Targeted analyses of very large genome-wide data collections

Young-suk Lee

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
Adviser: Professor Olga Troyanskaya

April 2016

Abstract

Genome-scale experiments provide an overwhelming amount of molecular information for biologist. New computational methods are needed for specific analysis and interpretation of such high-dimensional data. Here we take advantage of the massive public repositories to quantify the tissue-specific signals in gene expression profiles, characterize distinctive molecular features of human diseases, deconvolve the latent cell-type-specific factors in mixed clinical samples, and automatically integrate heterogeneous data sources in the context of a specific genome-wide dataset. First, we describe URSA (Unveiling RNA Sample Annotation) that incorporates the known tissue/cell-type relationships to better estimate the specific signal in any given gene expression profile. Our ontology-aware method combines independent discriminative classifiers in a Bayesian framework, outperforming other machine learning methods. We provide a molecular interpretation for the tissue and cell-type models learned by URSA, enabling a data-driven view of molecular processes specific to particular tissues and cell types. Then, we extend this work for human diseases. We use thousands of clinical disease-specific expression profiles in public repositories to quantify distinctive functional and anatomical characteristics of human diseases. Through our data-driven analysis, we explore the complexity of the human disease landscape and propose exploratory hypothesis for drug repurposing even for rare disease with no prior genetic knowledge. Lastly, we describe YETI (Your Evidence Tailored Integration) for targeted integration of heterogeneous genome-wide data sources. Biomedical researchers generate genome-wide datasets for data-driven exploration of specific questions but such analyses are disconnect from big public data collections. YETI is the first automatic integration method that effectively constructs functional networks specific to a genome-scale dataset. We show that the resulting integration reflect the biological context of the user-provided dataset while providing accurate prediction for functional interactions.

Acknowledgements

I would like to thank my advisor Olga Troyanskaya for her continuous support and thoughtful guidance during my graduate career at Princeton University. She was the first to introduce me to microarrays and needless to say, none of this work would have been possible without her. Many thanks to Barbara Engelhart for her constant encouragement during my last years at Princeton. I'd also like to thank the rest of my committee Thomas Funkhouser, Mona Singh and John Storey for their critical feedback and comments in completing my dissertation.

Much of this work was done amongst many friends and colleagues of the Troyanskaya lab: Patrick Bradley, Ruth Dannenfelser, Dima Gorenshteyn, Jonathan Goya, Casey Greene, Yuanfang Guan, Max Homilius, Arjun Krishnan, Chris Park, Ana Pop, Aaron Wong, Vicky Yao, Ran Zhang, Jian Zhou, and Qian Zhu. Thank you for all the insightful discussions that shaped the narrative of this work. In particular, I want to thank Arjun Krishnan, Chris Park and Aaron Wong for their mentorship and honest advice, especially during the beginning of my graduate career. I'd like to thank Bong-Ihn Koh for walking me through the often overlooked details in molecular biology experiments. Many thanks to Allison Chaney, Rajesh Ranganath, and Arman Suleimenov for sharing their expertise in the fast growing field of machine learning and human-computer interaction.

Special thanks to Nicki Gotsis and Melissa Lawson from the Department of Computer Science and Barbara Chinery, Marybeth Fedele and John Wiggins from Lewis-Sigler Institute for Integrative Genomics for their technical and administrative support.

I must say this work would not have been possible without all the distractions that kept me balanced during the compilation of this work. Many thanks to: Sehyoun Ahn, Eun Jeong Choi, Song Ha Joo, Taehee Han, Jonghun Kam, In Song Kim, Jeong-Ho Kim, Ringi Kim, Donghun Lee, Taewook Oh, and Yeje Park. I'm surely indebted to

Mungja Kim, Munhee Lee, Teddy Son and other members of the Princeton Korean Community Church for their prayers and spiritual support.

To my parents, Chang-yong Lee and Soon-won Nam for their unconditional love. And my brother Niels Lee for always being there for me. Thank you all for patiently listening through my silly ideas.

Foremost, I thank God who sent his only son Jesus Christ who die on the cross for me. Not much has gone according to my original plan during my time at Princeton. A lot more failed than succeeded. Small triumphs helped, but it was His comfort that gave me the courage to look at the big picture and take the next step.

I was supported by the the Princeton graduate fellowship, NSF CAREER award (DBI-0546275), NIH grants (R01 GM071966, R01 HG005998 and T32 HG003284) and NIGMS Center of Excellence grant (P50 GM071508).

To my family.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The amount of data is growing and has grown to a point where we now call it 'big' data. This 'big' data is greeted by both the academic and commercial sectors with much optimism and hope in that all our questions may be solved in a robust data-driven manner. Very large data collections such as 1000 Genome Project, the Encyclopedia of DNA Elements (ENCODE), Genotype-Tissue Expression (GTEx), Gene Expression Omnibus (GEO), the Human Protein Atlas, and the Cancer Cell Line Encyclopedia (CCLE) survey the biological variation at a molecular and genome-wide level in hopes to elucidate the complete molecular map of the human cell [20, 21, 86, 28, 142, 6]. To meet such need, new infrastructures such as data centers and clouding computing platforms are being developed to handle the storage and analysis of these large data collections [119, 135]. However, the accessibility and interpretation of these databases have been more of a catalog of experimental results much like Amazon.com and eBay are for commercial products. It's left for the biologist to detect relevant patterns of biological significance of this overwhelming amount of molecular information. New computational methods are needed to better utilize these very large genome-wide data collections in a "targeted" manner that uncovers patterns of biological relevance.

The work presented here are three examples of targeted analysis of very large genome-wide data collection that is only made possible via the joint analysis of publicly available data collections. No single laboratory or institution is able to generate such vast data collections, and new biological guidance and insights are provided by applying these novel computational approaches. Here I enumerate specific computational challenges for each example such as handling batch effects and accounting for spurious correlations in these data collections. Then, I show how these new computational methods take advantage of the high-dimensionalty and heterogeneity of the data collections to: (1) quantify and detect tissue-specific signals, (2) characterize the human disease landscape and (3) infer functional gene-gene interactions tailored for a specific biological question.

1.1 Background

All living organism begin with one cell and one DNA sequence. That single cell becomes two and then four cells ultimately leading to a multi-cellular unit with distinct functional anatomical part such as the heart, kidney and brain. That is, all through an intricate molecular process. Malfunction of this molecular process leads to various human diseases such as cardiac arrhythmia, neuroblastoma, and Alzheimer's disease. Understanding the underlying molecular mechanism of tissue-specific function and disease is key to better diagnosis and targeted treatment.

The central dogma of molecular biology states that DNA sequence *replicates* to preserve molecular information and DNA *transcribes* to RNA which are used as templates for *translation* to protein. A *gene* is a region of the DNA that encodes a functional RNA, and the *genome* is the set of hereditary molecules in living organisms and is often used interchangeably with the DNA sequence. The Human Genome

Project revealed more than $20,000$ human genes, and yet the function of most genes are still unknown [72].

Thousands of genes interact in a dynamic molecular network and enable response to external stimulus and tissue-specific function of complex biological processes. High-throughput experiments such as microarrays or mass spectrometry provide a genome-wide snapshot of this molecular network in action. In particular, DNA microarrays are used to simultaneously measure the amount of expression (i.e. amount of transcripts) of thousands of genes [16]. Such transcriptional *profiling* provides a genome-wide perspective of the molecular activity across different tissue and cell-types and functional abnormalities in human disease samples.

In detail, a DNA microarray is a lab-on-a-chip with a designated collection of DNA spots (also known as *probes*). Each probe consist of picomoles of specific DNA sequences that correspond to 100-1000 bases long sections of the genome - often unique sections of a gene. Purified RNA from the sample of interest is converted to complementary DNA (cDNA) and then labelled with fluorescence dyes. When these labelled cDNAs are exposed to the DNA microarray, complementary sequences between cDNAs and probes bind (i.e. hybridize). The dye intensity of each spot represents the amount of the corresponding RNA in that original sample. Multiple microarray experiments thus provide a genome-wide perspective of a particular dynamic system such as response to external stimulus, different phases of the cell-cycle, heterogeneity of tumor samples [67, 141, 129].

More recently, next-generation sequencing (NGS) technology termed RNAseq has been used to directly *sequence* the RNA and explicitly count the number of RNA transcripts for each gene as oppose to infer abundance based on dye intensity [94, 144]. No predefined collection of probes are needed for RNAseq. While the number of publicly available sequence-based datasets are limited compared to the well-established mi-

croarray datasets, efforts to develop platform-independent methods are needed with such advancement of high-throughput experiments.

1.2 Challenges

1.2.1 High dimensional genome-wide data

The curse of dimensionality refers to the challenges of visualization and computational analysis of data in high-dimensional space. As the dimensionality of the data increases, the volumne of the data-space increases exponentially. As a consequence, the amount of data needed for statistically robust multivariate analysis grows exponentially. In genome-wide data, the dimensionality of the data is in the order of thousands, and genome-wide microarrays measure ten thousands of genes simultaneously. Dimension reduction methods such as Principal Component Analysis (PCA) or Factor Analysis have been used for quality control and exploratory data analysis [107]. Incorporating prior biological knowledge is needed to coupe with this high dimensionality of genome-wide data.

1.2.2 Unknown technical batch effects

Genome-wide profiling such as microarrays or RNAseq experiments is subject to technical noise or batch effects [79]. That is, the data clusters better by the date or author of the experiment instead of the sample type such as its tissue type or even organism. The clustering of mouse data against human data have been shown to be merely due to batch effects in the data [34]. Data normalization methods have been developed to address such concern but most effective for when those batch variables are known and available [12, 53, 78]. These batch variables are not always available in public data repositories. When analyzing genome-wide data collection, additional care must be taken to not over-estimate the method's performance due to batch effects.

1.2.3 Spurious genome-wide relationships

Spurious relationships are common in high dimensional data with batch effects. Thousands of genes are found statistically significant in a standard differential gene expression analysis, even after multiple hypothesis correction. To mitigate such phenomenon, independent positive and negative controls (i.e. gold standards) are need to evaluate and develop better methods. Identifying reproducible relationships (i.e. data redundancy) by integrating multiple datasets is another approach to account for spurious relationships in individual data.

1.3 Contributions

Leveraging gene expression data through large-scale integrative analyses for multicellular organisms is challenging because most samples are not fully annotated to their tissue/cell-type of origin. A computational method to classify samples using their entire gene expression profiles is needed. Such a method must be applicable across thousands of independent studies, hundreds of gene expression technologies and hundreds of diverse human tissues and cell-types. In Chapter 2, we present URSA (Unveiling RNA Sample Annotation) that leverages the complex tissue/cell-type relationships and simultaneously estimates the probabilities associated with hundreds of tissues/cell-types for any given gene expression profile. URSA provides accurate and intuitive probability values for expression profiles across independent studies and outperforms other methods, irrespective of data preprocessing techniques. Moreover, without re-training, URSA can be used to classify samples from diverse microarray platforms and even from next-generation sequencing technology. Finally, we provide a molecular interpretation for the tissue and cell-type models as the biological basis for URSA's classifications. This work has been published in Bioinformatics [77].

Complex diseases are driven by multiple genetic changes and characterized by genome-wide perturbations of cellular pathways and functions. Gene expression profiling experiments have been potent in shedding light on the molecular pathology of diseases. Most studies typically focus on a single disease and contrast disease samples to their normal controls. However, such one disease at a time approaches disregard similarities and differences in pathological deregulations underlying different complex diseases and are thus unable to identify attributes unique to each particular disease, which is critical for developing targeted therapy. In Chapter 3, We have developed a unified probabilistic framework URSA$^{\mathrm{HD}}$ (URSA for Human Diseases) to identify and quantify distinctive disease signals based on gene expression profiles of clinical samples. Leveraging thousands of disease-specific profiles from public repositories, this data-driven approach identified distinctive molecular-level characteristics of each disease from both the functional and anatomical perspectives. Our framework can be used to distinguish between closely-related diseases, identify discerning genes and processes, associate rare-diseases to the nearest well-studied disease, and track the effectiveness of therapy. No curated set of genes were used in our data-driven approach, and so it can easily be extended to any human disease for which high-throughput expression data can be generated. We found that the most predictive genes identified by our method are significantly under-studied in the biomedical literature, demonstrating that many key biological processes underlying human pathophysiology are in fact in critical need of further investigation. This work has been submitted for publication and is currently under review.

Integration of heterogeneous genome-wide data sources has been used to generate functional networks, predict gene function, and study human disease. Most biomedical researchers have specific questions they want to answer with such integration, and these questions are usually accompanied by a user-produced genome-scale dataset. However, no computational approach exists to enable such user's dataset-guided inte-

gration of large genome-wide data collections. In Chapter 4, we develop an automatic integration method YETI (Your Evidence Tailored Integration) that constructs functional networks specific to a genome-scale dataset. We show that the resulting integrations reflect the biological context of the user-provided dataset while also providing accurate functional predictions. YETI's dataset-specific networks are unbias to the size of the dataset and reproducible across biological replicates. As such, YETI's networks revealed putative functional network modules regulated by distal eQTLs hidden in co-expression networks. YETI's integration framework streamlines the integration process for biologists to easily access and take advantage of very large genome-wide data collections in the context of their specific but genome-wide question. A version of this work will be submitted for publication.

Chapter 2

Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies

This chapter describes work done with critical support and comments from Arjun Krishnan and Qian Zhu.

2.1 Introduction

Genome-scale expression profiling is an invaluable technique for quantifying gene-level activity under different experimental conditions. For more than a decade, researchers and clinicians have submitted their experimental data to public repositories such as NCBI's Gene Expression Omnibus (GEO) [7] and EBI's ArrayExpress [115]. These repositories now include almost half a million human expression profiles from multiple laboratories and hospitalsonly to further grow with the advent of next-generation se-

quencing technologies. Large but independent microarray datasets have been used to discover tissue-specific patterns [87, 125], establish breast cancer subtypes [99, 24] and delineate the transcriptome response to candidate drugs [45, 71]. Previous integrative studies have leveraged these independent datasets and have developed methods based on correlation [46], differential expression [29], supervised learning [40] and data integration [146]. However, directly dealing with multicellularity is paramount for precisely defining human homeostasis, disease manifestation and pharmacokinetics/pharmacodynamics. To some effect, few studies have focused on certain sample characteristics such as disease or phenotype [47, 120]. Yet, to take full advantage of the entire compendia in all the above contexts, we must explicitly uncover tissue/cell-type-specific signals in genome-wide expression data.

The current exponential rate of data submission nevertheless makes manual annotation impractical, leaving a curated annotation index for only a small fraction of samples 2.1. Text-mining sample descriptions are often unreliable due to the lack of standardized nomenclature and structured descriptive information [68]. Furthermore, textual information may not reflect the potential specificity and heterogeneity that are concealed in the molecular-level expression measurements of these samples. Therefore, we need a scalable and robust computational method to discover the tissue/cell-type signals in each gene expression profile deposited in these large heterogeneous data compendia.

In practice, tissue/cell-type annotation of gene expression profiles relies on the expression of known biomarker genes. Although pervasive, this approach is limited by the number of sufficient (or often any) known discriminative expression biomarkers and ignores potential specific signals in the entire transcriptome. Machine learning methods that model genome-wide expression have emerged as promising alternatives [82], but so far have only been applied in the context of classifying tumor subtypes (e.g. ALL versus AML) in single datasets [57, 108, 137]. Applying such

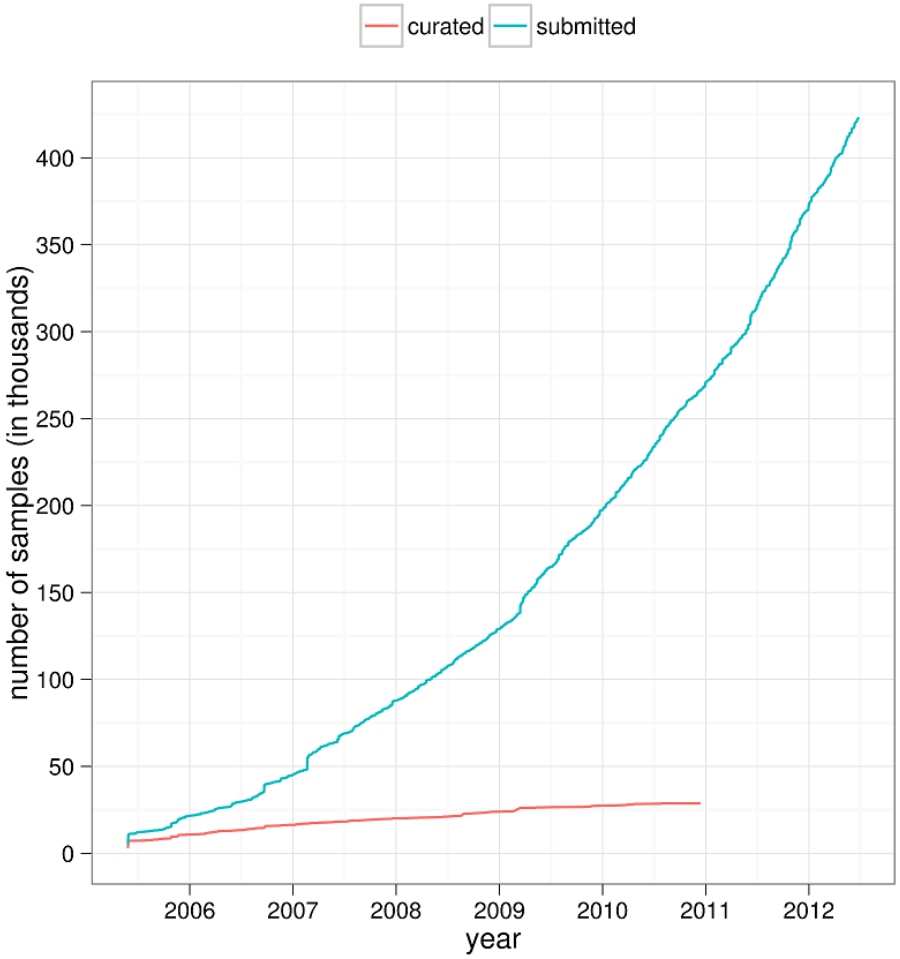


Figure 2.1: Manual curation is unable to keep up with the number of submissions. The line plot shows the number of submitted and curated human genome-scale experiments publicly available in GEO over time. Sample submissions have been growing exponentially in contrast to their manual annotation.

methods across a large collection of datasets is impeded by the dataset, platform and technology biases [79, 113]. The only successful attempt at addressing dataset biases is a nearest -neighbor (NN) classification method based on the barcode algorithm [92, 151].

The task of indexing these large heterogeneous data collections by tissues/cell-types presents substantial challenges. First, a successful method for this task should be able to classify the variety of human tissues/cell-types, not just the better-studied large tissue classes. For example, classifying blood from brain is a relatively easy problem, but discriminating among different subtypes of blood is a much harder one. Second, the method should maintain consistency with the developmental and anatomical relationships between these tissues and cell-types. Third, the method must be robust across independent datasets to overcome study/laboratory biases. Finally, with emerging profiling technologies, the method should be readily applicable to novel platforms/technologies. No existing approach, to our knowledge, addresses all these challenges.

Here, we present a computational algorithm Unveiling RNA Sample Annotation (URSA) that is the first to leverage the relationships between tissues and cell-types (based on a tissue ontology) and accurately identifies specific tissue/cell-type signals present in a given gene expression profile. URSA constructs individual tissue/cell-type classifiers based on ontology-aware sample labels and uses Bayesian Network Correction (BNC) [10] to integrate these individual classifiers. We demonstrate that URSA substantially outperforms barcode-based NN classification (the only prior approach to this problem) [151], as well as independent classifiers that do not use the tissue ontology. Furthermore, although URSA is trained on data from the single most popular microarray platform, it is able to make tissue/cell-type predictions (without re-training) for samples measured by other microarray platforms and even by next-generation RNA sequencing.

In the process of classification, our approach learns tissue/cell-type signals without the use of any tissue-specific gene database such as the human protein reference database [105]. Thus, by examining the biological pathways enriched among the feature-weights in each tissue/cell-type classifier, we are able to provide a molecular-level interpretation of URSA's predictions.

2.2 Methods

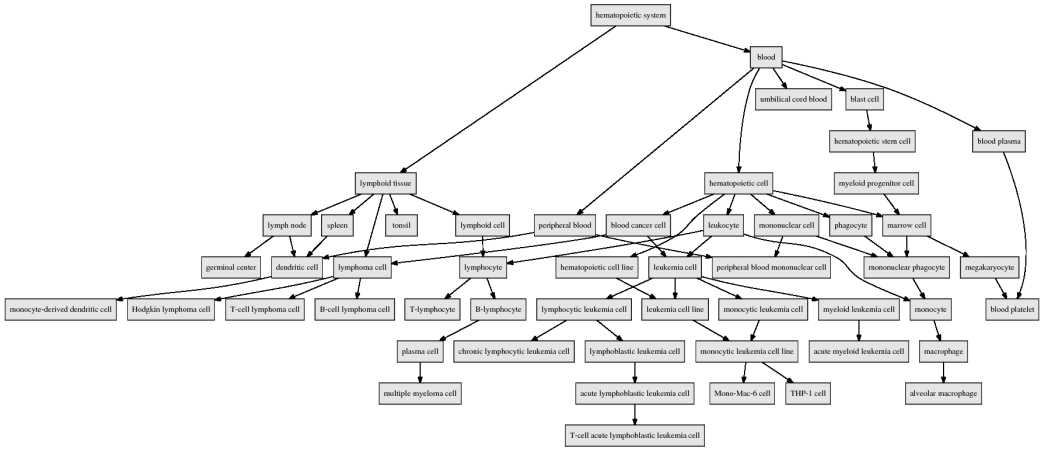


Figure 2.2: The full hematopoietic system sub-ontology used in URSA. URSA incorporates this complexity to provide meaningful, associative probability values for all terms.

We setup the tissue/cell-type signal classification problem as a hierarchical multilabel classification problem. From a curated collection of samples, we first label samples into positives and negatives based on the tissue ontology to train an individual classifier for each tissue/cell-type. We then aggregate these individual classifiers (in a Bayesian framework) based on their ontological relationships (Figure 2.2) [41]. Each individual classifier identifies indicative features (i.e. genes) for that tissue or cell-type, and the Bayesian network models the probabilistic relationship between classifiers to refine those individual predictions. We have previously demonstrated that such BNC improves classification accuracy in other settings, including gene

function prediction and geometric shape classification [9, 10, 44, 100]. URSA uses the BNC approach to tackle the challenges in tissue and cell-type prediction: limited gold standards for many general (e.g. leukocyte) and specific (e.g. T-cell acute lymphoblastic leukemia cell and monocyte-derived dendritic cell) tissues/cell-types, and heterogeneity and diversity in large expression compendia.

