carroll, J. S. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res.* **23,** 12–22 (2013).

111. Zhu, Q. *et al.* Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat. Methods* **12,** 211–214 (2015).

112. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

113. Muggerud, A. A. *et al.* Evaluation of MetriGenix custom 4D arrays applied for detection of breast cancer subtypes. *BMC Cancer* **6,** 59 (2006).

114. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27,** 1160–7 (2009).

115. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22,** 1813–1831 (2012).

116. Neph, S. *et al.* BEDOPS: High-performance genomic feature operations. *Bioinformatics* **28,** 1919–1920 (2012).

117. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC*

Bioinformatics **10,** 48 (2009).

118. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158,** 1431–1443 (2014).

119. Zambelli, F., Pesole, G. & Pavesi, G. PscanChIP: Finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic Acids Res.* **41,** (2013).

120. Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat. MethodsPerspective* **10,** (2013).

121. Sarrió, D. *et al.* Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res.* **68,** 989–997 (2008).

122. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat. Genet.* **43,** 27–33 (2011).

123. Prenzel, T. *et al.* Estrogen-dependent gene transcription in human breast cancer cells relies upon proteasome-dependent monoubiquitination of histone H2B. *Cancer Res.* **71,** 5739–5753 (2011).

124. Khaled, W. T. *et al.* BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. *Nat. Commun.* **6,** 5987 (2015).

125. Adam, R. C. *et al.* Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature* **521,** 366–70 (2015).

126. H, Z., F, M., G, L. & B, Z. Forkhead transcription factor foxq1 promotes epithelial-mesenchymal transition and breast cancer metastasis. *Cancer Reseach* **71,** 1292–1300 (2011).

127. Fleischer, T. *et al.* Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol.* **15,** 435 (2014).

128. Weigman, V. J. *et al.* Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res Treat* **133,** 865–880 (2012).

129. Kwon, M. & Shin, Y. Regulation of ovarian cancer stem cells or tumor-initiating cells. *Int. J. Mol. Sci.* **14,** 6624–48 (2013).

130. Steg, A., Bevis, K. & Katre, A. Stem cell pathways contribute to clinical chemoresistance in ovarian cancer. *Clin. Cancer Res.* **18,** 869–81 (2012).

131. Homayouni, R., Rice, D. & Curran, T. Disabled-1 interacts with a novel developmentally regulated protocadherin. *Biochem. Biophys. Res. Commun.* **289,** 539–47 (2001).

132. Katoh, Y. & Katoh, M. WNT antagonist, SFRP1, is Hedgehog signaling target. *Int. J. Mol. Med.* **17,** 171–5 (2006).

133. Schreck, K. C. K. *et al.* The Notch target Hes1 directly modulates Gli1 expression and Hedgehog signaling: a potential mechanism of therapeutic resistance. *Clin. cancer Res.* **16,** 6060–70 (2010).

134. Jalali, A., Bassuk, A. & Kan, L. HeyL promotes neuronal differentiation of neural progenitor cells. *J. Neurosci. Res.* **89,** 299–309 (2011).

135. Chuang, P. & McMahon, A. Vertebrate Hedgehog signalling modulated by induction of a Hedgehog-binding protein. *Nature* **397,** 617–21 (1999).

136. Athar, M., Tang, X., Lee, J. L., Kopelovich, L. & Kim, A. L. Hedgehog signalling in skin development and cancer. *Experimental Dermatology* **15,** 667–677 (2006).

137. Michno, K., Boras-Granic, K., Mill, P., Hui, C. C. & Hamel, P. A. Shh expression is required for embryonic hair follicle but not mammary gland development. *Dev. Biol.* **264,** 153–165 (2003).

138. Park, S. Y., Tong, M. & Jameson, J. L. Distinct roles for Steroidogenic factor 1 and Desert hedgehog pathways in fetal and adult leydig cell development. *Endocrinology* **148,** 3704–3710 (2007).

139. St-Jacques, B., Hammerschmidt, M. & McMahon, A. P. Indian hedgehog signaling regulates proliferation and differentiation of chondrocytes and is essential for bone formation. *Genes Dev.* **13,** 2072–2086 (1999).

140. Kosinski, C. *et al.* Indian hedgehog regulates intestinal stem cell fate through epithelial-mesenchymal interactions during development. *Gastroenterology* **139,** 893–903 (2010).

141. Daya-Grosjean, L. & Couve-Privat, S. Sonic hedgehog signaling in basal cell carcinomas. *Cancer Lett* **225,** 181–192 (2005).