2.2.1 Gold standard generation by manual sample annotation

GEO provides a structured sample description with dedicated entries such as Title, Source name, and Characteristics but enforces no controlled vocabulary. As a result, authors can freely use their nomenclatures or acronyms to fill-in any of these structured entries. For example, sometimes patient's ethnicity would be included in the Source name instead of the tissue/cell-type information. Often patient information would be included in the sample description with no structural distinction between patient and sample information. Excluding patient cancer-type information in free-text is difficult: peripheral blood mononuclear cell samples from patients with gastrointestinal and/or brain cancer, for example. Thus relying solely on text-mining methods is prone to many mis-annotations and requires post-manual curation.

In order to utilize the tissue relationships, hgu133plus2 gene expression experiments were annotated to a term or terms in the Brenda Tissue Ontology. After an initial substring text-mining of sample descriptions in GEO, term-to-experiment pairs were manually curated based on their sample descriptions and associated publication(s) to exclude incorrect or ambiguous pairs. Mixed tissue samples and non-human samples were excluded. An effort was made to annotate experiments to their most specific term in the ontology, although wasn't systematically enforced. Samples of specific tissue/cell-type not included in the tissue ontology were annotated to a higher-level term; for example, CD4+ T cell samples were annotated to T-

lymphocyte. The associated publication (i.e. original paper) was examined when the sample descriptions were ambiguous. Sample annotations were then propagated to more general terms based on the tissue ontology.

The scope and depth of our analysis is defined and hence limited by the completeness of the tissue ontology and extent of manual curation of public hgu133plus2 microarray experiments. Gene expression experiments were only annotated to tissue/cell-type terms included in the tissue ontology as the ontology terms were used as a controlled vocabulary. As a result, terms absent in the ontology (such as CD4+ T cell) were not used in manual curation. In addition, only tissue/cell-types with experiments available from at least three independent studies were modeled and evaluated to avoid overestimating the performance due to dataset bias. Our manual annotations are available on our website: ursa.princeton.edu

A set of high-quality tissue and cell-type annotations is needed for training accurate classifiers within the URSA framework. To this end, we manually annotated the cell-type(s) of ¿14 000 microarray experiments ranging over 500 GEO series/datasets from the hgu133plus2 platform. These annotations are based on the sample descriptions and other textual information available in GEO as well as the associated publications. Tissue and cell-type terms in the BRENDA Tissue Ontology (BTO) were used as the controlled vocabulary for sample annotation [41]. Detailed description of the manual sample annotation process is provided in the Supplementary Information. In our manual annotations, 71 tissue/cell-type terms were represented by at least 3 GEO series and 95 terms were represented in at least 2 GEO series. We excluded the term connective tissue from the ontology because it had many children terms, and thus appeared unresolved.

With an ontology of tissues, these manually curated annotations can be hierarchically propagated: e.g. monocyte samples can also annotated to leukocyte and blood (Figure 2.2). The minimal subgraph (i.e. directed acyclic graph) that is rooted at the

whole body term and includes all cell-type terms covered in our manual annotations was identified, and our manual annotations for 95 tissues/cell-types were then propagated up to their ancestors based on the tissue ontology, hence providing examples for over 244 different tissue/cell-type terms.

2.2.2 Expression data preparation

The Supplementary raw CEL files of gene expression samples were downloaded from GEO, and their probes were mapped to Entrez GeneIDs using the BrainArray Custom CDF [7, 25]. To compare methods across different preprocessing techniques, expression data were processed using each of the three alternative preprocessing algorithms: MAS5.0, fRMA and Barcode [48, 91, 92]. Default parameters and subroutines were used for each preprocessing approach. Additionally, the absolute expression values from the standard MAS5.0 were log transformed. As our method aims at classifying single expression profiles, series-based preprocessing techniques (i.e. RMA) were excluded from our study [53]. The Illumina Human Bodymap 2.0 RNA-seq data (GSE30611) was downloaded from GEO and mapped to NCBI's transcript reference using the Bowtie and Tophat alignment algorithms with default parameters [73, 138]. For tissue/cell-type prediction, FPKM transcript expression values were given as input to our hgu133plus2-trained method. Data transformation and significance test for RNA-seq (and cross platform) experiments are explained later in this section.

2.2.3 Individual tissue and cell-type classifiers

Labeling positive and negative samples correctly is essential for any accurate classifier. Conventional multilabel classification assumes that all labels are mutually exclusive. For example in our study, macrophage samples would be considered negative examples when classifying for leukocytes, ignoring the fact that macrophages are merely a specific type of leukocytes (Figure 2.3a). Using the tissue ontology, we thus re-

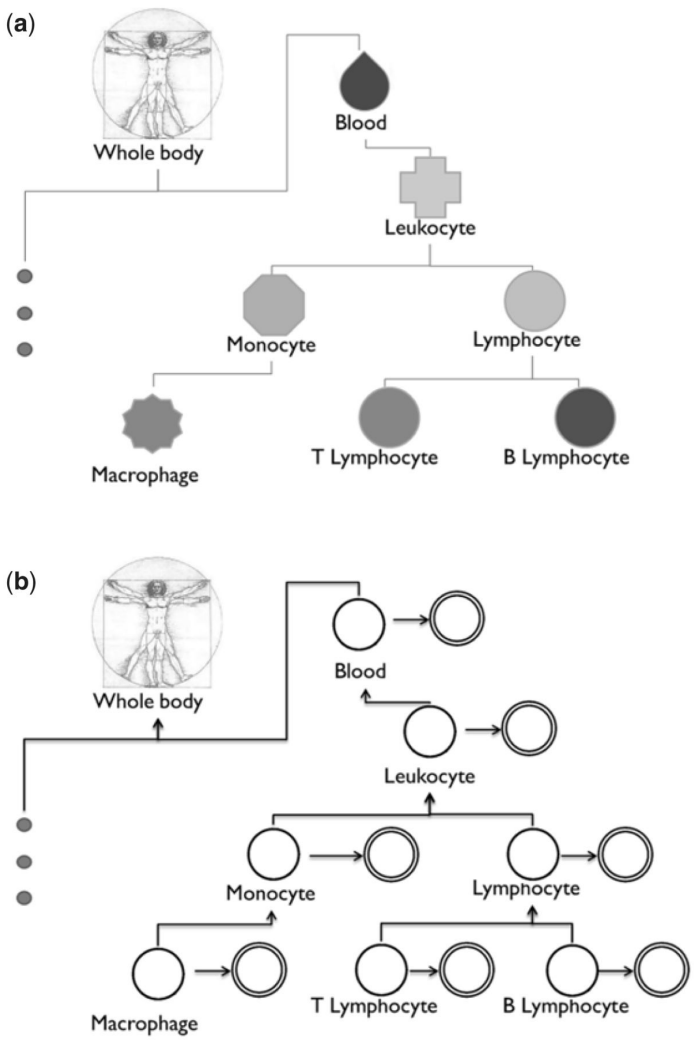


Figure 2.3: Leveraging the complex relationship between tissues and cell-types. (a) A small sub-tree of the BTO. The full hematopoietic system sub-ontology is shown in Figure 2.2. This complexity has yet been incorporated in tissue and cell-type-specific studies. (b) Our aggregation method uses this ontological structure to model the potential dependencies between individual cell-type models. The double circles indicate the noisy individual model predictions $\hat{y}_i$, and the single circles indicate the latent calibrated predictions y_i

consider the positive and negative samples for each individual tissue and cell-type classifier. For a given tissue term, samples annotated directly to that term or any of its descendant terms (i.e. cell-types) are now considered positive; samples annotated to only its ancestor terms are excluded from training; and the remaining samples annotated to other term in the ontology including sibling terms are considered negative. Now, macrophage samples would be considered positive examples for the leukocyte classifier. This re-labeling is based on the very design of the tissue ontology, and consequently expands the number of positive examples and removes ambiguous examples.

Each individual tissue or cell-type is first classified using an independent one-versus-all support vector machine (SVM) classifier using the ontology-aware training standard. SVM maximizes the margin between positive (i.e. $y_i = 1$) and negative (i.e. $y_i = 1$) examples and finds a linear decision boundary without any assumptions of the probability distributions [17]. Given l pairs (i.e. samples) of expression data x_i and its label y_i, we use the L2 linear SVM (with the default cost parameter) implemented in the LIBLINEAR software [30]:

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{l} \max(1 - y_i w^T x_i, 0)^2 \tag{2.1}$$

where C ¿ 0 is the cost parameter, and w the linear decision boundary (i.e. feature weight vector). Bayesian correction (explained later in text) is trained and applied using the SVM outputs $\hat{y}_1, \ldots, \hat{y}_N$ of these N cell-type-specific models.

2.2.4 Bayesian network correction

We use the structure of the tissue ontology as a framework of the Bayesian network (Figure 2.3a). We model each term's SVM output as a random event $\hat{y}_i$ and treat it as a noisy observation of a latent binary event y_i representing the true label (i.e. cell-

type) of a given sample (Figure 2.3b). The edges from y to $\hat{y}$ impose the independence of the noisy random variable $\hat{y}_i$ to all other noisy variables $\hat{y}_j$ $(i \neq j)$ given its true label y_i. This allows us to calculate the likelihood:

$$\mathrm{P}(\hat{y}_1, \ldots, \hat{y}_N | y_1, \ldots, y_N) = \prod_{i=1}^{N} \mathrm{P}(\hat{y}_i | y_i) \tag{2.2}$$

The distribution of positive and negative unthresholded SVM outputs varies across different terms (i.e. cell-types), and so the output values were dynamically binned based on the number of positive examples and their range. These empirical distributions represent the conditional probabilities $\mathrm{P}(\hat{y}_i | y_i = 0)$ and $\mathrm{P}(\hat{y}_i | y_i = 1)$. The conditional probability table for each term was estimated based on a 2-fold cross-validation that never split datasets between folds to mitigate potential batch effects. Laplace smoothing was applied for robustness.

The parentchild conditional probability tables were defined as in the original Bayesian correction method [10]. Intuitively, constant priors of 0.5 were assigned to leaf nodes, and the whole-body root node was assigned a probability of 1. This root assignment allows potential dependencies between every latent variable. This allows us to calculate the prior:

$$\mathrm{P}(y_1, \ldots, y_N) = \prod_{i=1}^{N} \mathrm{P}(y_i | \mathrm{ch}(y_i)) \tag{2.3}$$

where $\mathrm{ch}(y_i)$ is child labels of y_i.

Finally, we infer the posterior probabilities $\mathrm{P}(y_i | \hat{y}_1, \ldots, \hat{y}_N)$ for each term i using the Lauritzen algorithm as implemented in the SMILE library [27, 75]. These posterior probabilities for each term (i.e. cell-type) are the estimated probabilities that our method uses to annotate gene expression samples.

2.2.5 Method training and testing

Genomic experiments are known to suffer from potential laboratory and dataset biases [79, 151]. Not controlling for this bias (during evaluation of any method applied to these data) may result in an overestimation of performance and overfitting to dataset-specific biases at the expense of the desired signals. Therefore, for each cell-type, the series/datasets of the manually annotated samples were partitioned into three sets with each set containing roughly the same number of samples. Two partitions were used as the training set and the other as the testing set. Never splitting a single series/dataset between training and test sample sets ensures that our approach does not identify signals specific to particular studies, but rather those reflective of cell-types and tissues.

2.2.6 Cross-platform prediction

The individual classifiers in URSA were trained on samples only from the most popular Affymetrix Human Genome U133 Plus 2.0 platform (hgu133plus2). URSA has not been explicitly modified or tuned for predicting across other platforms. As input to our method, a gene expression profile from other array-based and sequence-based platforms were quantile transformed to generate a hgu133plus2-like expression profile. Additionally, a permutation test was performed to correct for potential biases from gene coverage differences across platforms.

Quantile transformation

The individual cell-type models in URSA have been trained on one microarray platform (hgu133plus2). To detect cell-type-specific information from other gene expression platforms, we must first transform those expression values to a comparable expression space. If we can effectively transform those values, our method without any modifications or retraining may be able to measure cell-type-specific signals

in these cross-platform experiments. The individual expression values xi may not be comparable across different platform technologies (especially between array-based and sequence-based platforms), but signals based on the relative abundance between genes should be more or less preserved irrespective of the technology used. Therefore, we quantile transform these cross-platform samples to preserved their relative gene abundances (or gene order) and compute hgu133plus2-like expression values based on a hgu133plus2 reference distribution. This reference distribution was constructed by averaging the expression value of each quantile across 1000 random hgu133plus2 arrays.

The most crucial bottleneck for cross-platform annotation is the bias in gene coverage across platforms. The hgu95v2 microarray platform (covering 12 000 genes), for example, covers about two-thirds the genes covered by hgu133plus2 (18 000 genes). Classification is handicapped by thousands of missing values, and hence, the mean expression value of the reference distribution was used to impute missing gene values [139].

Permutation test

A simple permutation test was performed to select significant predictions. The input data x_j consist of real and imputed gene values. Introducing noise to the actual data will blur any real signal and decrease its associated probability value. Thus, we permute only the sample data $\pi_1(x_j), \ldots, \pi_K(x_j)$ to generate a null distribution of SVM outputs $\pi(\hat{y}_i) = (\pi_k(\hat{y}_1), \ldots, \pi_k(\hat{y}_N))$, where $\pi_k(\hat{y}_i) = w_i^T \cdot \pi_k(x_j)$. This null distribution is then used to call out questionable annotations: any tissue annotation $\mathrm{P}(y_i|\hat{y}_1, \ldots, \hat{y}_N)$ with a value lower than even a single random annotation $P(y_i|\pi_k(\hat{y}_1), \ldots, \pi_k(\hat{y}_N))$ is considered insignificant and assigned a value of 0.

2.2.7 Double-blind evaluation of sample annotations

In addition to the evaluation based on our manual sample annotations, we conducted a rigorous double-blind literature-based study to evaluate the quality of URSA's novel predictions. To control for any subjective bias, we must also evaluate a random group of predictions in the same literature-based study. First, 120 hgu133plus2 array experiments from GEO that were not in our manual annotation were randomly selected. These experiments were partitioned into three groups. URSA annotations were made for all samples, but only group 1 predictions were retained and group 2 samples were assigned predictions from group 3. This procedure provides random annotations while ensuring the same apparent behavior of predictions as true predicted annotations. We use this conservative background to completely blind the evaluator from distinguishing original from random annotations.

The quality of predicted annotations should be judged based on retrieval of both the most precise tissue term and more general terms consistent with the precise term. For example, an acute lymphocytic leukemia (ALL) sample predicted to ALL but also other non-blood related terms such as urinary bladder and colon is precise but not consistent, whereas the same sample predicted to blood cancer cell or leukocyte in addition to ALL is both precise and consistent. Estimated annotations in group 1 (i.e. original) and group 2 (i.e. random) were evaluated as precise and/or consistent based on associated publications and textual sample descriptions. We also repeated this double-blind study for other microarray platforms: hgu133a, hgu95v2 and hugene1.0st.

2.3 Results

We address the cell-type prediction challenge as a multilabel classification problem with hierarchical constraints to account for the diverse nature of biological samples.

We assess the impact of incorporating the tissue ontology in our method and the method's robustness across different microarray preprocessing methods. Although our method can be readily retrained to any additional expression technologies given manually curated samples, we find that our method is capable of precisely annotating samples across platforms (including next-generation sequencing-based assays) without any modifications to the original method or its parameters. We finally show that our tissue/cell-type predictions are interpretable based on the biological processes enriched among learned informative genes.

2.3.1 URSA uses the tissue ontology to accurately predict tissue/cell-type signals

To address the challenge of limited gold standards and high noise levels in the tissue/cell-type classification problem, URSA incorporates the BTO to better predict tissue/cell-type signals in a given gene expression sample. BTO systematically defines parent-to-child relationships between tissue and cell-type terms [41]. URSA wields the complexity of this ontology to both systematically label samples to train tissue/cell-type SVM classifiers and also apply BNC to make consistent predictions.

To measure the impact of incorporating the ontology, we compare URSA with individual (i.e. independent) one-versus-all SVM classifiers whose outputs are converted to estimate probability values using logistic regression [104]. For these one-versus-all SVMs, whole blood samples are considered as negatives in a leukocyte classifier, for example. Both methods were trained on $\sim$ 9000 samples and tested on 5000 independent samples (Fig. 2a). The top-predicted term for each sample was evaluated and automatically considered incorrect if the estimated probability value was below a cutoff. Multiple cutoffs from 0 (i.e. no cutoff) to 0.9 (i.e. high-confidence cutoff) were tested (Fig. 2a). This setup simulates the user experience with a predefined cutoff and penalizes correct top predictions with a low probability value.

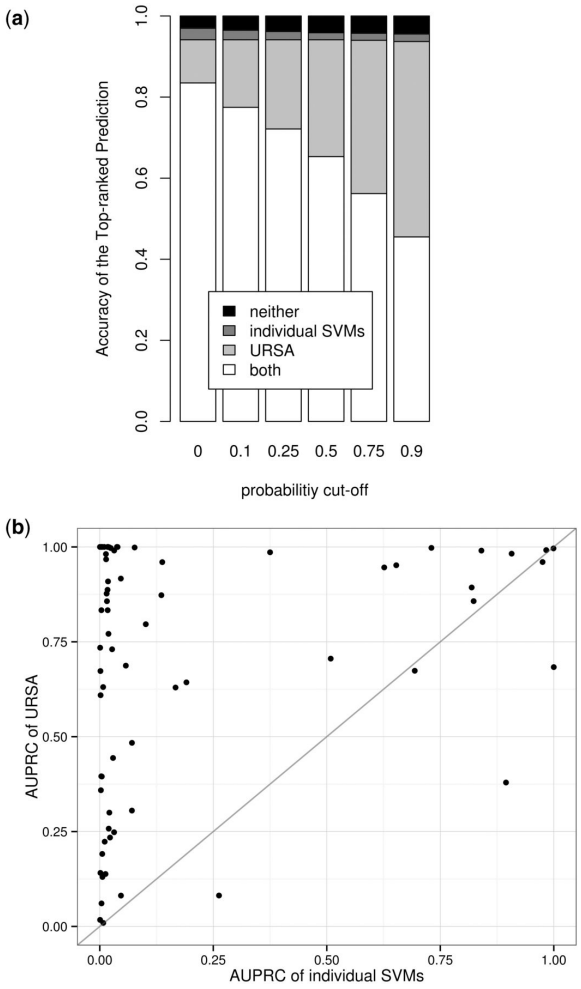


Figure 2.4: Prediction accuracy improves after integrating the tissue ontology. MAS5 was used for preprocessing. (a) Accuracy of the most probable estimation above a range of probability cutoffs. Estimations below the probability cutoff are discarded. The Bayesian framework corrects many of the mistakes made by individual SVMs and provides meaningful probability values. (b) Scatter plot comparison between URSA and individual SVM classifiers. Each point represents a unique tissue/cell-type with direct sample annotations, and the size of the point represents the number of samples curated to that particular tissue or cell-type. Points above the diagonal correspond to improvements by our method. URSA's improvements are independent of term size.

Across the entire range of probability cutoffs, URSA offers accurate top predictions for more samples in the holdout set. Without a cutoff on the estimated probabilities, both naïve SVM and ontology-aware URSA show considerable accuracy of the top-predicted term over the heterogeneous evaluation set (Figure 2.4a, leftmost bar). However, URSA accurately predicts an additional ~ 550 samples misclassified by the independent SVMs. Furthermore, URSA conveniently computes a probability value for each predicted tissue/cell-type annotation that provides a natural intuition about the strength of the predicted tissue/cell-type signal present in a given sample. Although probabilities can also be obtained for the individual SVMs, URSA's Bayesian framework provides a unified probabilistic model that enforces potential dependencies between distant and close tissues. This abstraction consequently computes consistent parameter estimations: e.g. if the probability for leukocyte is high, then the probability for blood should also be high, but not necessarily vice versa. Lending import to the calibrated probability values calculated by BNC, the proportion of URSA's accurate corrections of SVM's mis-annotations increases with higher probability cutoffs (Figure 2.4a). In case of high confidence predictions (0.9 cutoff), URSA provides accurate annotations for 94% of the test samples, 45% (> 2200) of which were misclassified by the individual SVMs.

The performance over different probability cut-offs was shown to access the quality of the estimated probability values. Figure 2.4a represents the accuracy for the top-predicted term for each sample across all holdout samples (i.e. gene expression experiments). The white bar represents the proportion of holdout samples that both individual SVMs and URSA predicted accurately. The light gray bar represents the proportion of holdout samples that only URSA predicted accurately; the dark gray bar represents the proportion of holdout samples that only individual SVMs predicted accurately. The black bar is the proportion of samples that neither URSA nor the individual SVMs predicted accurately. In other words, URSA accurately predicted

more than 94% of the holdout samples (white + light gray), whereas the individual SVM predicted 86% correctly (white + dark gray). Neither URSA nor SVM were able to accurately predict the remaining 3% of holdout samples. As shown in Figure 2.4a, the number of samples correctly predicted only by URSA increases with higher probability cut-offs and so highlights URSA's ability to make accurate predictions with meaningful probability values.

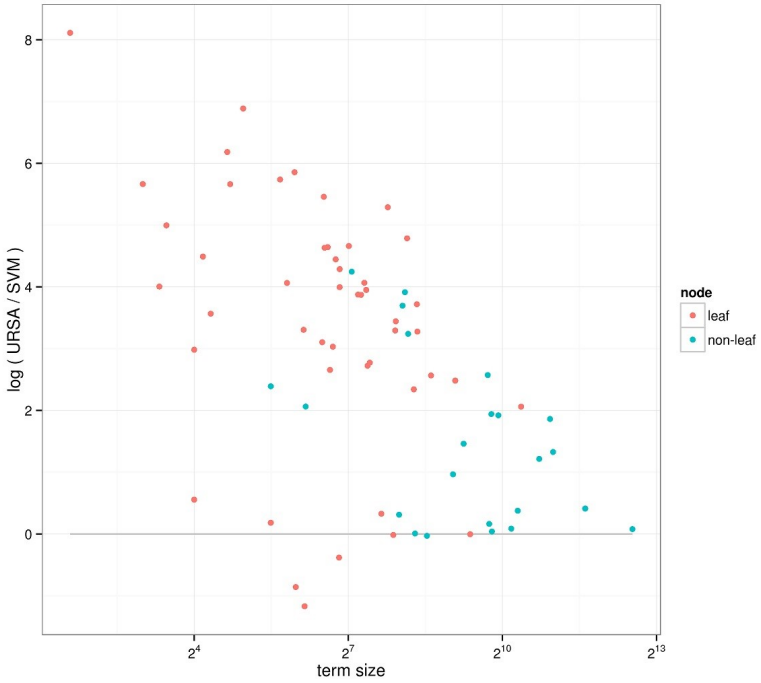


Figure 2.5: URSA's improvements of ranking accuracy over individual SVMs is greater for specific leaf nodes/terms in the tissue ontology. Each point on the scatter plot represents log-fold improvement for one unique cell-type. Points above the grey horizontal line correspond to improvements by our method. The greater improvement for small, leaf nodes demonstrates the need of our ontology-ware aggregation of leaf and non-leaf classifiers to detect subtle tissue/cell-type signals.

In addition to the overall performance evaluation, it is important to consider how annotation accuracy depends on the number of expression profiles available for training for each tissue term (namely 'term size'). Term size also serves as an appropriate estimation of the term's specificity in the tissue ontology, as sample annotations were propagated based on the same ontology. URSA's ontology-aware Bayesian frame-

work aggregates multiple individual classifiers so that classifiers for large terms (such as blood) could help classify related specific small terms (such as T-cell acute lymphoblastic leukemia cell). Using area under the precision-recall curve (AUPRC), we compare the entire ranking accuracy of URSA and the individual SVM classifiers across tissues/cell-types. URSA provides increased performance for 65 of the 71 tissue terms spanning both large general terms (such as B-lymphocyte > 0.98, breast > 0.89 and lung > 0.95) and small specific terms (such as T-cell acute lymphoblastic leukemia cell $= 1$, HeLa cell > 0.91 and bronchial epithelial cell > 0.83) (Figure 2.4b). Decreased performance for a few terms could be attributed to the incompleteness of the tissue ontology (e.g. the missing parental relationship between hepatocyte and hepatoma cell). URSA's improvements over individual SVMs are greater for leaf nodes than non-leaf nodes (Figure 2.5). The observed inverse relationship and larger improvements for leaf nodes than for non-leaf nodes highlights the need for URSA especially for the specific terms where individual classifiers often perform poorly due to the lack of training data. Thus, although the number of training samples affects the quality of individual models, our results show that exploiting the known cell-type associations enables URSA to be reasonably immune to this effect.

We examine the performance of the leaf nodes/terms in URSA's subgraph. URSA provides increased performance compared to independent SVMs for 43 of the 48 leaf terms including cell lines (such as HeLa cell > 0.91, T-47D cell $= 1$, and MCF-7 cell > 0.85), cancer types (such as cervical carcinoma cell > 0.98, chronic lymphocytic leukemia cell > 0.87, and renal cell carcinoma cell > 0.83), and stem cells (such as embryonic stem cell line > 0.73 and mesenchymal stem cell > 0.60). This greater improvement for small and specific terms is common across a wide range of tissues and thus highlights the need for our method especially for specific leaf terms (Figure 2.5). In essence, the additional intermediate (i.e. non-leaf) classifiers help improve the predictions for specific leaf nodes/terms over individual SVMs. It is worth emphasizing

that our approach is completely oblivious of known biomarker genes, but only relies on known ontological tissue relationships to delineate specific signals unique to each tissue/cell-type. It is not surprising that 48 of the 71 terms we evaluated are leaf nodes because the comparisons between URSA and individual SVMs were made for only terms with direct sample annotations. Also, missing associations between terms contribute to having related terms end up as leaf nodes. For example, both breast epithelial cell and MCF-10A cell are leaf nodes because the current tissue ontology is missing the parent-child relationship between breast epithelial cell and mammary epithelial cell. Often branch length is used to estimate the specificity of a term in a hierarchy. Unfortunately, branch length is not a good estimate of specificity in the tissue ontology used in URSA. For example, branch length for embryonic stem cell line is 4 while that for BJ cell is 8.

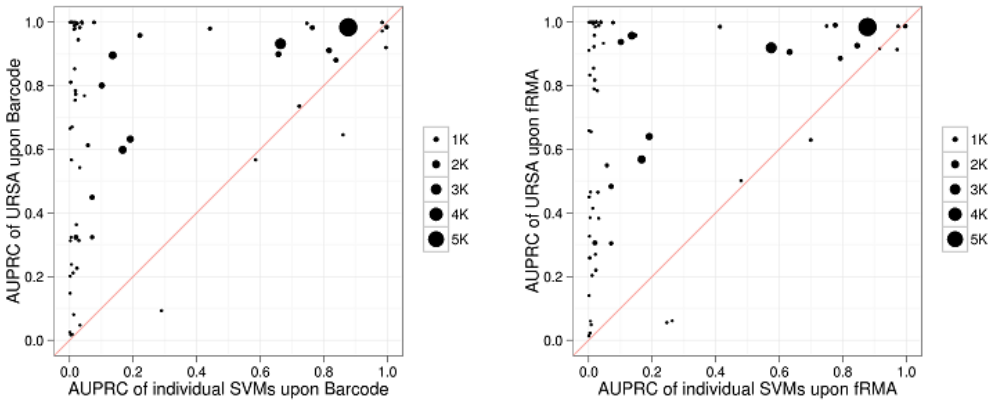


Figure 2.6: Scatter plot comparison between URSA and individual SVMs with different preprocessing algorithms. Each point represents a unique cell-type, and the size of the point represents the number of samples curated to that particular cell-type. Points above the diagonal correspond to improvements by our method. URSA consistently outperforms individual (i.e. independent) SVMs across many cell-types independent of term size and preprocessing methods.

Even without the use of the tissue ontology, independent SVMs perform reasonably well for easy problems such as discriminating blood samples, and so the improvement of our approach is relatively small (AUPRC of 0.9072 for individual SVMs versus AUPRC of 0.9823 for URSA). However, independent SVMs are unable to effectively discriminate more specific cell-type samples such as T-cell acute lymphoblastic leukemia cell samples (SVM AUPRC 0.0034), whereas our ontology-aware approach accurately classifies holdout samples of this specific blood cancer subtype (URSA AUPRC 1.0). This improvement of URSA can be attributed to the effective incorporation of the ontological complexity (Figure 2.2). Notice this improvement also holds true across a wide range of non-blood cell-types such as prostate gland (0.0317 vs. 0.9906), bronchial epithelial cell (0.0174 vs. 0.8333) and mesenchymal stem cell (MSC) (0.0017 vs. 0.6093) (Figure 2.4b). The fact that these signals were learned in a completely data-driven approach not from known biomarkers indicates that our method can provide a data-driven estimation of specific blood (and non-blood) cell-type signals.

2.3.2 URSA's performance is robust to expression data preprocessing

Data preprocessing and normalization can have a significant impact on downstream analysis, including prediction of tissues/cell-type signals [151]. MAS5.0 and fRMA are the two most well-known algorithms for preprocessing single arrays [48, 91]. Additionally, the barcode preprocessing algorithm was shown to accurately estimate whether a gene is expressed in a given microarray experiment and in specific tissues [92, 151].

We test the robustness of URSA's ranking accuracy to different preprocessing methods. Our first evaluation (using AUPRC) shows that URSA improves performance over individual SVMs across all three data processing methods: MAS5.0 (Figure 2.4b), fRMA and Barcode (Figure 2.6). Next, we compare URSA with a NN

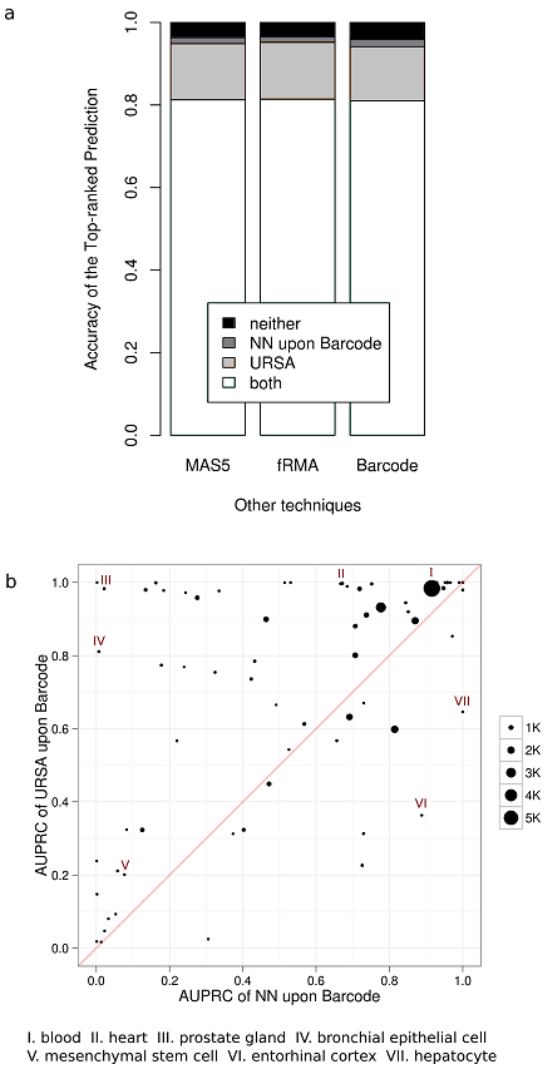


Figure 2.7: URSA's performance is robust across different preprocessing techniques. (a) We compare the accuracy of the most probable estimation from NN upon Barcode and URSA upon different preprocessing techniques. URSA outperforms the competing method across different preprocessing techniques. (b) Scatter plot comparison between URSA and NN upon Barcode preprocessing. Each point represents a unique tissue or cell-type, and the size of the point represents the number of samples curated to that particular tissue or cell-type. Points above the diagonal correspond to improvements by our method.

classifier after barcode processing, which is, to our knowledge, the only previous approach shown to predict cell-type [151]. It is important to note that the overall accuracy of the NN classifier relies on the accuracy of the barcode preprocessing algorithm. URSA correctly annotates $\sim 95\%$ of the test samples independent of the preprocessing algorithm used, with > 650 samples being correctly predicted exclusively using URSA (Figure 2.7a). Furthermore, our method returns better ranking accuracy for at least 50 of the 71 tissues/cell-types than the NN classifier (Figure 2.7b and 2.8). Again, the performance improvements appear to be robust to term size.

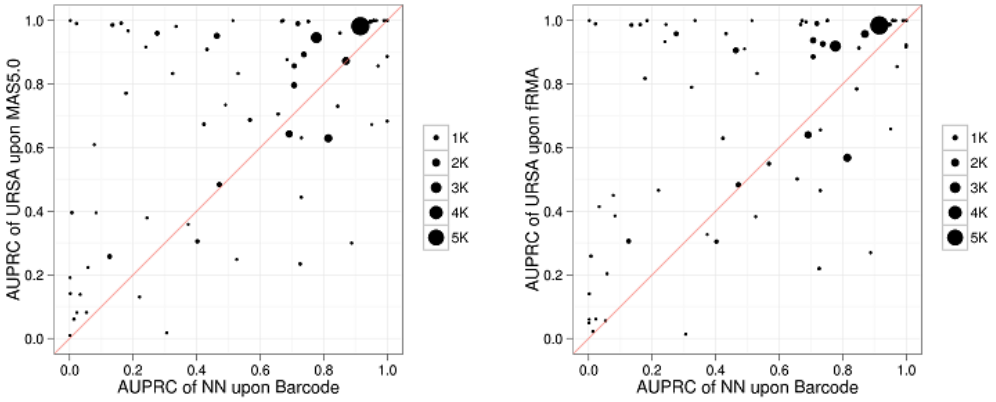


Figure 2.8: Scatter plot comparison between NN upon Barcode and URSA with different preprocessing algorithms. Each point represents a unique cell-type, and the size of the point represents the number of samples curated to that particular cell-type. Points above the diagonal correspond to improvements by our method. URSA consistently outperforms NN upon Barcode across many cell-types independent of term size and preprocessing methods.

These analyses show that URSA can adapt to both generic (e.g. MAS5.0) and specific (e.g. Barcode) preprocessing methods to discover tissue/cell-type-specific information in genome-scale experiments. Moreover, robustness to preprocessing suggests that URSA is modeling biological signals rather than systematic biases or data processing artifacts present in these large compendia. We focus our remaining analy-

ses using the most commonly used MAS5.0, chosen for its simplicity and application to many array platforms.

2.3.3 URSA is precise for experiments from other expression platforms

URSA is trained using data from the most popular gene expression microarray platform HG-U133 Plus 2.0 (hgu133plus2) with $\sim$ 70000 samples (from 2500 datasets/series) in GEO. We have shown that URSA performs well for samples from this platform, but there exist many other expression datasets that use other platforms, with new ones emerging continuously. The Affymetrix Human Genome U133A (hgu133a), for example, is arguably the second most common microarray platform, associated with $\sim$ 1000 studies in GEO. Other genome-wide array platforms such as HG-U95Av2 (hgu95av2) and HuGene 1.0 ST (hugene1st) have been used for their focused gene coverage. As hundreds of such platforms have been used for human gene expression measurements, re-training classifiers for each platform is impractical. Instead, the challenge is to overcome technical differences across platforms and predict tissue/cell-type signals in a platform-independent manner.

We test URSA's potential to measure the tissue-specific signatures in profiles from other array-based platforms without re-training its parameters. For this, we quantile-transform input data from cross-platform samples and filter final predictions by using a permutation test (see Methods). To evaluate these predictions in a manner that best emulates an end-user's experience, we conduct a double-blind literature study on *original* and *random* annotations. The evaluation shows that the majority of URSA's predicted annotations are both precise and consistent regardless of the microarray platform (Figure 2.9). Despite missing expression values for $>$ 10000 genes (due to limited gene coverage), our method is still able to provide high-quality annotations even for hgu95av2 samples. These consistent trends illustrate URSA's potential to

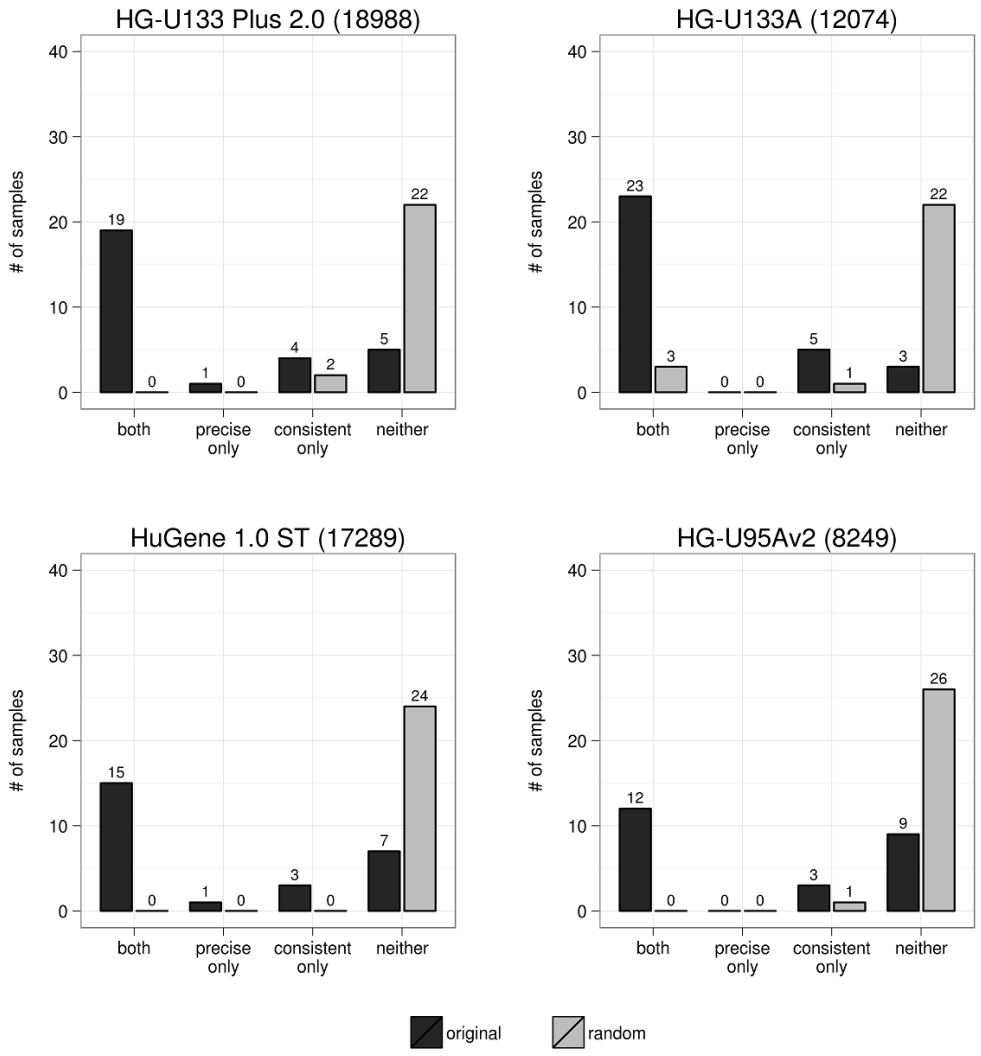


Figure 2.9: Consistent and precise original annotations for different array-based platforms. Original annotations were evaluated based on their associated sample descriptions. Random annotations were made on real expression data but evaluated based on random sample descriptions. The number of shared genes is denoted in parentheses. See Methods for more details.

detect cell-type-specific signals across microarray platforms rising above technical biases and even substantial gene coverage differences.

Next-generation sequencing is another rapidly growing technology for transcriptome profiling. A sample annotation method that can be applied to this burgeoning technology is also of great interest, and yet the current number of available tissue/cell-type-specific experiments limits the prospect of effectively training classifiers specifically for RNA-seq experiments. To address this problem, we test URSA's ability to detect tissue-specific signatures in RNA-seq experiments using the model trained on microarray data. At the outset, this is a challenging task due to the substantial technical differences between microarrays and RNA-seq. We challenge URSA to annotate RNA-seq experiments in the Illumina Bodymap 2.0 reference dataset (GSE30611), which consists of a diverse set of samples from 16 different tissues, generated with both single-end and pair-end sequencing methods. URSA correctly predicts the tissue of origin for all single-end and pair-end samples, except for adrenal gland and adipose tissue samples (Figure 2.10). For adrenal gland, URSA ranked adrenal gland as the second most significant tissue signal for adrenal gland samples (and not for any of the other tissue types such as kidney or thyroid gland). Although URSA can eventually be re-trained to better fit growing next-generation sequencing data, its robustness across platforms and technologies demonstrates URSA's promise to remain applicable and relevant to emerging experimental approaches and data processing methods.

2.3.4 URSA's tissue and cell-type-specific models are biologically interpretable

With accurate models constructed from $>$ 14000 diverse samples representing over 244 tissue/cell-type terms, URSA's discriminative features (i.e. genes) could paint a molecular portrait of tissue/cell-type-specific gene expression. To test this hypothesis, we use the PAGE algorithm [64] to examine the Gene Ontology (GO) biological

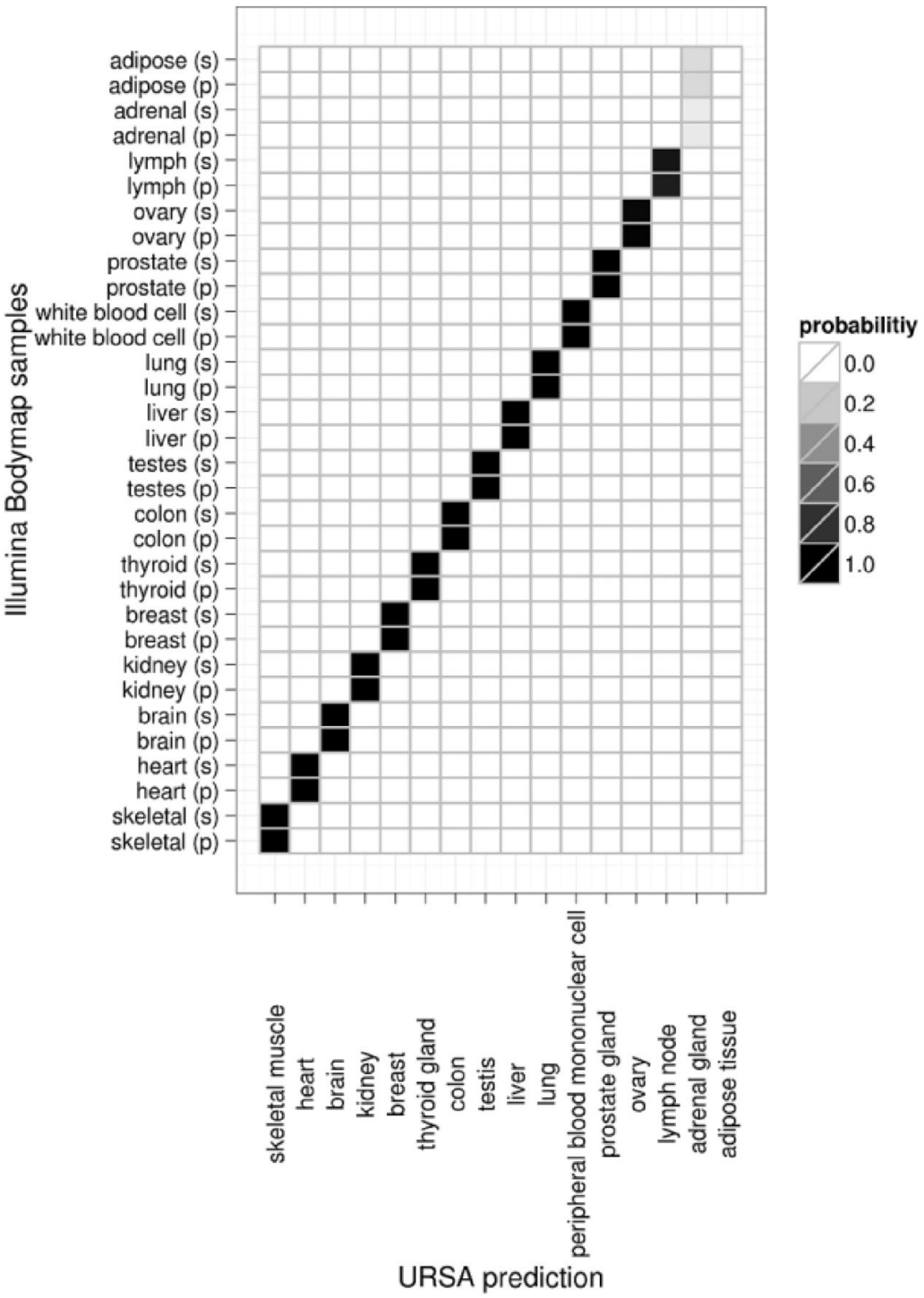


Figure 2.10: Accurate prediction of tissue of origin for RNA-seq samples. Heatmap of URSA's estimated tissue probabilities of 32 RNA-seq experiments (16 different tissues) in the Illumina Bodymap dataset. The rows are the individual samples, either single-end (s) or pair-end (p), and the columns are the estimated cell-types.

processes [4] represented in each tissue/cell-type model. In effect, the analysis summarizes the gene weights and offers a rich description of the models. Testing all 244 cell-type models, we find that many of the processes enriched among most informative genes in these models appear to be the relevant cell-type-specific GO terms. For example, the top GO term for the B-lymphocyte model is B cell activation (adjusted P < 0.001), whereas the top GO term for the B-cell lymphoma cell model is regulation of inflammatory response (adjusted P < 0.001). GO terms astrocyte differentiation, regulation of synaptic transmission and behavior are enriched in the brain model (adjusted P < 0.001).

Certain associations are not necessarily obvious. The top GO terms in the MSC model include mesenchymal-specific developmental processes such as skeletal system development, cartilage condensation and muscle organ morphogenesis. The enrichment of glycosaminoglycan biosynthetic process in the MSC model has some support in that glycosaminoglycans regulate osteoblast differentiation of bone marrow-derived human MSCs and chondrogenesis in mouse MSCs [63, 90]. The top specific GO terms in the embryonic stem cell (ESC) model include calcium-dependent cellcell adhesion, positive regulation of Wnt receptor signaling pathway and glutamine family amino acid metabolic process. During mouse embryogenesis, inner mass formation and cell surface polarization is regulated by the calcium-dependent cellcell adhesion system [124]. Highly conserved Wnt family proteins play a key role in embryogenesis and oncogenesis, but moreover the positive regulation (i.e. activation) of Wnt signaling maintains the pluripotency in human ESCs [85, 103, 118]. L-glutamine is needed for the culture and maintenance of human ESCs and is shown to inhibit mouse embryogenesis in high concentrations [2, 60, 97]. The enrichment of these non-trivial and specific biological processes demonstrates the expressive (and accurate) interpretation of URSA's predictions.

Figure 2.11: Tissue-specific biological processes enriched in URSA's skeletal muscle and heart models. Barplot of enrichment z-scores of top GO terms in the two models are shown. Both skeletal muscle and heart are primarily populated by muscle cells; yet, the heart tissue model selects genes specifically involved in cardiac muscle processes.

Based on the enriched biological processes (i.e. GO terms), we examine whether the models are specific enough to distinguish even closely related cell-types such as skeletal muscle cells and heart cells (Figure 2.11). Skeletal muscle and heart are among the most studied human tissues, and thus are appropriate examples to test the specificity of our models, which are based solely on genome-wide expression experiments. Both skeletal muscle and heart are comprised of muscle cells, and so one might expect that the top GO terms for both tissue models would be general muscle-related GO terms such as actin-mediated cell contraction. Instead, we find that although all top enriched processes for skeletal muscle are general muscle GO terms as expected, the top processes for heart (e.g. ventricular cardiac muscle tissue development and heart contraction) are specific to heart cells (Figure 2.11). Thus, without prior knowledge of tissue and cell-type-specific genes, URSA's models identify genes involved in corresponding cell-type-specific biological processes. This approach

could be extended for understanding poorly characterized cell-types including specific cancer subtypes. Our analysis altogether provides biological intuition and credence to the basis for URSA's tissue and cell-type annotations.

2.4 Discussion

In multicellular organisms, integrative analysis leveraging large gene expression compendia requires accurate annotations of samples to their tissue and cell-type of origin. In this article, we present a scalable computational method URSA that predicts tissue/cell-type signals in expression profiles across platforms and technologies. Key to its performance is the incorporation of the tissue ontology. Much of URSA's improved performance can be attributed to the construction of more than one hundred additional intermediate (i.e. non-leaf) classifiers, which are then integrated using a Bayesian framework.

URSA can be used to automatically annotate samples in public gene expression repositories where most samples are currently lacking tissue/cell-type-specific information. Researchers can discover specific signals in their own samples via our interactive interface at `ursa.princeton.edu`. Others interested in integrative studies can download the URSA C++ software and annotate samples on a large scale.

Despite URSA's current applicability to a wide variety of tissues/cell-types, its predictions can be further improved as the ontology used for integration adds additional terms and associations. For example, immunologists may be interested in the signal of specific T-lymphocytes such as CD4+ T cells, Th17 cells, germinal B cells, and so forth. Unfortunately, the current BRENDA ontology (which was used as a controlled vocabulary and the ontology structure of our method) does not include such terms. Nonetheless, URSA's ability to delineate tissue/cell-type signals without known biomarker genes makes it naturally extendable to such specific cell-types as

the BRENDA ontology is extended with more terms and associations. We plan to regularly maintain and update the software with new tissue and cell-type annotations and the latest version of the BRENDA ontology.

Both the strength and the limitation of our method across platforms and technologies depend on the amount of tissue signal in the gene order and the number of missing values. For a given gene expression profile from a different platform, quantile transformation is applied to compute hg133plus2-like expression values. In consequence, our method is robust to different normalization techniques used because only the information of relative gene abundance is transferred. However, specific signals associated with the particular gene expression value may be lost, and properly incorporating such signals may provide greater prediction accuracy. Furthermore, expression values for genes not measured in hg133plus2 could affect the accuracy of our method, although simple mean imputation seems to alleviate that effect.

URSA's tissue and cell-type-specific models provide a biological interpretation of its predictions. As such, URSA could potentially be used to test and identify possible sample contaminations, resolve cancer samples of unknown primary origin and perhaps provide insight into the molecular basis of poorly characterized clinical subtypes.

Chapter 3

Genome-wide characterization of the human disease landscape

This chapter describes work done with critical support and comments from Arjun Krishnan. Literature curation and feedback were provided by Chandra Theesfeld, Christie Chang, Rose Oughtred, and Jennifer Rust.

3.1 Introduction

Gene expression profiling has been used for two decades now to capture the genome-wide dysregulation in a number of human diseases. A typical gene-expression study for a particular disease is carried out first by profiling a group of disease samples and a comparable group of normal control samples, and then contrasting disease samples against controls. The resulting differential mRNA abundance of thousands of genes is valuable in capturing the genome-wide perturbations of genes and pathways that underlie the disease of interest. However, complex diseases fall along a continuous landscape of molecular phenotypes, sharing with each other several of their underlying genetic and functional changes. Therefore, from the myriad of observed expression

changes, it is impossible to tease apart those unique to a disease when the disease gene expression is analyzed in isolation.

Fortunately, hundreds of disease gene expression datasets created in the last two decades have been deposited in public repositories like NCBI GEO [7]. Integration of these individual studies offers a promising path towards better understanding the characteristics of multiple human diseases [1, 120]. In fact, several efforts have been made to integrate multiple studies by scaling-up differential expression analysis (comparing disease to healthy samples) to quantify the abnormalities in multiple diseases [47, 109, 132]. However, by comparing diseases only post-analysis, such approaches do not explicitly address disease-disease relationships, thereby failing to identify features distinctive to each disease. For example, multiple diseases are related to the immune system, but the specific immune component in each disease is poorly understood [18]. Therefore, a unified framework is needed to tackle the challenge of understanding the functional and anatomical context of each disease in the context of all other related diseases. This framework needs to be comprehensive, covering a large number of diseases, and data-driven, taking advantage of thousands of clinical gene expression datasets, in order to uncover subtle differences between similar diseases and highlight identifiable aspects of rare diseases.

Here we present URSA$^{\mathrm{HD}}$ (Unveiling RNA Sample Annotation for Human Diseases), a systematic framework that mines hundreds of individual clinical datasets to explicitly compute the distinctive characteristics of 309 human diseases, including 20 rare diseases. Leveraging the hierarchical relationships among human diseases and thousands of disease-specific gene expression datasets, URSA$^{\mathrm{HD}}$ builds individual disease-specific models and integrates them in a probabilistic framework to provide hierarchically consistent estimates of disease signals. URSA$^{\mathrm{HD}}$ can then accurately characterize the disease signals in any gene-expression sample, providing a predictive probability of this sample being associated with each disease. The rigorous processing

and evaluation settings in URSA^{HD} allow it to overcome potential patient, dataset and profiling-technology biases, and learn discerning models (i.e. genome-wide weight vector) even for rare diseases with limited numbers of samples. These disease-specific models effectively characterize molecular signals specific to each disease in the context of not only associated normal tissues and anatomical regions but also other human diseases. This completely data-driven approach does not rely on literature-based disease-gene associations and is solely based on thousands of gene expression experiments of clinical samples, both normal and disease.

In the rest of the paper, we describe the probabilistic framework behind URSA^{HD} and systematically show that URSA^{HD} outperforms other approaches of using individual disease genes or the typical normal/disease differential expression model in quantifying disease signals. We reveal how, in addition to accurate sample predictions, URSA^{HD} provides interpretable, molecular models in terms of discerning biological processes and associated tissues. Finally, we tackle two central problems for drug development: tracking therapeutic effect by expression profiling and associating rare diseases to their nearest well-studied human diseases for the purpose of drug repositioning. We have implemented URSA^{HD} in a publicly available web-server at `ursahd.princeton.edu`, where biomedical researchers can submit their gene expression data to obtain data-driven quantification of disease signals.

3.2 Methods

3.2.1 Documented disease and anatomical genes

Gene2mesh uses curated MeSH annotations of PubMed articles to find genes that are statistically significantly studied with a particular MeSH term `http://gene2mesh.ncibi.org`. We used MeSH terms under the Anatomy MeSH tree structure as anatomical MeSH terms, and terms under the Diseases MeSH tree structure as disease

MeSH terms. 396 anatomical MeSH terms and 509 disease MeSH terms had at least 10 associated genes. Disease and anatomical genes were downloaded from gene2mesh on May 14 2014.

3.2.2 PubMed article gene annotations

Human gene annotations to PubMed articles were downloaded from the National Center for Biotechnology Information (NCBI) on Oct 31 2014 [106, 134]. The number of unique PubMed article associations for each gene is used as a proxy to estimate how well the gene is studied and characterized. $436,945$ PubMed articles had at least one gene annotation. $33,454$ unique human genes were annotated to at least one PubMed article. The most studied gene was tumor protein p53 (TP53) with $6,592$ associated PubMed articles.

3.2.3 Genome-wide expression data processing

The Human Genome U133 Plus 2.0 Array (hgu133plus2) raw CEL files were downloaded from Gene Expression Omnibus (GEO) [7]. Probes were mapped to Entrez GeneIDs using the BrainArray Custom CDF ver. 18. MAS5.0 with default parameters and subroutines were used for normalization, and then log-transformed [25, 48]. Therapeutic treatment datasets (GEO: GSE10281, GEO: GSE16879, GEO: GSE28844, GEO: GSE53552) used only for analysis were also pre-processed and normalized using the same pipeline [3, 22, 114, 143]. Clinical information (patient id, diagnosis, treatment type, response type) were from the author-provided sample description in GEO.

3.2.4 Gold standard construction by manual sample annotation

High-quality sample annotations are needed to accurately compare and evaluate the performance of different approaches to estimate disease signals in genome-wide experiments. We manually annotated $8,359$ microarray experiments of clinical patient samples from 139 datasets from the hgu133plus2 platform. Available sample descriptions and other textual information in GEO and their associated publications were used for this curation step. Disease terms in the MeSH disease category were used as the controlled vocabulary. Normal or control (non-disease) samples were annotated as 'other.' For example, 'unaffected sites' (GEO: GSM404013) and 'surrounding noncancerous cells' (GEO: GSM490997) were annotated as 'other.' A total of 1996 samples were annotated as 'other.' Reference, xenograft, cultured, or cell-line samples were excluded to avoid learning extraneous signals. The manual annotations for 116 disease terms were then propagated based on the MeSH disease hierarchy, resulting the coverage of 335 disease terms.

3.2.5 Therapeutic chemical disease associations

Chemical disease associations were downloaded from the Comparative Toxicogenomics Database (CTD) on Mar 2 2015 [26]. CTD contains both curated and inferred chemical-disease interactions. Only curated associations with direct therapeutic evidence were used, a total of 27571 associations with 5852 unique chemicals to 2290 diseases. Hypertension (MESH: D006973) had the most associated therapeutic chemicals ($n = 343$).

List of rare diseases were downloaded from OrphaData V 0.9 on Nov 17 2014 at `http://www.orphadata.org` [5]. 20 of our models were for rare diseases. Out of the 20 rare disease models, 6 rare diseases had no documented therapeutic chem-

ical associations: Arrhythmogenic Right Ventricular Dysplasia (MESH: D019571), Enteropathy-Associated T-Cell Lymphoma (MESH: D058527), Collagenous Colitis (MESH: D046729), Limb-Girdle Muscular Dystrophies (MESH: D049288), Primary Cutaneous Anaplastic Large Cell Lymphoma (MESH: D054446), Extranodal NK-T-Cell Lymphoma (MESH: D054391).

3.2.6 Training and testing setup

Method evaluations are often done with a random holdout. However, genome-wide experiments are prone to laboratory and dataset biases, and so a simple random holdout might overestimate the performance of these methods [79, 113]. To control for this bias, the series/datasets of the manually annotated samples were randomly partitioned into training and testing for each term as done previously [77]. Only disease terms with at least two positive and negative samples in both the training set and the testing set were evaluated.

3.2.7 Individual disease prediction methods

Documented gene-based prediction method

A common method to predict disease signals is based on the expression of a documented disease gene. The documented gene-based method picks a documented disease gene (from gene2mesh) that best distinguishes the disease samples (positives) from their normal counterparts (negatives) based on its Area-Under-the-Precision-Recall-Curve (AUPRC) ranking accuracy. Positive samples are only from direct sample annotations, and negative samples are *other* (i.e. control) samples in those datasets with positive samples. Datasets with only disease samples or other control samples weren't included in training. This method represents a typical single gene-based approach that relies only on documented disease genes. This approach is similar to how

the sex of the sample can be identified based on the expression of Y-chromosome genes.

Normality-based prediction method

Genome-wide differential analysis between normal and clinical samples is a common approach for understanding the molecular abnormality and mechanistic dysregulations in the manifestation of that single disease (Ritchie et al., 2015). The normality-based prediction method mimics this typical normal vs. disease differential setup and is a Support-Vector-Machine (SVM) trained on disease samples (as positives) and normal samples (as negatives) (Burges, 1998). Likewise positive samples are only from direct sample annotations, and negative samples are other control samples in those datasets with positive samples. Notice that both normal and disease samples are needed in at least one dataset to train the disease model for this method.

3.2.8 URSA$^{\mathbf{HD}}$'s unified disease prediction method

We set URSA$^{\mathrm{HD}}$'s unified framework for disease prediction as a hierarchical multi-label classification problem [10, 77]. Each individual disease classifier characterizes the distinctive features of those expression profiles of the disease compared to that of all other control samples and unrelated disease samples. The Bayesian network then models the probabilistic relationship between those classifiers to calibrate the individual predictions and thus provides an interpretable list of disease (both cancerous and non-cancerous) predictions for a given clinical gene expression profile.

Hierarchy-aware characterization of individual diseases

The choice of positive and negative samples explicitly defines the learning criteria for the classifier, and so systematically setting the context of a particular disease is crucial for a unified framework. We use the MeSH disease hierarchy to set this context

and capture the distinctive characteristics of a particular disease. Samples annotated directly to the disease term or any of its descendant terms (i.e. more specific diseases) are considered positive; samples annotated to only its ancestor terms are excluded from training; and the remaining samples including those annotated to sibling terms are considered negative. The control samples (those annotated as emphother) are negative for all disease terms. A dataset of gene expression profiles across 65 healthy tissues (GEO: GSE3526, 353 samples) was used to ensure a comprehensive coverage over various tissue-types [111]. As a consequence, any tissue-specific signal in URSA$^{\text{HD}}$'s disease models represents the over-expression of those tissue-specific genes in the disease, even over its normal (or healthy) counterpart. Related disease terms share positive samples (such as between Adenocarcinoma (MESH: D000230) and Renal Cell Carcinoma (MESH: D002292)), but more specific terms (i.e. renal cell carcinoma) have a more exclusive set of positive samples. Given l pairs (i.e. samples) of expression data x_i and its label y_i, we use the L2 linear SVM (with cost parameter c = 20) [52, 56]:

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{l} \max(1 - y_i w^T x_i, 0)^2 \tag{3.1}$$

Notice that this labeling scheme defines the learning criteria for even general terms (such as Neoplasm (MESH: D009369)) with no direct sample annotations - including the most general 'Disease' term. These two hundred additional disease SVM models are then incorporated in the Bayesian network that combines these distinctive characteristics in a unified probabilistic framework.

Hierarchy-aware probabilistic aggregation of distinctive classification models

Each individual model is trained separately, and so the predictions - given an expression profile - aren't comparable without explicitly defining the relationships between

those predictions. We use the structure of the MeSH disease hierarchy and define these relationships in a Bayesian network [77]. We model each term's unthresholded SVM output as a noisy observation $\hat{y}_i$ of a latent binary event y_i representing the true label (i.e. disease) of a given sample. The edges from y to $\hat{y}$ establish conditional independence of an SVM prediction $\hat{y}_i$ to all other SVM predictions $\hat{y}_i$ $(i \neq j)$ given its true label y_i. This allows us to easily compute the likelihood:

$$\mathrm{P}(\hat{y}_1, \ldots, \hat{y}_N | y_1, \ldots, y_N) = \prod_{i=1}^{N} \mathrm{P}(\hat{y}_i | y_i) \tag{3.2}$$

The conditional probability tables $\mathrm{P}(\hat{y}_i|y_i)$ for each term represents the discriminative power of each term's SVM. Through 2-fold cross-validation, we estimate these conditional tables by counting the number of negative samples with smaller SVM outputs than that of a positive SVM output. Laplace smoothing is finally applied for robustness.

The parent-child conditional probability tables are defined similar to the original Bayesian correction method and so ensure that a label is true when any one of its children is true. When none of its children are true (including when it has no children), a constant prior of 0.1 is assigned. This allows us to compute the prior:

$$\mathrm{P}(y_1, \ldots, y_N) = \prod_{i=1}^{N} \mathrm{P}(y_i | \mathrm{ch}(y_i)) \tag{3.3}$$

where $\mathrm{ch}(y_i)$ is child labels of y_i.

We use the loopy belief propagation algorithm implemented in the SMILE library to infer the posterior probabilities $\mathrm{P}(y_i|\hat{y}_1, \ldots, \hat{y}_N)$ for each disease term [27]. These posterior probabilities are the estimated probabilities that our method uses to annotate gene expression samples.

3.2.9 TCGA mRNA-Seq sample prediction

URSA^{HD} disease models were trained on hgu133plus2 samples and so have not been specifically tuned for predicting sequence-based expression profiling experiments. In order to account for this difference, samples from sequence-based technologies were quantile transformed as done previously [77]. The approximate maximum expression value 15 was used to impute missing values in the quantile transformed sample. A permutation test was performed to filter insignificant predictions that might have arisen from technical biases. Only the non-imputed values were permuted to compute random predictions of the null distribution. This conditional permutation controls for imputation bias. Insignificant predictions were assigned a value of 0.

TCGA's RNASeq Version 2 IlluminaHiSeq normalized gene expression data (Data level 3) was downloaded on July 18 2014 [49, 81]. 15 different cancer-types were covered by both TCGA's RNASeq Version 2 and the current disease models at the time. Predictions were made for a total of 6172 RNASeq samples.

3.2.10 Inferred URSA^{HD} disease model and gene set associations

Each individual URSA^{HD} disease model is a hyperplane that best separates the positive and negative samples. This hyperplane is a high-dimensional vector with coefficients (or weights) for each dimension (or genes): $\vec{w} = \{w_1, w_2, \ldots, w_m\}$ where m is the number of genes covered by the gene expression profile assay. The PAGE enrichment algorithm that is based on the central limit theorem is used to calculate the association between a given gene set and a gene weight vector [64]. Given a disease model $\vec{w}_d$ and gene set S_t, the enrichment score z_{td} for term t and disease d is:

$$z_{td} = \frac{\bar{x}_{td} - \mu_d}{\sigma_d} \tag{3.4}$$

where

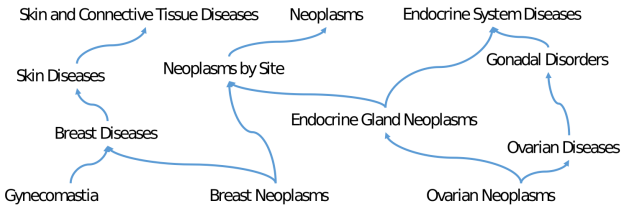
$$\bar{x}_{td} = \frac{\sum_{g \in S_t} w_g}{n} \tag{3.5}$$

where n is the size of S_t and μ_d and σ_d is the population mean and standard deviation of $\vec{w_d}$. This enrichment score estimates the statistical significance of the mean weight of genes associated to term t compared to the mean weights of random genes of the same size. All functional, disease, and anatomical associations are based on this enrichment score.

3.2.11 Drug repurposing evaluation based on disease model and disease gene set associations

Curated chemical disease associations from CTD were used to evaluate the utility of the disease model and its disease gene set associations for drug repurposing. Known multi-purpose uses of chemicals were used as the gold standard for evaluation. For a human disease M with an URSA$^{\text{HD}}$ model, the expected association score with disease S that shares a therapeutic chemical is compared with the expected association score with a random disease. This comparison is similar to the test of independence in probability: $\mathrm{P}(A|B) = \mathrm{P}(A)$. The first expectation conditions on the known chemical association with both disease M and S, and the second expectation is marginalized. If the first expectation is greater than second, then therapeutic chemicals for diseases with high association scores statistically will have a therapeutic effect on disease M as well. It is worth mentioning that our evaluation circumvents the need for 'negative' chemical disease associations - that the chemical has no therapeutic effects on the disease.

a



b

Figure 3.1: (a) MeSH disease sub-hierarchy for Breast Neoplasms. Such disease complexity must be accounted for accurate characterization of specific disease signals. (b) Word cloud of 116 disease terms covered in manual curation of 8359 gene expression profiles. Size of term corresponds to the number of profiles annotated to the term. Text color is set arbitrarily for visualization.

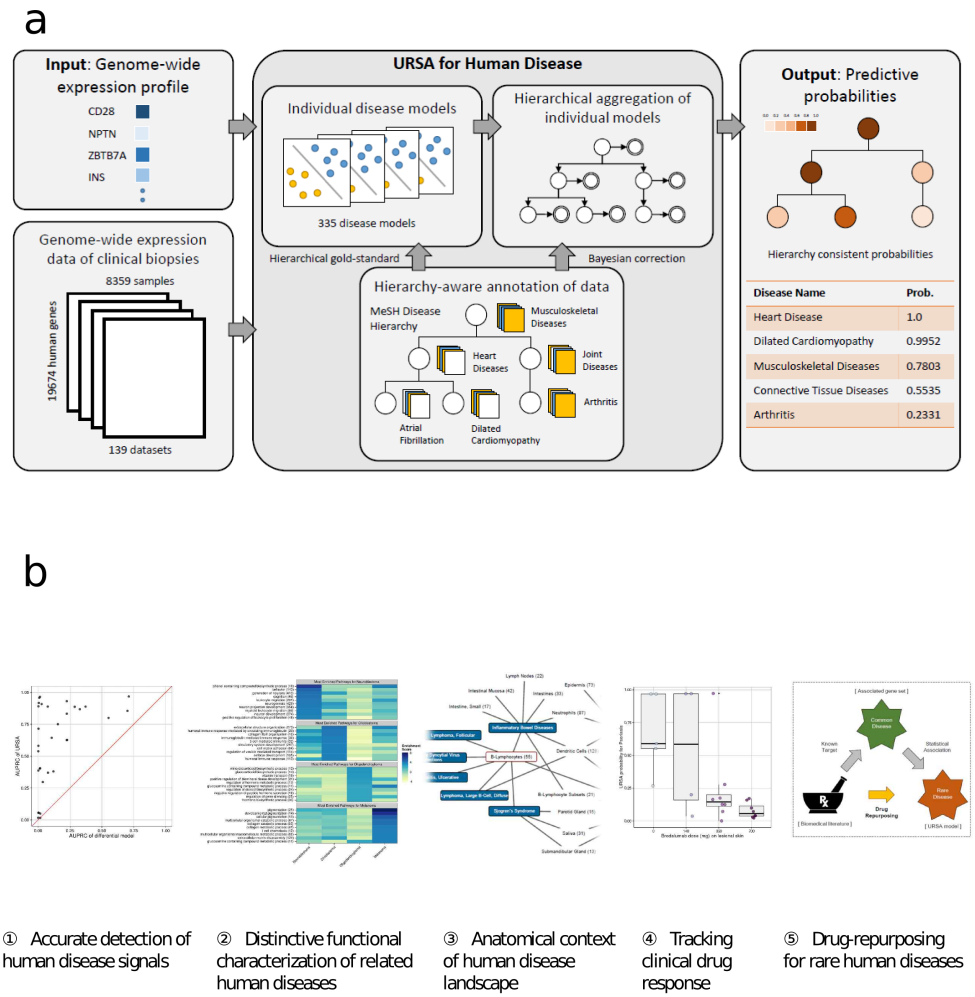


Figure 3.2: (a) URSA^{HD} provides hierarchy-consistent predictive probabilities for 335 disease terms respective of a given gene expression profile. URSA^{HD} integrates 8359 disease and normal clinical samples to quantify distinctive disease signals for 335 disease terms under the MeSH disease category. Hierarchy-aware annotation is applied to effectively characterize individual disease models, and these models are later aggregated into a unified Bayesian framework consistent with the known hierarchical relationships. Note that no feature selection method or known gene sets are used in our approach. (b) We demonstrate URSA^{HD}'s ability for accurate disease signal detection, specific functional and anatomical characterization of each individual disease, tracking therapeutic drug treatments from gene expression experiments, and repurposing known drugs for the treatment of rare human diseases.

3.3 Results

3.3.1 Hierarchy-aware characterization of the human diseases from clinical biopsies

Accurate characterization of diverse molecular pathologies underlying human complex diseases requires identification of signals that distinguish both the particular disease-state from its corresponding healthy condition and its context in the human disease landscape (Figure 3.1). To identify each human disease, we first model the distinctive, genome-wide features of each individual disease by contrasting expression profiles of clinical disease samples to that of all control (or healthy) samples and unrelated disease samples (Figure 3.2a). URSA$^{\mathrm{HD}}$ then integrates 355 individual disease models into a unified Bayesian network based on the structure of the MeSH hierarchy to provide hierarchically-consistent estimates of the specific disease signal (Figure 3.2a). This genome-wide characterization of the human disease landscape is achieved by leveraging large public data repositories. Most publicly available genome-wide datasets are associated with a single disease, sometimes including a paired control set of normal samples. In order to organize these datasets in a single compendium, we've manually annotated 8359 gene expression experiments across 136 clinical datasets to 116 MeSH disease terms which are organized within the MeSH hierarchy (Figure 3.2a, Figure 3.1b). These data-driven models implicitly identify and up-weight genes that are consistently expressed differently in positive profiles than in negative profiles, and at the same time, shrink weights of non-discriminative genes (see Methods). This specificity of weighted genes characterizes the distinctive genome-wide traits of each human disease without the use of any known disease-gene associations or prior feature selection methods.

3.3.2 URSA^{HD} accurately detects disease-state solely from gene expression profile

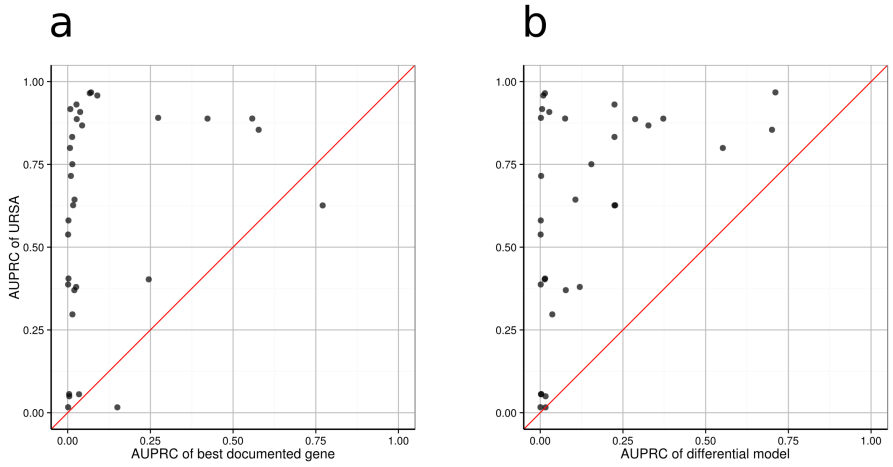


Figure 3.3: URSA^{HD} accurately detects human disease signals in gene expression profiles. (a) URSA^{HD} outperforms literature-documented disease gene method across multiple diseases. Scatterplot of AUPRC of URSA^{HD} (y-axis) and known disease gene method (x-axis). Each dot represents the comparative performance of a specific disease. Red line is the identity line, and so dots above the red line indicate diseases with greater performance. (b) URSA^{HD} outperforms typical normal/disease genome-wide differential expression approach for disease detection. Scatterplot of AUPRC of URSA^{HD} (y-axis) and typical differential approach (x-axis). Each dot represents the comparative performance of a specific disease. Red line is the identity line, and so dots above the red line indicate diseases with greater performance.

The disease-state of a given clinical sample is often inferred by the expression of single disease gene. While the expression of these genes hints at abnormal molecular changes of the underlying tissue/cell-type, many known disease genes aren't exclusive to individual diseases and thus limit the use of known genes to distinguish a particular disease from others. Tumor necrosis factor (TNF), for example, is the most common human disease gene, being documented in literature with 96 human diseases such as psoriasis, non-Hodgkin lymphoma, and obesity [13, 80, 110, 112]. In contrast, URSA^{HD} uses a data-driven genome-wide approach to generate disease

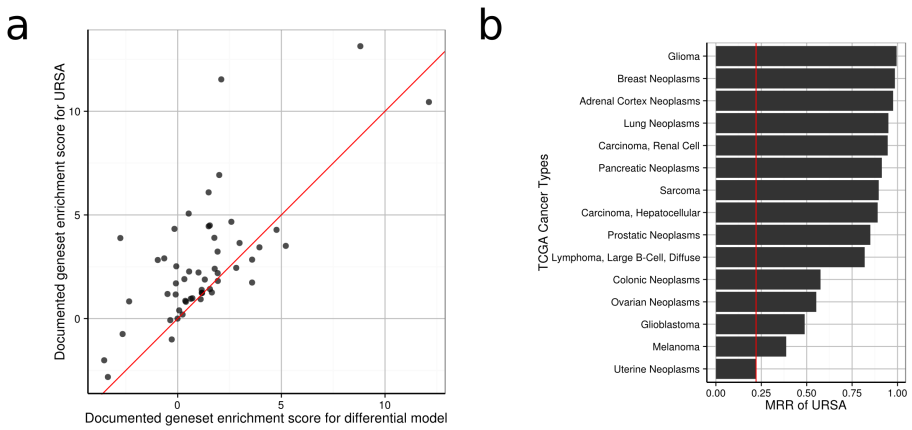


Figure 3.4: (a) Documented disease genes (i.e. gene2mesh genes) are more enriched in URSA^{HD} models over typical normal/disease differential models. Scatterplot comparison of documented disease gene set enrichment in URSA models (y-axis) and normal/disease differential models (x-axis). Red line is the identity line and so dots above the red line indicate diseases with greater enrichment score. (b) Without retraining, URSA^{HD} accurately predicts different cancer-type samples from TCGA's RNASeq collection. 6172 samples across 15 different cancer-types were predicted. Mean reciprocal rank of the correct prediction is shown for each cancer-type. Red line indicates performance of random prediction.

models (i.e. genome-wide weight vector) reflective of the molecular characteristics that is specific to each disease. As such, these models effectively differentiate among other human disease-states and healthy-states, outperforming the best documented (via gene2mesh) single genes for 30 of the 32 diseases with an independent holdout set (Figure 3.3a). For 75% of these diseases, URSA^{HD} models were over 10 fold more accurate than the best-performing documented disease gene (Figure 3.3a).

Many human disease studies have used gene expression profiles to systematically quantify genome-wide changes between healthy samples and disease samples. However, such definition of control (i.e. healthy samples) restricts our understanding to the abnormal changes brought by the disease and not its precise manifestation. Instead, URSA^{HD} takes the entire set of multiple healthy tissue samples (in addition to the corresponding healthy samples) and other disease samples to identify its context

in the human disease landscape (see Methods). Indeed URSA^{HD} outperformed the typical normal/disease differential analysis for all 32 diseases with an independent holdout set (Figure 3.3b). This performance difference appears to arise from URSA's ability to identify distinctive characteristics of each human disease. For example, both dilated cardiomyopathy models (AUPRC of URSA^{HD} = 0.9069, AUPRC of typical differential model = 0.0752) were enriched for heart-related Anatomical MeSH terms such as 'left ventricular hypertrophy', 'atrial fibrillation' and 'heart atria.' However, the 30 dilated cardiomyopathy genes documented in the literature were only enriched in URSA^{HD} model ($z = 4.326$) and not the typical normal/disease differential model ($z = -0.155$). Note that both approaches are data-driven and not based on documented disease genes. This lack of specificity underlies the limitation of the typical normal/disease differential approach, and such trend persists across other human diseases including rare diseases (Figure 3.4a). Without retraining the models, URSA^{HD}'s predictions were also consistent and accurate for TCGA's RNASeq samples further demonstrating the biological relevance of URSA^{HD}'s disease models, independent of profiling platform (Figure 3.3b, see Methods).

3.3.3 URSA^{HD}'s characterization of neuroblastoma and other diseases of ectodermal origin is distinct and specific

URSA^{HD}'s disease models identify the distinctive genome-wide characteristics of each human disease by up-weighting genes with peculiar expressions compared to other diseases and normal tissue samples (see Methods). Understanding the molecular basis of human diseases landscape from associated gene(s) is limited as shown by the significant but ubiquitous disease gene set similarities (Figure 3.5a). In contrast, our approach provides a data-driven functional perspective of human diseases,

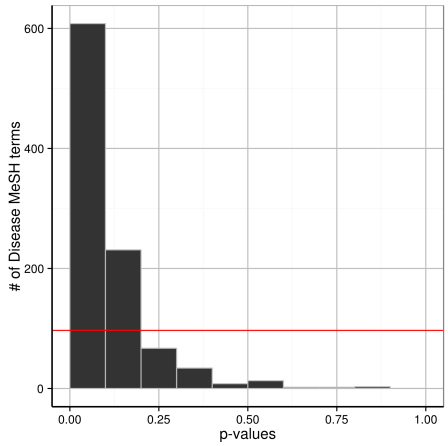


Figure 3.5: Skewed distribution of documented disease gene set overlap with documented neuroblastoma genes. Distribution of p-values of disease terms associated with at least 10 documented genes. Red line indicates uniform distribution. The amount of skewness could merely represent our comprehension of generic cancer-related pathways and relative lack of targeted functional understanding.

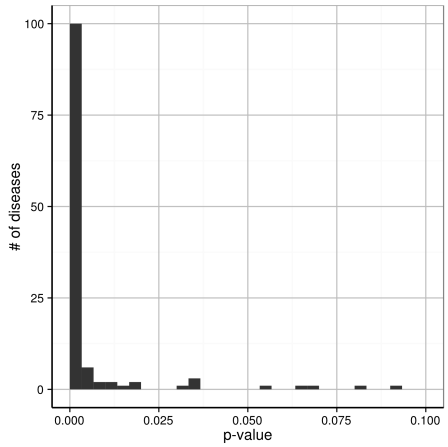


Figure 3.6: URSA$^{\mathrm{HD}}$'s top model genes are statistically understudied compared to documented disease genes. Wilcoxen rank-sum test for each human disease (with at least 10 documented genes) between the numbers of publications associated with each top model gene and the numbers of publications associated with each documented disease gene. Sample sizes were matched. Type 2 diabetes mellitus, breast cancer, Alzheimer disease, Rheumatoid arthritis were among the top human diseases with statistically understudied top model genes.

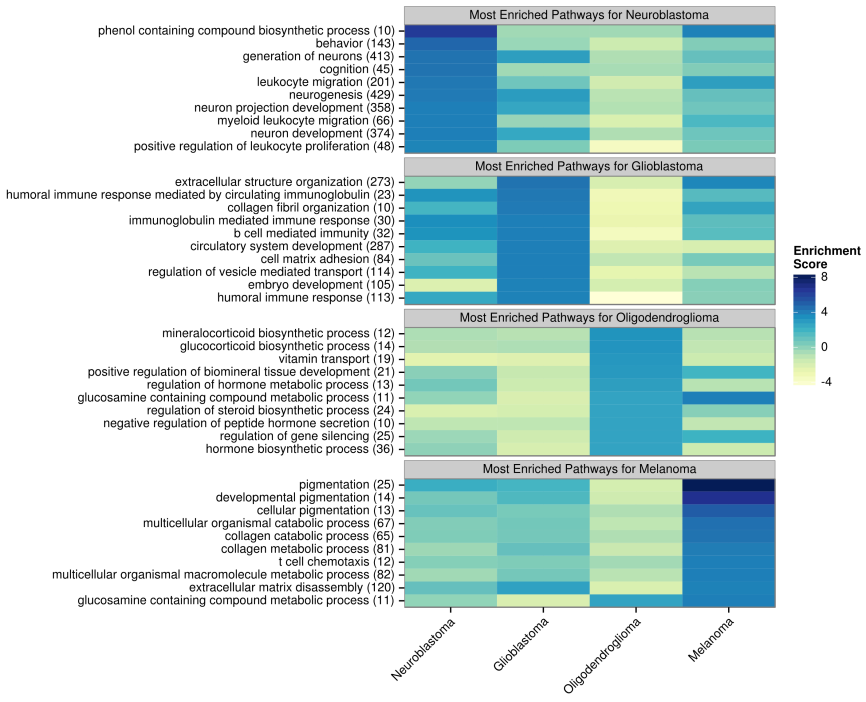


Figure 3.7: Distinctive functional characterization of neuroblastoma and other diseases of ectodermal origin. Each panel (i.e. heatmap) summarizes the top functional (i.e. GO biological process) enrichments for one (titled) of URSA^{HD}'s models: neuroblastoma, glioblastoma, oligodendroglioma, and melanoma. Each ectodermal disease is characterized by their distinctive functional associations.

especially in the context of closely-related diseases. Neuroblastoma, glioblastoma, oligodendroglioma and melanoma all originate from ectodermal tissues and exhibit similar characteristics of tumor-host interaction at the molecular-level [127]. Nonetheless, each ectodermal diseases exhibit unique signs and symptoms used for diagnosis, targeted treatment, and prognosis. URSA^{HD}'s disease models for the four ectodermal diseases were enriched with functional characteristics specific and consistent to known literature for each individual disease (Figure 3.7). URSA^{HD}'s neuroblastoma model was enriched with neuron development-related processes and leukocyte migration-related processes recapitulating its known neural crest-derived origins and lymphocytic infiltration [148, 74]. URSA^{HD}'s glioblastoma model was specifically enriched with distinctive pathways relevant to its strong dysregulation of circulat-

URSA Top 20 Genes	**Descriptions**	**Pubmed IDs**
PHOX2B	Major susceptibility factor (Orphanet)	15901893
LIN28B	Major susceptibility factor (Orphanet)	23042116
GATA3	Prognostic marker	25351211
PHOX2A	Biomarker	18949361;15901893;19212675
HAND2-ASI	Antisense RNA to HAND2	18171985
NPY	Biomarker	20676138;9802408
ISL1	Biomarker	23503646;23417100
NNAT	Prognostic marker	17762496
FABP6	Prognostic marker	16989664
STMN2	Biomarker	23333500
TH	Biomarker	12507966; 4399798
DLK1	Biomarker	3470797; 8095043; 15798773
CHRNA3	Biomarker	2336208; 9009220; 23417100
PRPH	Biomarker	6399022; 8381395
FAM163A (NDSP)	Biomarker	19671756
ARHGAP36	Lightly characterized protein	None found
LOC1005070194	Uncharacterized lncRNA	None found
MAB21L1	(Expressed in neural crest derivatives)	(10556287)
MAB21L2	(Expressed in neural crest derivatives)	(10495284, 10556287)

Table 3.1: Descriptions and literature evidence for URSAHD's top 20 Neuroblastoma genes. Information in parenthesis indicate indirect evidence.

ing immunoglobulin, extracellular matrix structure and angiogenesis to aggressively invade the brain parenchyma [149, 102, 36, 59]. Note that low enrichment only indicates the relative absence or involvement and not a complete lack of related gene expression. Interestingly, glucocorticoid metabolism-related biological processes were most enriched in URSA^{HD}'s oligodendroglioma models. Such association encourages further investigation of glucocorticoid metabolism and its role in oligodendroglioma, especially in the context of MYOC (myocilin, trabecular meshwork inducible glucocorticoid response or also known as TIGR) a known mediator of oligodendrocyte differentiation [70, 19].

The functional and anatomical contexts set by our disease models are primarily driven by these top model genes (see Methods). Of the top 20 up-weighted genes for URSA^{HD}'s neuroblastoma model, two (i.e. PHOX2B and LIN28B) are known sus-

ceptibility genes for neuroblastoma, 13 are highly expressed, serving as biomarkers for diagnosis and prognosis, and 3 are associated with embryonic and nervous system development (Table **??**). Such specificity warrants further investigation of the remaining two uncharacterized genes (ARHGAP36 and LOC100507194) that were also up-weighted in URSA^{HD}'s neuroblastoma model. Overall, these top model genes were significantly less studied than documented disease genes but nonetheless associated with the specific disease, thus providing a data-driven avenue for better understanding the genetic basis of human diseases (Figure 3.6). It is unlikely that these top model genes are causal genes, but more likely genes amplified at the end of a deregulated signaling pathway. Closely studying the function and structure of these genes could help unravel specific pathogenesis that are distinctive of a particular human disease such as diabetes, Alzheimer, colorectal cancer, and many others.

3.3.4 Anatomical context of the human disease landscape is well-summarized using URSA^{HD} models

Understanding the anatomical site of each disease is crucial for accurate diagnosis and treatment of the disease. A unified human disease framework must account for such anatomical characteristics while not over-fitting for tissue-specific signals. A data-driven approach may be over-optimistic and identify the responsible tissue-type rather than the specific disease signal. In order to control for such bias, corresponding normal tissue samples are used as negative samples to discourage any discrimination derived only from the tissue-specific differences between an unrelated disease and the disease of interest (see Methods). Nonetheless, many tissue-specific associations are found with both cancerous and noncancerous diseases (Figure 3.8). Figure 3.9 summarizes the disease model associations with T-lymphocyte specific genes. Not surprisingly, mycosis fungoides (a common form of cutaneous T-cell lymphoma) and peripheral T-cell lymphoma were exclusively associated with T-lymphocytes and not

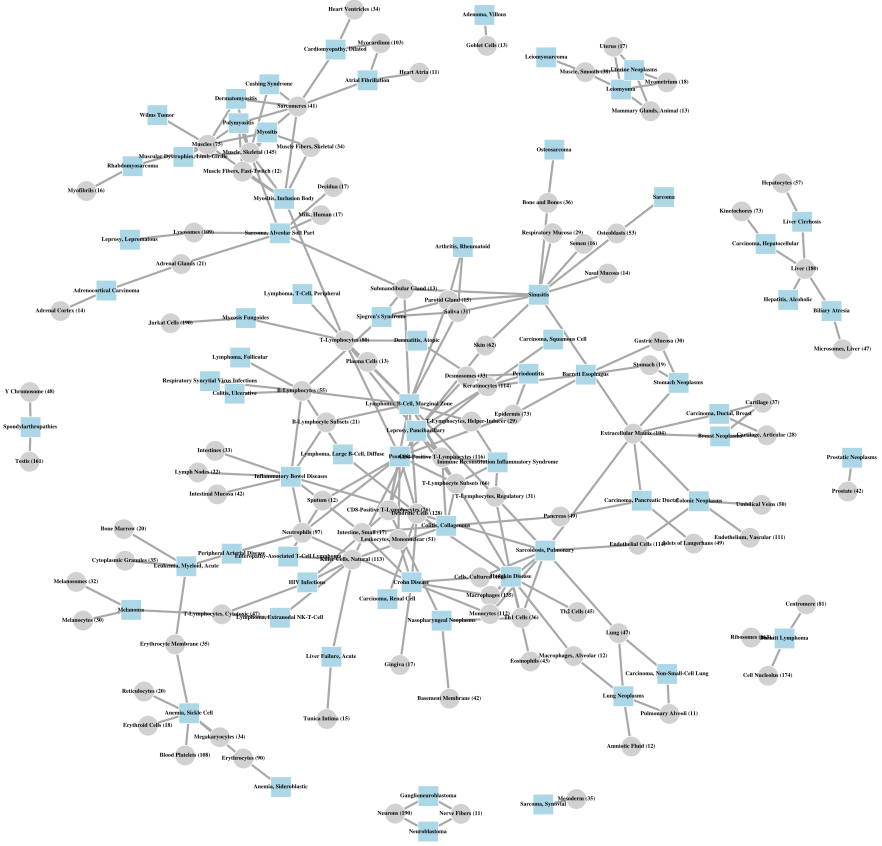


Figure 3.8: Anatomical context defined by URSA$^{\text{HD}}$'s disease models. Bipartite graph of disease terms (blue squares) and anatomical MeSH terms (grey circles). Association based on enrichment score ¿ 5. Heart diseases are connected to heart-related tissues/cell-types; and tissue-specific cancers are connected to their appropriate tissue of origin.

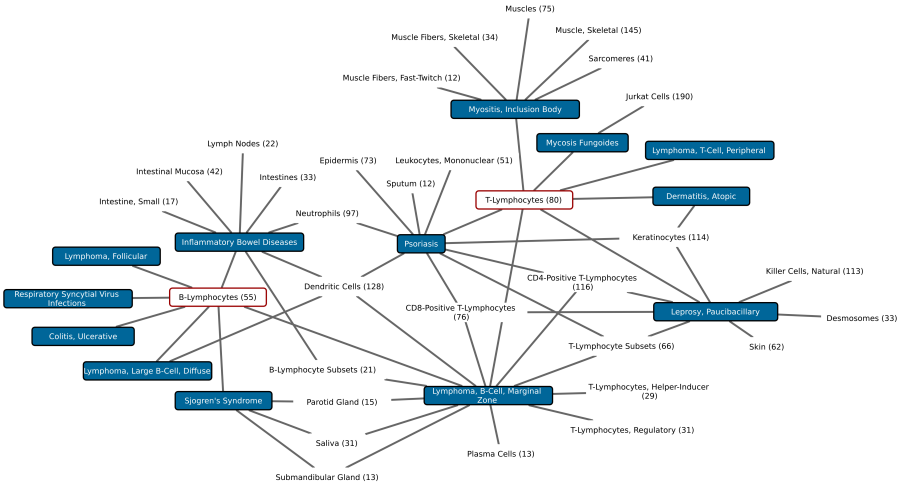


Figure 3.9: Diseases associated with T-lymphocytes and B-lymphocytes (red borders). Bipartite graph of disease terms (blue squares) and anatomical MeSH terms (black font). Association based on enrichment score ¿ 5. Both show associations with multiple immune-related diseases, but not many diseases are associated with both T-lymphocytes and B-lymphocytes.

B-lymphocytes. The anatomical context from URSA's inclusion body myositis (IBM) a type of inflammatory myopathy characterized by the invasion of T cells to muscle fiber tissue - model was well represented in this local bipartite graph - connecting T-lymphocyte and skeletal muscle-related anatomical terms. It is worth emphasizing that no gene selection or prior knowledge of IBM is used to construct URSA^{HD}'s IBM model. B-lymphocyte genes were instead over-represented with B-cell lymphomas such as follicular lymphoma and diffuse large B-cell lymphoma (Figure 3.9). Autoimmune or immune-mediated pathogen diseases associated with B-lymphocytes were Sjogren's syndrome, inflammatory bowel diseases, and respiratory syncytial virus infections. This separate clustering among immune-related diseases shows the distinctive, anatomical context set by URSA^{HD}'s data-driven disease models to characterize the human disease landscape. See Supp. Table 5 for the complete list of the anatomical enrichment scores for all disease models.

3.3.5 URSA^{HD} detects molecular disruptions in clinical samples after effective therapeutic treatment

Accurately quantifying a treatment's effectiveness is crucial for understanding drug-specific resistance and developing personalized medicine. However, treatments are often poorly understood at the molecular level and more so for its patient-specific outcomes. To our knowledge, no computational method has quantified the efficacy of therapeutic treatments in gene expression profiles. Here we tested URSA^{HD}'s ability to recognize the genome-wide disruption caused by therapeutic drugs in gene expression profiles of post-treatment clinical samples. Note that URSA^{HD} models don't use any post-treatment samples for training and so are completely oblivious to their potential outcome and response (see Method). We examined URSA^{HD}'s predictive probability of an ulcerative colitis dataset with both pre-treatment and post-treatment samples (GEO: GSE16879) [3]. Complete mucosal healing was clinically accessed 4 - 6 weeks after infliximab treatment. Both pre-treatment and post-treatment samples in this dataset were not used to learn/train URSA^{HD}'s disease models. We found that URSA^{HD} predictive probability distribution of the response groups are indistinguishable before treatment but differentiate after treatment, concordant with the independent clinical assessment (Figure 3.10a). We next examined URSA^{HD}'s predictive probability for a psoriasis dataset with brodalumab post-treatment samples (GEO: GSE53552) [114]. Brodalumab is an interleukin-17 antibody that prevents interleukin-17 ligands from binding to cell receptors. 3 skin biopsies (pre-treatment lesional, post-treatment lesional, and non-lesional) were collected from 25 patients with moderate to severe plaque psoriasis. Patients were divided into groups and treated with a single-dose of brodalumab (n = 4, 140mg subcutaneously; n = 8, 350mg subcutaneously; n = 8, 700mg intravenously; n = 5, placebo). URSA^{HD}'s estimate for psoriasis signal was low (essentially 0) for non-lesional samples and high for pre-treatment lesional samples (Figure 3.10b, left). Again, URSA^{HD} is oblivious

of the effects of brodalumab treatments or any other treatment effect (see Methods). Nonetheless, we found a significant decrease of URSA^{HD}'s psoriasis signal for post-treatment samples in a dose-dependent manner (Figure 3.10b). Because URSA^{HD} is general and not tuned for a specific drug, URSA^{HD} can be used to track the efficacy of any drug in any treatment setting for which clinical biopsies are available.

3.3.6 Repurposing drugs for the treatment of rare diseases using URSA^{HD}'s distinctive models.

Understanding the molecular behavior of human diseases leads to a more targeted development of new medicine. However, the therapeutic drug development process can often take more than 10 years after drug discovery, pre-clinical trials, clinical trials, and FDA review. The development process for rare diseases is particularly challenging as they have been under-characterized compared to other common diseases. Drug repurposing is a common approach to expedite this process. With no prior knowledge, our unified approach provides a data-driven, genome-wide characterization with distinctive functional and anatomical associations even for rare diseases. We test whether URSA^{HD}'s disease models and its disease associations (based on known disease genes) could be used to prioritize existing drugs for the treatment of another disease (Figure 3.11a). We find that the association scores of diseases with a common therapeutic drug interaction are statistically greater than the scores of random diseases (paired ranked-sum test, negative log p-value = 23, see Methods). It is worth mentioning that the random association scores are low between -0.4 and 0.8 and statistically insignificant. Accordingly, therapeutic drugs for a common disease could be used to treat rare diseases based on the disease associations derived from URSA^{HD}'s disease models.

'Sideroblastic Anemia' and 'Refractory Anemia with Excess Blasts' (RAEB) are both conditions in which the blood does not have enough healthy red blood cells

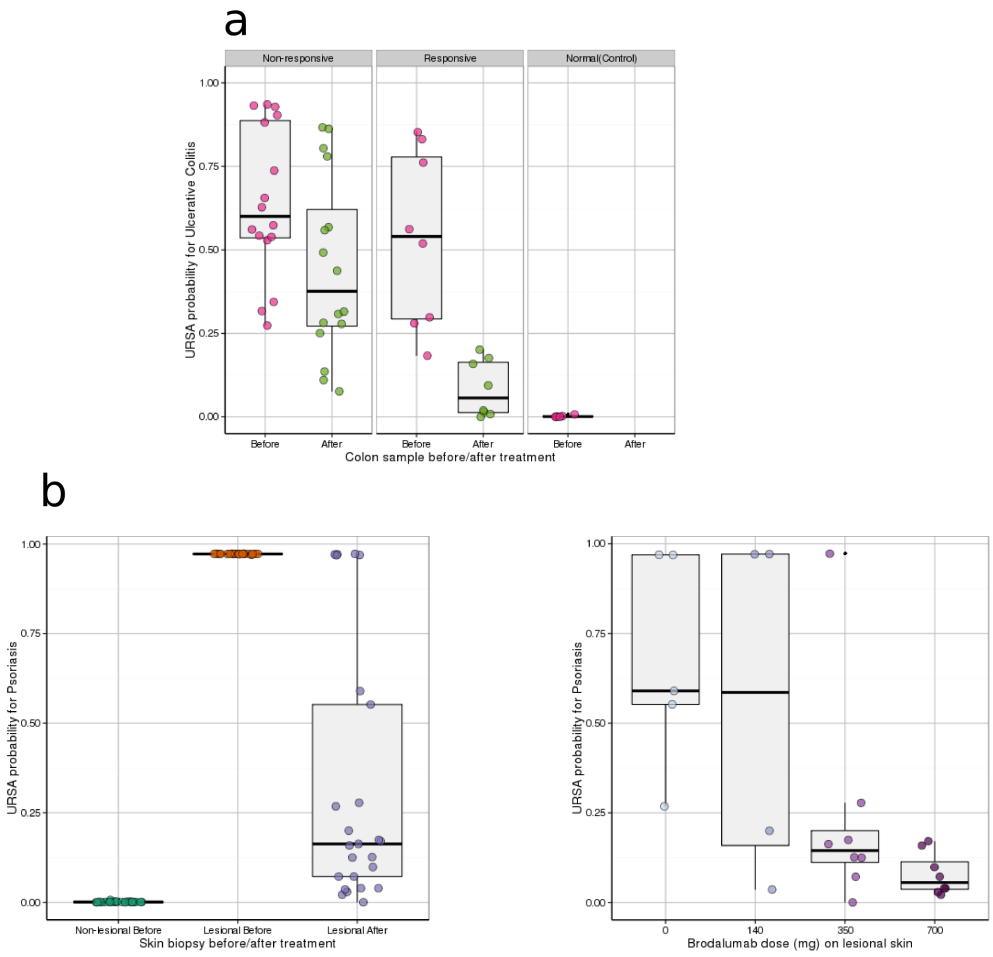


Figure 3.10: (a) URSA^{HD}'s Ulcerative Colitis predictions of ulcerative colitis samples before and after treatment (GEO: GSE16879). Each panel plots non-responsive, responsive, and control samples, respectively. URSA^{HD} estimated high predictive probabilities for both responsive and nonresponsive samples before treatment, but only the probabilities for the responsive samples decreases after treatment. (b) URSA^{HD}'s psoriasis predictions of skin biopsy samples before and after treatment (GEO: GSE53552). (left) URSA^{HD} estimated high predictive probabilities for lesional samples before treatment and low probabilities of samples after treatment. Non-lesional samples are used as a control. (right) The variation of estimated Psoriasis signal negatively correlates with dose of brodalumab treatment.

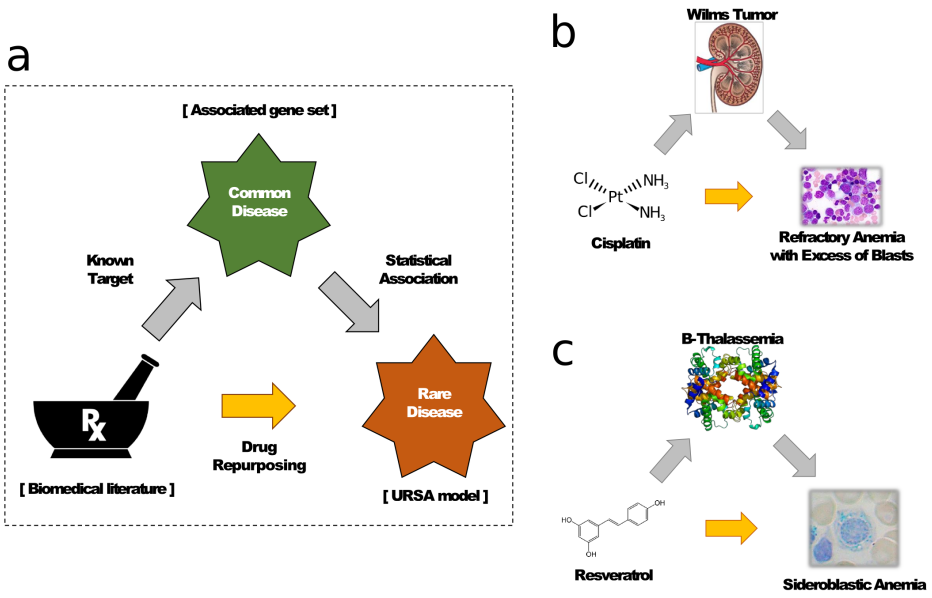


Figure 3.11: (a) Drug repurposing schematic (b,c) Two novel target predictions for two different anemias.

to carry oxygen [117] and so exhibit similar gene expression profiles (median sample correlation within 0.891, 0.875 and between 0.875) based on anemia samples in our curated gene expression compendium. The underlying molecular mechanism for these diseases are distinct, and their difference is well-characterized by URSA^{HD}'s sideroblastic anemia and RAEB disease models. RAEB is a myelodysplastic syndrome (MDS) that frequently progresses to acute myeloid leukemia [23, 33]. Based on its statistical association with Wilms tumor, this distinction was recapitulated in URSA^{HD}'s drug predictions for cancer chemotherapy drugs such as cisplatin, etoposide, melphalan, tretinoin, and vincristine (Figure 3.11b). These drugs have previously been shown to have an effect on RAEB transformation to acute myeloid leukemia (AML) or on RAEB itself [54, 62, 69, 145]. One of the hallmarks of RAEB is the aberrant hyper-methylation of gene promoters and such methylation-related GO terms were enriched in URSA^{HD}'s RAEB model (Figure 3.12) [55]. Cisplatin has been recommended for treating a variety of cancers including head and neck, bladder,

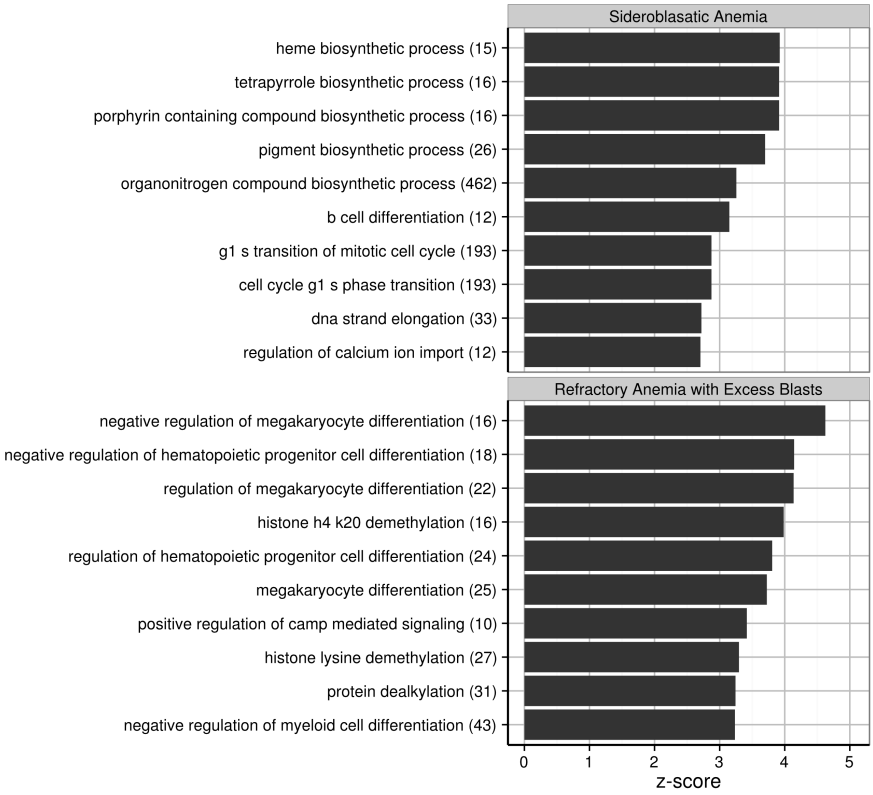


Figure 3.12: Enrichments for URSA^{HD}'s Sideroblastic Anemia model and Refractory Anemia with Excess Blasts (RAEB) model accurately describes the inefficient binding and/or transportation of the heme molecule in Sideroblastic Anemia, and the misregulation of hematopoiesis in RAEB.

lung, testicular and ovarian [140]. In particular, it has been shown to reverse hypermethylation of target genes in cervical cancer, further warranting the effectiveness of cisplatin for targeting RAEB or AML [128]. These drug predictions for sideroblastic anemia and RAEB follow the distinctive GO term enrichments of URSA^{HD}'s data-driven, integrative analysis of human diseases. Hence, our approach could inform the development of drug repurposing for the treatment of rare diseases without prior knowledge and given only a few gene expression profile samples from patients with the rare disease.

In sideroblasatic anemia, iron in the blood accumulates in the mitochondria due to the defects in heme biosynthesis, mitochondrial protein biosynthesis, iron metabolism

or [Fe-S] cluster biosynthesis [88, 123]. This molecular trait is indeed represented in URSA^{HD}'s sideroblastic anemia model and supports its disease associations and drug predictions (Figure 3.12). Iron chelators, hematopoiesis stimulants, and anti-oxidants (such as deferiprone, resveratrol, and mangiferin) were among the top drug treatment predictions for sideroblastic anemia based on its association with -Thalassemia and Iron Overload (Figure 3.11c). Such drug treatments are consistent with current therapeutic treatment of iron chelators and anti-oxidants to prevent iron overload and oxidative stress caused by sideroblastic anemia [14, 89]. In particular, resveratrol is especially interesting as it has been shown to have both anti-oxidant and hemoglobin-activating activity. In cultured human erythroid progenitor cells and K562 cells, resveratrol has been shown to accelerate erythroid maturation and increase hemoglobin levels [31, 32]. In a mouse model of β-thalassemia, resveratrol reduces ineffective erythropoiesis and increases red cell survival in the presence of oxidative damage [32].

3.4 Discussion

The amount of genome-wide experiments in clinical studies is growing and so opening the possibility of various integrative analysis of complex diseases. There are many questions that this large compendium of clinical data could answer. Here, we formulate it in a unified framework to identify distinctive characteristics of hundreds of human diseases. No pre- or post- feature selection method is used for a data-driven and unbias analysis of the compendium. The most predictive genes identified by our data-driven approach were significantly under-studied in the biomedical literature and thus providing a novel perspective for investigating the genetic basis of human diseases. Understanding the pathogenic role of these highly weighted genes in each individual disease model may shed insight for targeted detection and treatment.

Typical genome-wide normal/disease differential analysis have helped us better understand the abnormality of complex diseases at a molecular level. However, these complex diseases are similar and different at multiple levels and so an understanding of the distinctive characteristics of each diseases is need especially for closely-related diseases. We demonstrated here that our integrative method is able to identify this distinction in a data-driven manner. URSA^{HD}'s disease models map the heterogeneous landscape of multiple diseases to their proper functional and anatomical context. The sensitivity and specificity of our method is highlighted by capturing the genome-wide effect of therapeutic treatments for various clinical samples. Such *in-silico* sample evaluation can assist researchers and clinicians uncover subtle patient-specific molecular dysregulations in response to specific treatments.

Extending our approach to other diseases is straightforward and sidesteps the need for prior knowledge of causal genes or tissue of origin. In fact, URSA^{HD} models the expression phenotype of the disease and thus is complementary to genotyping studies. The overall approach incorporates a large collection of healthy tissues and so automatically accounts for any tissue-specific signals of the disease. This flexibility is particularly useful for studying rare diseases or diseases of unknown origin. In some sense, the functional and anatomical characteristics of any of the 7,000 rare diseases could be identified with an addition of two gene expression profiles.

Chapter 4

Dataset-specific integration of the public data compendium

This chapter describes work done with critical support and comments from Christopher Park and Aaron Wong.

4.1 Introduction

Genome-wide databases of physical/genetic interaction, gene expression, and perturbation data offer multiple perspectives of the underlying molecular system [130, 61, 83, 93, 28, 131]. Data integration methods combine these vast but complimentary assays to infer the molecular network of the genome [51, 76, 146, 101]. Much progress have been made to construct both accurate and system-specific networks taking into account the network dynamics in tissues, immune system and interaction types [37, 39]. Yet, biologists have specific genome-wide questions ranging from knock-down experiments to disease progression patterns in clinical samples.

Gene expression profiling is one of the most popular genome-wide experiments to answer these specific genome-wide questions [8]. Gene-gene correlation-based methods infer the co-expression patterns in the specific gene-gene network but plagued by

false positives. Functional associations other than co-regulated genes are discounted as the experiment focuses on the transcriptional change and not other complimentary assays. A new computational method is needed that both captures the specificity of the question that also automatically integrates the large, heterogeneous data compendium for accurate network inference.

Here we present YETI (Your Evidence Tailored Integration), an automatic data integration framework that utilize heterogeneous data sources to infer functional gene-gene interactions relevant to the biologist's dataset. Independent of the input dataset, YETI first constructs context-specific genome networks to survey the landscape of the dynamic molecular network. Then, YETI identifies dataset-relevent context networks by recasting it as a regression problem and then builds a single dataset-specific functional network. Through this framework, we demonstrate the YETI networks are not only accurate but also specific to the original genome-wide question. The selected contexts for each dataset are distinct and reproducible and can be used as *functional barcodes* to link other similar genome-wide datasets.

In the rest of the paper, we describe YETI in detail and the systematic comparison with previous network inference methods. We show that YETI networks are unbias to the size of the dataset with wider coverage of the human genome than the input dataset. In particular, the YETI network based on a brain eQTL study effectively discounts spurious gene-gene correlations and infers reproducible network modules regulated by distal eQTLs. We assess the relevance of the distinct context selection and its robustness over biological replicates.

4.2 Methods

4.2.1 Heterogeneous genome-wide data source

Gene expression

980 public microarray datasets were collected from NCBI Gene Expression Omnibus (GEO) consisting more than 22000 experiments [8]. Probes were pre-processed and normalized as done previously [51, 25, 101]. For each gene pair, pearson correlation was computed, transformed using Fisher's transformation and then standard normalized.

Physical and genetic interaction

Physical and genetic interaction data were downloaded from BioGRID, IntAct, MINT, and MIPS and were encoded based on the support of the gene-gene interaction [130, 61, 83, 93]. Transcriptional regulatory interactions were estimated based on TF binding site motifs from Jaspar and further processed using FIMO as done previously [116, 38, 101].

Perturbation data

Curated chemical and genetic perturbation data and motif-based microRNA target data were downloaded from GSEA and encoded based on the normalized co-occurrences of gene pairs for each dataset [131].

Protein sequence data

Protein sequence similarity data was downloaded from Biomart, and protein domain information was downloaded from PfamA and Prosite and then binarized [58, 11, 50].

4.2.2 Dataset-specific functional relation network construction

Context-sensitive data integration and network construction

We integrated heterogeneous genome-wide data in a context-sensitive manner adhering to the functional variation in genome-wide data shown previously in [96, 46, 150]. For each expert-selected fringe GO biological process ($n = 237$), we applied context-sensitive Bayesian integration to predict the context-dependent functional relations of 25825 genes covered by the processed genome-wide data [96, 51]. Regularized naïve Bayes classifier was used for integration to account for the inter-dependence of large-scale genome-wide data [51]. For training, gene pairs co-annotated to the fringe GO biological process were considered as known functional interaction standards (i.e. positive examples) and those not co-annotated to any terms in BioCyc, the GO fringe, KEGG, and PID were considered as non-interacting pairs (i.e. negative examples). See [101] for details in constructing the gold standard used for training the classifier.

Dataset-specific selection of context-specific functional networks

Optimal covariate (i.e. context network) selection is NP-hard and so we approximate the optimal via lasso [98, 136]. We formulate the dataset-specific selection problem as a regression problem where the dependent y variable is the genome-wide dataset and the independent variables x are the context-specific networks. Specifically, we compute the distance correlation of a genome-wide dataset for every known functional gene-gene interactions and assume the correlations to be noisy observations of the dataset-specific functional interactions [133].

For gene A and gene B, let (A_1, B_1), (A_2, B_2), (A_3, B_3) be identically distributed random variables. The distance correlation between gene A and gene B is:

$$
\begin{aligned}
\mathrm{dCor}(A,B) &= \frac{\mathrm{dCov}(A,B)}{\sqrt{\mathrm{dVar(A)dVar(B)}}} \\
\mathrm{dCov}^2(A,B) &= \mathrm{E}[\,||A_1 - A_2||\,||B_1 - B_2||\,] \\
&\quad + \mathrm{E}[\,||A_1 - A_2||\,]\,\mathrm{E}[\,||B_1 - B_2||\,] \\
&\quad - 2\mathrm{E}[\,||A_1 - A_2||\,||B_1 - B_3||\,] \\
\mathrm{dVar}^2(A) &= \mathrm{E}[\,||A_1 - A_2||^2\,] \\
&\quad + \mathrm{E}^2[\,||A_1 - A_2||\,] \\
&\quad - 2\mathrm{E}[\,||A_1 - A_2||\,||A_1 - A_3||\,]
\end{aligned}
$$

where $||\cdot||$ denotes Euclidean norm. While our method is not limited to distance correlation, we use distance correlation because of its robustness to false positives [126].

We use lasso to compute a sparse solution of dataset-relevant context networks. Both the distance correlations y and the networks x were logit-transformed for dataset-specific selection. Let $y = \{y_{i,j}\}$ for any gene i and gene j with known functional gene-gene relationship. Let x be the corresponding functional association score in those 237 context-specific functional networks. Lasso optimizes the following:

$$
\min_{|w|} \frac{1}{2N}||y - Xw||_2^2 + \lambda||w||_1 \tag{4.1}
$$

where N is the number of known functional gene pairs. The λ free parameter is usually fitted via cross-validation, but here we use the covariance test for lasso to select significant covariates (i.e. context networks) instead of minimizing least square error [84]. Covariates with $p < 0.01$ were selected for 100 lars steps. When less than 20 covariates were selected, covariates up to the minimum p-value between steps 20 and 80 were added for robustness. Finally, selected dataset-relevant context networks

were averaged (i.e. edge weights) to construct the final dataset-specific functional networks.

4.2.3 Systematic evaluation of dataset-specific functional networks

We evaluated the accuracy of dataset-specific functional networks by computing the network density of dataset-relevant genes. Genes annotated to disease MeSH terms that are then associated to the GEO GDS dataset were considered dataset-relevant genes. Disease MeSH term annotations were obtained from gene2mesh (`http://gene2mesh.ncibi.org/`). Disease MeSH terms indexing the associated publication of the GDS dataset were considered dataset-relevant MeSH terms.

Given a graph (i.e. network) $G = \{V, E\}$ with vertex set $V = \{g_1, g_2, \ldots, g_n\}$ and edge set $E = \{e_{ij}\}$ for $i, j \in V$, the density ρ of gene set $S = \{g_1, g_2, \ldots, g_m\}$ in G were defined as the following:

$$\hat{\rho}_G(S) = \frac{2}{m(m-1)} \sum_{g_i, g_j \in S \cup V} e_{i,j} \tag{4.2}$$

$$\rho_G(S) = \frac{\hat{\rho}_G(S)}{\hat{\rho}_G(V)} \tag{4.3}$$

Notice that $\hat{\rho}_G(V)$ is the density (i.e. average edge score) in G. This normalization is for comparison between networks with different global edge score distributions.

4.2.4 Analysis of distal eQTL associated gene modules

We present a network approach for prioritizing distal eQTLs associated with functional modules over those inevitable distal eQTLs from multiple hypothesis testing and spurious gene-gene correlations. Network density ρ of distal eQTL-associated

genes were used to evaluate and distinguish putative functional modules from spurious modules.

Gene and SNP location (hg19), processed normal brain genotype and expression data were obtained from seeQTL [147, 95]. Sample covariate data (i.e. gender, age, PMI) were obtained from the original paper [95]. SNP within a million distance were considered local gene-SNP pairs. MatrixEQTL was used to test gene-SNP associations [122]. Local eQTL's with $p < 10^{-2}$ and distal eQTL's with $p < 10^{-5}$ were considered significant. A distal eQTL with at least 10 significant gene associations were considered a *module regulating* distal eQTL.

Subset of the data ($n = 50, 75, 100$) were re-analyzed with the same parameters to assess the reproduciblility of these distal eQTL modules. For each subset, putative distal eQTL modules were ranked based on its density score ρ in either YETI's network or distance correlation network. Accuracy of the ranks were evaluated by comparing to distal eQTL modules found using all the data.

4.2.5 Statistical robustness of functional inference

We evaluated the robustness of our dataset-specific network inference method by comparing the dataset-specific context selection and network density of dataset-relevant MeSH terms. Dataset-specific networks were constructed via bootstrapping ($n = 30$) known functional interaction standards to assess its robustness to known functional interactions. Dataset-specific networks were also constructed via subsampling ($n = 30$) biological replicates in the genome-wide dataset to assess its robustness to the available biological replicates in the experimental study. Of the GEO GDS1733 dataset, only control samples across 6 time-points with 3 replicates each were used. Replicates from each time-point were subsampled ($m = 1, 2$) and individual networks constructed to assess the statistical robustness of the method.

4.2.6 Relevant public genome-wide datasets based on context network selection

We retrieve dataset-relevant public genome-wide datasets based on similar context network selection. Dataset-specific network selection was computed for 464 GEO GDS human datasets. Hypergeometric test was used to calculate the statistical significance of selection overlap with that of the user's dataset. GEO GDS datasets with $p < 0.001$ were considered statistically significant and so dataset-relevant.

4.2.7 Implementation

Software used for public data integration and distance correlation calculation has been implemented in the open source *Sleipnir* library available at `http://libsleipnir.bitbucket.org` [52]. R packages *lars* and *covTest* were used for lasso regression and significance test [136, 84].

4.3 Results

We developed a general method YETI (Your Evidence Tailored Integration) for dataset-specific data integration and network prediction. We systematically evaluated the functional accuracy and dataset-relevance of our method over hundreds of public genome-wide expression data across different experimental procedures and technical platforms. Our method enables experimental study-specific integration of large and heterogeneous data and genome-wide exploration of the particular functional landscape.

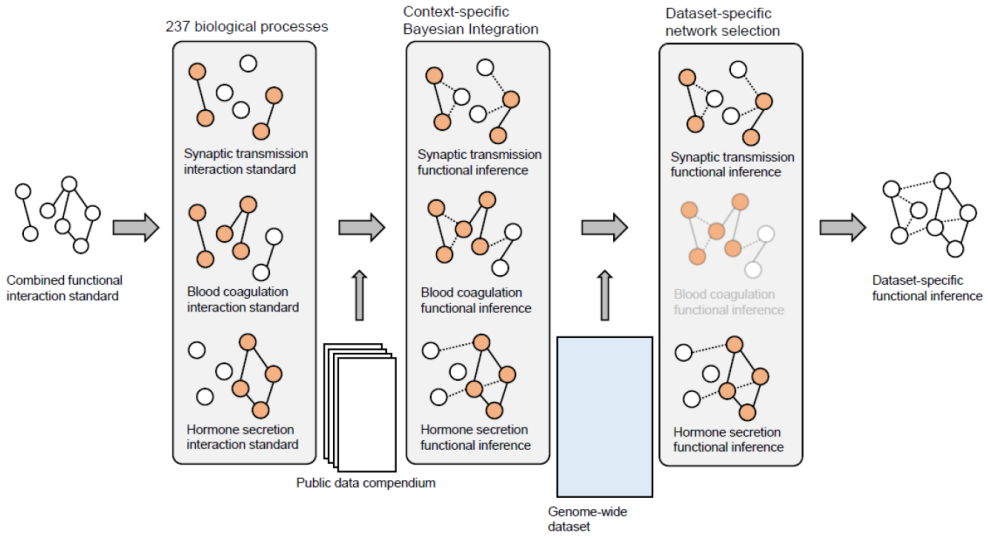


Figure 4.1: Flowchart for dataset-specific aggregation of context-specific functional interaction networks. YETI's dataset-specific network construction involves a dataset-independent integration step followed by a dataset-relevant context selection step.

4.3.1 Dataset-specific integration of the public human data compendium

YETI uses putative gene-gene functional associations from the genome-wide dataset to then select predicted functional associations derived from the large heterogeneous data compendium (i.e. sequence similarity, physical interaction, co-expression, etc) (Figure 4.1). Diverse contexts of 237 GO fringe biological processes were used to represent the known functional interaction standards. Based on these standards, YETI first predicts dataset-independent but context-specific functional network maps to effectively leverage the context-dependent functional variation in genome-wide data (Figure 4.2). Of these 237 context-specific networks, the method selects dataset-relevant contexts by formulating it as a regression-based feature selection problem. Briefly, the putative, dataset-specific gene associations is modeled as a noisy random variable manifested by the combination of dataset-relevant context networks. These

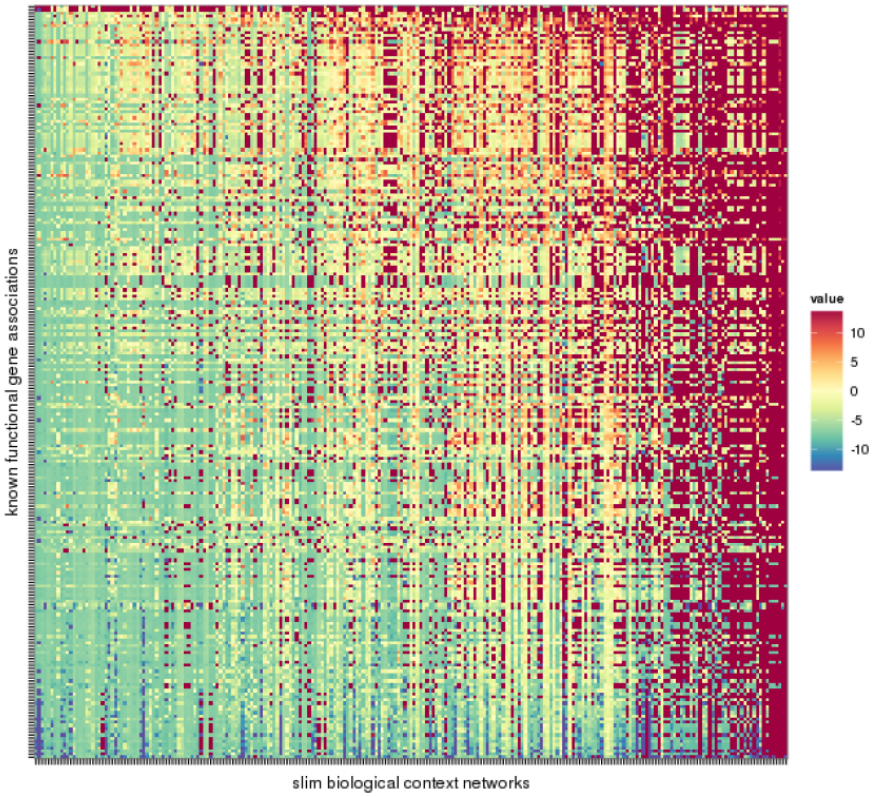


Figure 4.2: Heatmap of edge scores of 237 context-specific networks and subsample of known functional gene-gene associations. The strength of association for each known functional gene-gene pair across GO fringe context networks are shown as a heatmap. The edge score (i.e. posterior probability score) for each gene-gene pair were logit-transformed for visualization. YETI selects these context networks (in columns) that represent the dataset-specific functional associations latent in the noisy genome-wide dataset.

dataset-relevant context networks are aggregated to estimate the dataset-specific functional network. Notice that no prior gene or pathway information is required to model the genome-wide functional characteristics of the experimental study.

4.3.2 Dataset-specific networks retrieved dataset-relevant functional and disease network modules

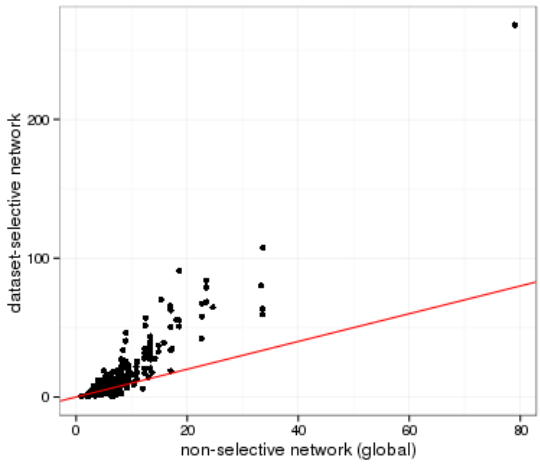


Figure 4.3: Increased statistical power to detect dataset-relevant network modules. Improved subgraph density (i.e. expected within edge weight) of dataset-relevant disease genes in dataset-specific network over global (non-selective) network. Red line indicates the identity line.

YETI's dataset-specific functional networks favored the association of dataset-relevant functional and disease modules. We anticipated the reinforcement of dataset-specific functional associations with the proper fitting of the human public compendium to the given dataset. We constructed dataset-specific networks from 462 GEO human GDS datasets and evaluated the network association of modules (i.e. genesets) of dataset-relevant diseases. These GDS datasets were across various biological systems and technical platforms. Dataset-relevant diseases were based on pubmed

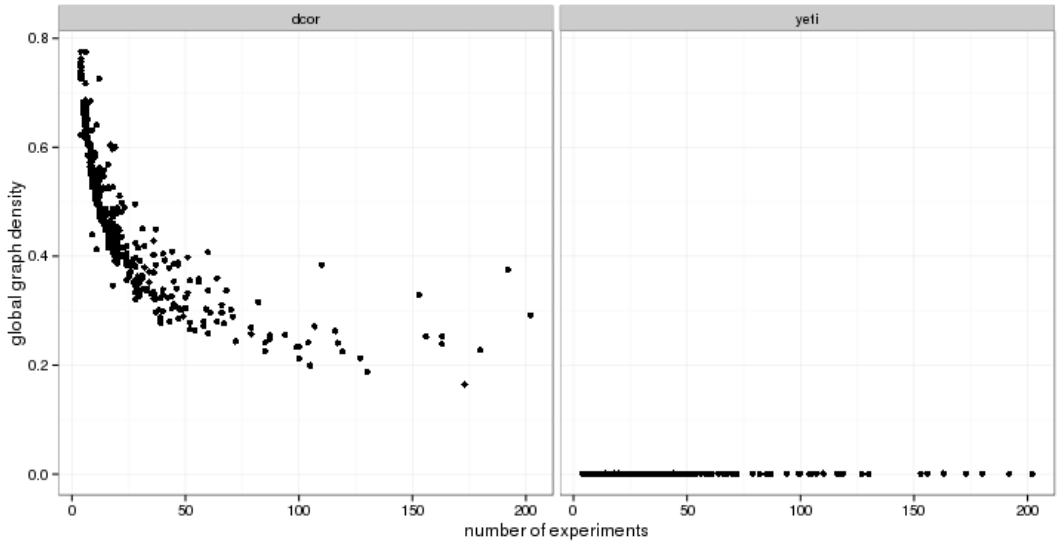
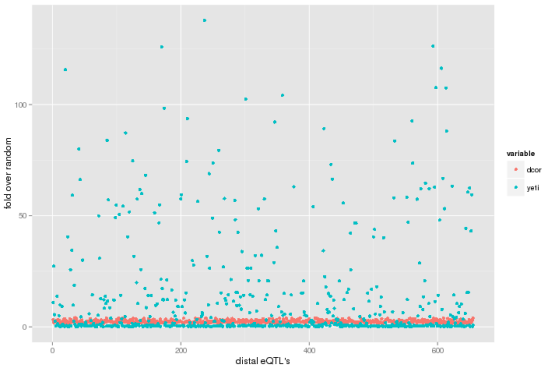
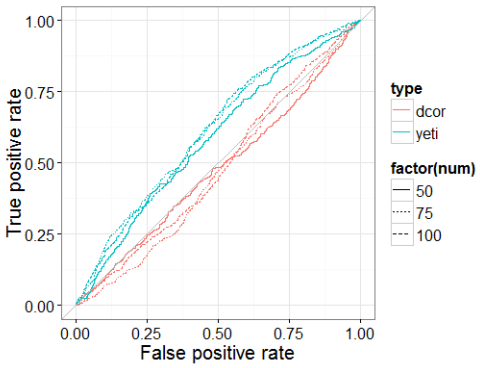


Figure 4.4: Whole graph densities of correlation-based networks were heavily biased by the size (i.e. number of experiments) of the dataset while well-controlled in YETI's dataset-specific networks.



(a) Network density of modules (i.e. genesets) associated with putative distal eQTL in distance correlation network (in red) and YETI network (in blue)



(b) Reproducibility analysis via subseting the genotype and expression data

Figure 4.5: YETI's dataset-specific networks prioritize putative functional modules regulated by distal eQTL's. Subgraph densities of distal eQTL modules in correlation-based network and YETI's dataset-specific functional network. Prioritizing distal eQTL modules based on YETI's dataset-specific network were more likely to be reproduced in a subsampling statistical analysis.

annotations and so not used in the construction of dataset-specific networks (see Methods). Dataset-relevant disease modules were more enriched in YETI's dataset-specific networks than in the human global (i.e. non-selective) functional network (Figure 4.3). Much of this improvement could be attributed to YETI's effective selection of dataset-specific gene-gene associations and indifference towards systematic biases. The expected edge weight of the coexpression networks were heavily biased by the size (i.e. number of experiments) of the dataset as expected, while fairly consistent in YETI's dataset-specific networks (Figure 4.4).

YETI's dataset-specific networks reveal putative functional modules regulated by distal eQTLs. Distal eQTLs (or putative trans-eQTLs) regulate multiple genes at large genomic distances or different chromosomes [35]. Large-scale eQTL studies are popular methods for distal eQTL discovery but often suffer from multiple-testing and spurious gene-gene correlations of the genome-wide expression dataset. Using YETI, we tailored the human data compendium to the eQTL expression dataset and found a few distinctive distal eQTL modules enriched in the dataset-specific network (Figure 4.5a). We hypothesized these network-enriched modules to be functional modules regulated by the distal eQTL while others to be risen by spurious associations from the expression dataset. To confirm this, we conducted a subsampling-based statistical analysis and evaluated the module prioritization based on YETI and coexpression. We found that modules prioritized by YETI were more likely to be reproducible ($AUC \approx 0.62$) while the prioritization based on coexpression were close to random ($AUC \approx 0.5$) (Figure 4.5b). In fact, YETI outperformed distance correlation even with half the number of samples (i.e. gene expression profiles). Such functional accuracy illustrates YETI's effectiveness to incorporate heterogeneous genome-wide data for dataset-specific functional network inference.

4.3.3 Selection of dataset-relevant biological context networks were distinct and consistent while also shared among similar genome-wide studies.

Dataset-specific context selections were distinct with consistent preference for specific biological processes. Diverse context-specific networks were constructed to model the complex functional interactions across different biological processes. Context-sensitive Bayesian integration of heterogeneous genome-wide data was used to estimate specific functional associations of 237 GO fringe terms. These GO fringe terms were not mutually exclusive but selected for maximal coverage of the diverse functional interactions (see Methods). Strength of putative functional associations were in fact diversified across these context networks (Figure 4.2). Of this wide coverage of functional associations, YETI selects less than 80 context-networks with functional interactions that in combination mimic those detected in the dataset's co-expression structure. YETI's selections for 464 genome-wide datasets were sparse and fairly uniform across those 237 context networks (Figure 4.6). Strong preference for particular contexts were found over subsampling biological replicates ($m = 1, 2$) in a heat-shock response time-course dataset (Figure 4.8). In fact, even with 1 of the 3 biological replicates, YETI selected *microtubule anchoring* and *mitosis* over all 30 subsampled datasets and consistently ignored 71 biological processes such as .

YETI's dataset-specific selection linked genome-wide datasets with similar genome-wide functional interactions. The probability of two datasets randomly sharing a significant number of contexts follows the hypergeometric distribution. In fact, the actual probability is even smaller given YETI's dataset-specific selection preference (as shown above). Nonetheless, statistically significant selection overlaps between genome-wide datasets were found. The heat-shock time-course dataset shared 11 contexts with a lesional and non-lesional skin biopsy samples from 13

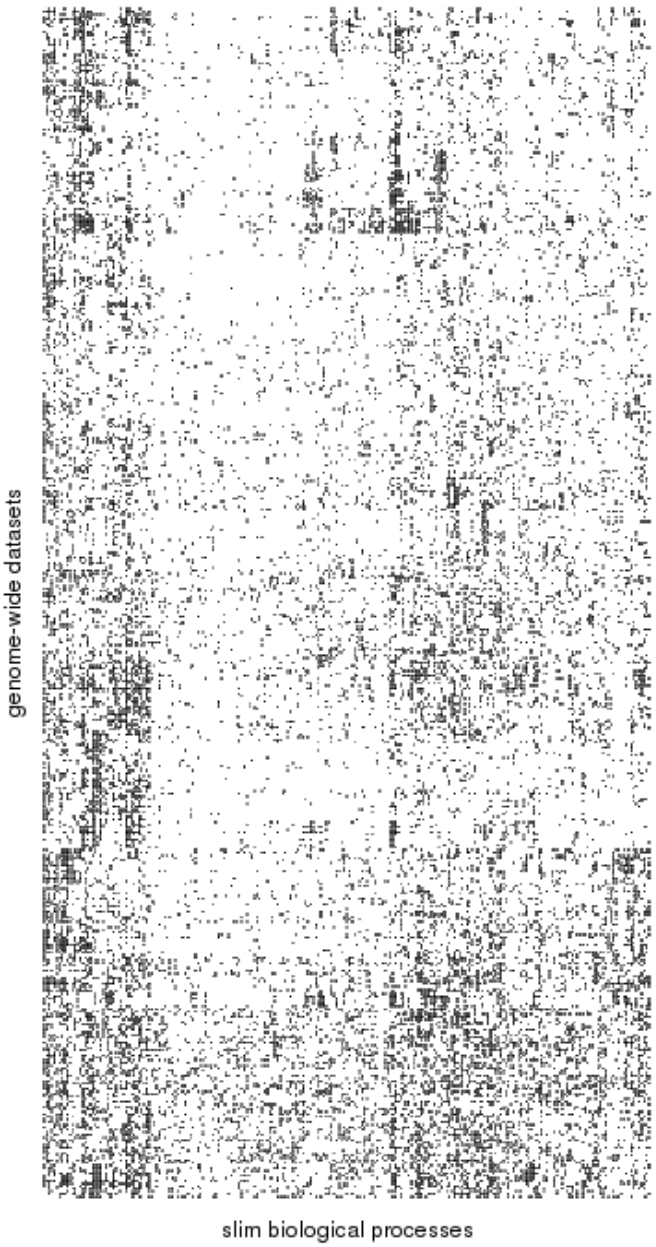


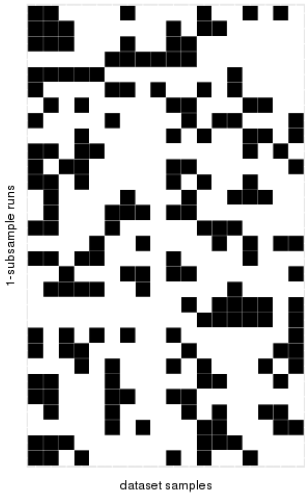
Figure 4.6: Dataset-specific selections were distinct and sparse with non-exclusive selection between genome-wide datasets. Selection (in black) of 237 context networks (in columns) for 464 genome-wide dataset (in rows) were hierarchically clustered with the manhattan distance metric.

Figure 4.7: Selected (in black) context networks (in columns) for the 1-subsampling run (in rows). 71 context networks were repeatedly not selected for multiple 1-subsampling runs. Strong selection preference was found despite the uniform random selection of biological replicates. Such consistent trend suggests proper identification of dataset-specific functional associations conditioned on known global functional associations.

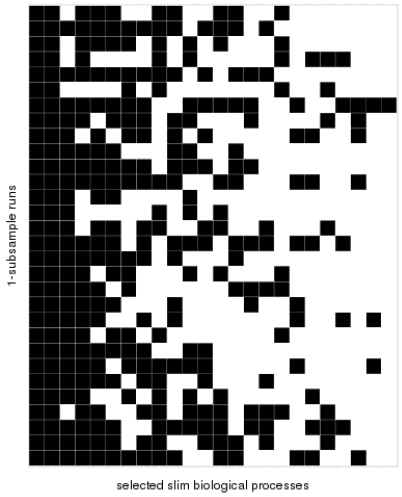
psoriatic patients (Fisher's exact test p-value: 0.000878). Its overlap with other public datasets ($n = 463$) were not statistically significant (p-value > 0.01). The directed differentiation time-course dataset (GEO GSE28191) was significantly linked (p-value < 0.001) to other cell differentiation datasets: hepatic stellate cells activation in response to liver damage (GDS3492), epidermal keratinocyte differentiation (GEO GDS2732), transdifferentiation in Barrett's esophagus (GDS3472), and chemotherapy resistance (GEO GDS2367, GDS3638). This directed differentiation dataset was also linked to a large ($n = 130$) primary squamous cell lung carcinoma dataset (GDS2373) suggesting potential shared functional modules between the driving factors of cardiomyocyte differentiation and low-risk vs high-risk prognosis. Note that the p-value distribution of context overlap was fairly uniform supporting our usage of the hypergeometric distribution to link genome-wide datasets.

4.4 Discussion

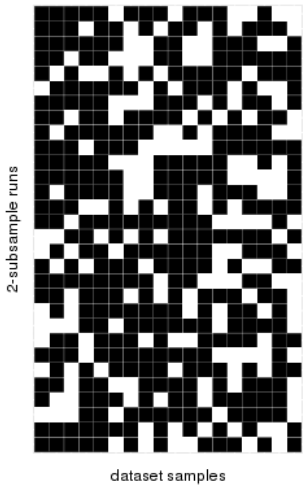
Our understanding of the dynamic functional interactions is incomplete. In fact, just recently have we been able to tap into the functional rewiring in the immune system and across different tissues [37, 39]. This has been accomplished by integrating the heterogeneous data compendium and easily extendable as more molecular data be-



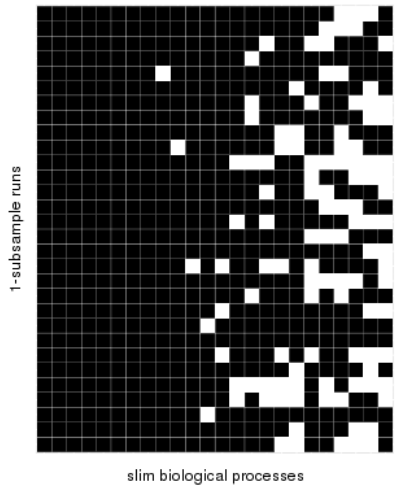
(a) 1-subsample replicate selection



(b) 1-subsample context selection



(c) 2-subsample replicate selection



(d) 2-subsample context selection

Figure 4.8: Dataset-specific selections were reproducible across subsamples of biological replicates for GDS1733: Time-course heat-shock response in HeLa cells with 3 replicates and 6 time points. Selected (in black) biological replicates (in columns) for each (1,2)-subsampling run (in row). Selected (in black) context networks among the those selected using all 3 replicates (in columns) for the (1,2)-subsampling run (in rows). Selection preference was, nonetheless, found even with 1 sample per timepoint.

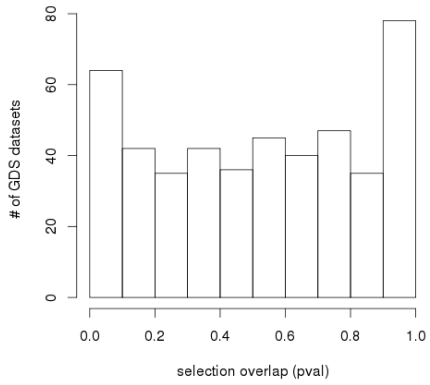


Figure 4.9: Histogram of p-value distribution of context selection overlap between differentiation dataset GSE28191 and 464 GDS datasets.

comes available. Nonetheless, there's a gap between broad functional networks and the biologist's study-design. These study designs target for specific functional interactions of cellular response to drug treatment or essential gene-gene associations in multiple tissue or cell lines. While the dataset could be incorporated as part of the compendium, no existing method tailors the vast compendium specific to the biologist's dataset and so most genome-wide studies under-utilize this publicly available molecular data compendium.

Here we developed an automatic data integration framework YETI that tailors the human data compendium to the input dataset without sacrificing accuracy. YETI's dataset-specific networks are more accurate than correlation-based networks by leveraging genome-wide databases that are complimentary to gene expression data. YETI's network accuracy is based on its construction of 237 diverse context (i.e. fringe GO BP terms) networks that span the dynamic molecular interactions in human cells. Only the specificity of the experimental study-designed is learned to avoid spurious gene-gene correlations in the input dataset. Distinct selection of these 237 context via the input dataset lead to specific network-level interpretation of the dataset and connection to other datasets with similar context selection. The modularity of YETI's

"infer-then-select" framework allows easy adaptation to improved context networks with more data in the future. Extensions for other types of input data such as proteomics or DNA methylation are also worth exploring for complementary data-driven exploration.

Genome-wide data are mostly organized in a single tab-deliminated spreadsheet or sql-like database for most genome-wide molecular databases. These organized databases have grown and will continue to grow. To better use these large data collections, scalable computational methods such as YETI are needed to bring the data compendium closer to the lab bench.

Chapter 5

Conclusion

Much work is needed to draw the complete molecular map of the human cell. This effort to map the dynamic and multi-faceted human cell cannot be completed with simple generation of data but coupled with effective analysis of these very large data collections. In this thesis, I presented three specific genome-wide analysis of molecular data that offers a data-driven perspective of these data collections. In Chapter 2, I developed a method to quantify tissue-specific signals in gene expression profiles by incorporating the tissue ontology. In Chapter 3, I used thousands of clinical samples deposited in GEO to characterize the human disease landscape. In Chapter 4, I developed a framework that integrates heterogeneous molecular data to infer functional gene-gene interactions specific to a particular genome-wide dataset.

More data will become available and so, much thought on how we could consume this data is needed. Computational methods for transferring information of common disease to rare diseases would offer a scalable approach to studying the thousands of rare diseases that is affecting millions of people worldwide. Rare diseases affect about 1 out of 2,000 people, yet more than 30 million people in the US are affected by a rare disease [65, 42, 15]. More than 5,400 rare diseases are catalogued in the Orphanet databases [5]. A comprehensive map of the interrelationships of human diseases is

needed to connect each individual rare disease in the context of the entire biomedical information. Linking and indexing this information in a semantic ontology allows for efficient extraction in the vast but unorganized data collections. Much manual, expert-driven construction of an ontology for human diseases is being made to make this information more accessible and interpretable [66, 121]. A computational, data-driven method for this construction is needed - in conjunction with expert curation - for effective management of the growing 'big' biomedical data. Such effort would enable repositioning of treatment for well-known diseases to closely-related rare diseases.

Identifying candidate drugs for repurposing is also an potential area of research that would benefit with effective analysis of large biomedical data collection. Drugs effect different parts of the body depending on the medium (oral or intravenous) of treatment and accurate understanding of the molecular response of human cells is needed for drug development, targeted treatment and drug repurposing. Yet, little is known for even the most commonly used drugs such as NSAIDs - the acting chemical in painkillers [43]. A future goal is to aggregate drug-related genome-wide data collections and identify the canonical genome-wide response distinctive of a particular drug treatment. Identifying the canonical response in human cells would be a backbone of understanding tissue/cell-type specific effects (or side-effects) of the various mediums of the drug treatment.

Bibliography

[1] David Amar, Tom Hait, Shai Izraeli, and Ron Shamir. Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. *Nucleic acids research*, 43(16):7779–7789, 2015.

[2] Michal Amit and Joseph Itskovitz-Eldor. Maintenance of human embryonic stem cells in animal serum-and feeder layer-free culture conditions. In *Human Embryonic Stem Cell Protocols*, pages 105–113. Springer, 2006.

[3] Ingrid Arijs, Gert De Hertogh, Katleen Lemaire, Roel Quintens, Leentje Van Lommel, Kristel Van Steen, Peter Leemans, Isabelle Cleynen, Gert Van Assche, Séverine Vermeire, et al. Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *PLoS One*, 4(11):e7984, 2009.

[4] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[5] Ségolène Aymé, Bertrand Bellet, and Ana Rath. Rare diseases in icd11: making rare diseases visible in health information systems through appropriate coding. *Orphanet journal of rare diseases*, 10(1):35, 2015.

[6] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[7] Tanya Barrett, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Rolf N. Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets10 years on. *Nucleic Acids Research*, 39(suppl 1):D1005–D1010, 2011.

[8] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data setsupdate. *Nucleic acids research*, 41(D1):D991–D995, 2013.

[9] Zafer Barutcuoglu and Christopher DeCoro. Hierarchical shape classification using bayesian aggregation. In *Shape Modeling and Applications, 2006. SMI 2006. IEEE International Conference on*, pages 44–44. IEEE, 2006.

[10] Zafer Barutcuoglu, Robert E Schapire, and Olga G Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[11] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl 1):D138–D141, 2004.

[12] Benjamin M Bolstad, Rafael A Irizarry, Magnus Åstrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[13] Claudio Bonifati and Franco Ameglio. Cytokines in psoriasis. *International journal of dermatology*, 38(4):241–251, 1999.

[14] Sylvia S Bottomley and Mark D Fleming. Sideroblastic anemia: diagnosis and management. *Hematology/oncology clinics of North America*, 28(4):653–670, 2014.

[15] Kym M Boycott, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681–691, 2013.

[16] Patrick O Brown and David Botstein. Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21:33–37, 1999.

[17] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[18] Damien Chaussabel, Virginia Pascual, and Jacques Banchereau. Assessing the human immune system through blood transcriptomics. *BMC biology*, 8(1):84, 2010.

[19] Abbot F Clark, H Thomas Steely, Jaime E Dickerson, Sherry English-Wright, Karen Stropki, Mitchell D McCartney, Nasreen Jacobson, Allan R Shepard, John I Clark, Hiroyuki Matsushima, et al. Glucocorticoid induction of the glaucoma gene myoc in human and monkey trabecular meshwork cells and tissues. *Investigative ophthalmology & visual science*, 42(8):1769–1780, 2001.

[20] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[21] ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[22] Chad J Creighton, Xiaoxian Li, Melissa Landis, J Michael Dixon, Veronique M Neumeister, Ashley Sjolund, David L Rimm, Helen Wong, Angel Rodriguez, Jason I Herschkowitz, et al. Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features. *Proceedings of the National Academy of Sciences*, 106(33):13820–13825, 2009.

[23] Antonio Cuneo and GL Castoldi. Refractory anemia with excess blasts (raeb). 2004.

[24] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[25] Manhong Dai, Pinglang Wang, Andrew D. Boyd, Georgi Kostov, Brian Athey, Edward G. Jones, William E. Bunney, Richard M. Myers, Terry P. Speed, Huda Akil, Stanley J. Watson, and Fan Meng. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Research*, 33(20):e175, 2005.

[26] Allan Peter Davis, Cynthia J Grondin, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Thomas C Wiegers, and Carolyn J Mattingly. The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic acids research*, 43(D1):D914–D920, 2015.

[27] Marek J Druzdzel. Smile: Structural modeling, inference, and learning engine and genie: a development environment for graphical decision-theoretic models. In *AAAI/IAAI*, pages 902–903, 1999.

[28] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[29] Jesse M. Engreitz, Rong Chen, Alexander A. Morgan, Joel T. Dudley, Rohan Mallelwar, and Atul J. Butte. Profilechaser: searching microarray repositories based on genome-wide patterns of differential expression. *Bioinformatics*, 27(23):3317–3318, 2011.

[30] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874, 2008.

[31] Eitan Fibach, Eugenia Prus, Nicoletta Bianchi, Cristina Zuccato, Giulia Breveglieri, Francesca Salvatori, Alessia Finotti, Michele Lipucci di Paola, Eleonora Brognara, Ilaria Lampronti, et al. Resveratrol: Antioxidant activity and induction of fetal hemoglobin in erythroid cells from normal donors and β-thalassemia patients. *International journal of molecular medicine*, 29(6):974–982, 2012.

[32] Sara Santos Franco, Luigia De Falco, Saghi Ghaffari, Carlo Brugnara, David A Sinclair, Achille Iolascon, Narla Mohandas, Mariarita Bertoldi, Xiuli An, Angela Siciliano, et al. Resveratrol accelerates erythroid maturation by activation of foxo3 and ameliorates anemia in beta-thalassemic mice. *haematologica*, 99(2):267–275, 2014.

[33] Ulrich Germing, Corinna Strupp, Andrea Kuendgen, Manuel Aivado, Aristoteles Giagounidis, Barbara Hildebrandt, Carlo Aul, Rainer Haas, and Norbert Gattermann. Refractory anaemia with excess of blasts (raeb): analysis of reclassification according to the who proposals. *British journal of haematology*, 132(2):162–167, 2006.

[34] Yoav Gilad and Orna Mizrahi-Man. A reanalysis of mouse encode comparative gene expression data. *F1000Research*, 4, 2015.

[35] Yoav Gilad, Scott A Rifkin, and Jonathan K Pritchard. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in genetics*, 24(8):408–415, 2008.

[36] Sophie Godard, Gad Getz, Mauro Delorenzi, Pierre Farmer, Hiroyuki Kobayashi, Isabelle Desbaillets, Michimasa Nozaki, Annie-Claire Diserens, Marie-France Hamou, Pierre-Yves Dietrich, et al. Classification of human astrocytic gliomas on the basis of gene expression a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. *Cancer research*, 63(20):6613–6625, 2003.

[37] Dmitriy Gorenshteyn, Elena Zaslavsky, Miguel Fribourg, Christopher Y Park, Aaron K Wong, Alicja Tadych, Boris M Hartmann, Randy A Albrecht, Adolfo García-Sastre, Steven H Kleinstein, et al. Interactive big data resource to elucidate human immune pathways and diseases. *Immunity*, 43(3):605–614, 2015.

[38] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[39] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569–576, 2015.

[40] Casey S. Greene and Olga G. Troyanskaya. Pilgrm: an interactive data-driven discovery platform for expert biologists. *Nucleic Acids Research*, 39(suppl 2):W368–W374, 2011.

[41] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The brenda tissue ontology (bto): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, 39(suppl 1):D507–D513, 2011.

[42] Robert C Griggs, Mark Batshaw, Mary Dunkle, Rashmi Gopal-Srivastava, Edward Kaye, Jeffrey Krischer, Tan Nguyen, Kathleen Paulus, Peter A Merkel, et al. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular genetics and metabolism*, 96(1):20–26, 2009.

[43] Tilo Grosser, Susanne Fries, and Garret A FitzGerald. Biological basis for the cardiovascular consequences of cox-2 inhibition: therapeutic challenges and opportunities. *The Journal of clinical investigation*, 116(1):4–15, 2006.

[44] Yuanfang Guan, Chad L Myers, David C Hess, Zafer Barutcuoglu, A Caudy, and O Troyanskaya. Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome biology*, 9(Suppl 1):S3, 2008.

[45] Laura M Heiser, Anguraj Sadanandam, Wen-Lin Kuo, Stephen C Benz, Theodore C Goldstein, Sam Ng, William J Gibb, Nicholas J Wang, Safiyyah Ziyad, Frances Tong, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8):2724–2729, 2012.

[46] Matthew A Hibbs, David C Hess, Chad L Myers, Curtis Huttenhower, Kai Li, and Olga G Troyanskaya. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, 23(20):2692–2699, 2007.

[47] Haiyan Huang, Chun-Chi Liu, and Xianghong Jasmine Zhou. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proceedings of the National Academy of Sciences*, 107(15):6823–6828, 2010.

[48] Earl Hubbell, Wei-Min Liu, and Rui Mei. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, 2002.

[49] Thomas J Hudson, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al. International network of cancer genome projects. *Nature*, 464(7291):993–998, 2010.

[50] Nicolas Hulo, Amos Bairoch, Virginie Bulliard, Lorenzo Cerutti, Edouard De Castro, Petra S Langendijk-Genevaux, Marco Pagni, and Christian JA Sigrist. The prosite database. *Nucleic acids research*, 34(suppl 1):D227–D230, 2006.

[51] Curtis Huttenhower, Erin M Haley, Matthew A Hibbs, Vanessa Dumeaux, Daniel R Barrett, Hilary A Coller, and Olga G Troyanskaya. Exploring the human genome with functional maps. *Genome research*, 19(6):1093–1106, 2009.

[52] Curtis Huttenhower, Mark Schroeder, Maria D Chikina, and Olga G Troyanskaya. The sleipnir library for computational functional genomics. *Bioinformatics*, 24(13):1559–1561, 2008.

[53] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[54] Kuniaki Itoh, Tadahiko Igarashi, and Hisashi Wakita. Letter to the editor: Successful treatment with vincristine by slow infusion in a patient with refractory anemia and excess of blasts. *American journal of hematology*, 39(1):73–74, 1992.

[55] Ying Jiang, Andrew Dunbar, Lukasz P Gondek, Sanjay Mohan, Manjot Rataul, Christine O'Keefe, Mikkael Sekeres, Yogen Saunthararajah, and Jaroslaw P Maciejewski. Aberrant dna methylation is a dominant mechanism in mds progression to aml. *Blood*, 113(6):1315–1325, 2009.

[56] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226. ACM, 2006.

[57] Dejan Juric, Sanja Sale, Robert A. Hromas, Ron Yu, Yan Wang, George E. Duran, Robert Tibshirani, Lawrence H. Einhorn, and Branimir I. Sikic. Gene expression profiling differentiates germ cell tumors from other cancers and defines subtype-specific signatures. *Proceedings of the National Academy of Sciences of the United States of America*, 102(49):17763–17768, 2005.

[58] Arek Kasprzyk. Biomart: driving a paradigm change in biological data management. *Database*, 2011:bar049, 2011.

[59] LJ Kaufman, CP Brangwynne, KE Kasza, E Filippidi, Vernita D Gordon, TS Deisboeck, and DA Weitz. Glioma expansion in collagen i matrices: analyzing collagen concentration-dependent growth and motility patterns. *Biophysical journal*, 89(1):635–650, 2005.

[60] Lia Kent. Culture and maintenance of human embryonic stem cells. *Journal of visualized experiments: JoVE*, (34), 2009.

[61] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, et al. The intact molecular interaction database in 2012. *Nucleic acids research*, page gkr1088, 2011.

[62] Akira Kikuchi, Daisuke Hasegawa, Yoshitoshi Ohtsuka, Kazuko Hamamoto, Seiji Kojima, Jun Okamura, Tatsutoshi Nakahata, and Atsushi Manabe. Outcome of children with refractory anaemia with excess of blast (raeb) and raeb in transformation (raeb-t) in the japanese mds99 study. *British journal of haematology*, 158(5):657–661, 2012.

[63] Jae-Sung Kim, Zae Young Ryoo, and Jang-Soo Chun. Cytokine-like 1 (cytl1) regulates the chondrogenesis of mesenchymal cells. *Journal of Biological Chemistry*, 282(40):29359–29367, 2007.

[64] Seon-Young Kim and David J Volsky. Page: parametric analysis of gene set enrichment. *BMC bioinformatics*, 6(1):144, 2005.

[65] Andrew W Knight and Timothy P Senior. The common problem of rare disease in general practice. *Medical Journal of Australia*, 185(2):82, 2006.

[66] Sebastian Köhler, Sandra C Doelken, Christopher J Mungall, Sebastian Bauer, Helen V Firth, Isabelle Bailleul-Forestier, Graeme CM Black, Danielle L Brown, Michael Brudno, Jennifer Campbell, et al. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(D1):D966–D974, 2014.

[67] Juha Kononen, Lukas Bubendorf, Anne Kallionimeni, Maarit Bärlund, Peter Schraml, Stephen Leighton, Joachim Torhorst, Michael J Mihatsch, Guido Sauter, and Olli-P Kallionimeni. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature medicine*, 4(7):844–847, 1998.

[68] Martin Krallinger, Florian Leitner, and Alfonso Valencia. Analysis of biological processes and diseases using text mining approaches. In *Bioinformatics Methods in Clinical Research*, pages 341–382. Springer, 2010.

[69] Andrea Kuendgen, Sabine Knipp, Frank Fox, Corinna Strupp, Barbara Hildebrandt, Christian Steidl, Ulrich Germing, Rainer Haas, and Norbert Gattermann. Results of a phase 2 study of valproic acid alone or in combination with all-trans retinoic acid in 75 patients with myelodysplastic syndrome and relapsed or refractory acute myeloid leukemia. *Annals of hematology*, 84(1):61–66, 2005.

[70] Heung Sun Kwon, Naoki Nakaya, Mones Abu-Asab, Hong Sug Kim, and Stanislav I Tomarev. Myocilin is involved in ngr1/lingo-1-mediated oligodendrocyte differentiation and myelination of the optic nerve. *The Journal of Neuroscience*, 34(16):5539–5551, 2014.

[71] Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr,

Eric S. Lander, and Todd R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006.

[72] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[73] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biol*, 10(3):R25, 2009.

[74] I Lauder and W Aherne. The significance of lymphocytic infiltration in neuroblastoma. *British journal of cancer*, 26(4):321, 1972.

[75] Steffen L Lauritzen and Nanny Wermuth. Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, pages 31–57, 1989.

[76] Insuk Lee, U Martin Blom, Peggy I Wang, Jung Eun Shim, and Edward M Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research*, 21(7):1109–1121, 2011.

[77] Young-suk Lee, Arjun Krishnan, Qian Zhu, and Olga G Troyanskaya. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29(23):3036–3044, 2013.

[78] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[79] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[80] Craig L Leonardi, Jerold L Powers, Robert T Matheson, Bernard S Goffe, Ralph Zitnik, Andrea Wang, and Alice B Gottlieb. Etanercept as monotherapy in patients with psoriasis. *New England journal of medicine*, 349(21):2014–2022, 2003.

[81] Bo Li, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 2010.

[82] Tao Li, Chengliang Zhang, and Mitsunori Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, 2004.

[83] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, et al. Mint, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(D1):D857–D861, 2012.

[84] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[85] Catriona Y Logan and Roel Nusse. The wnt signaling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.*, 20:781–810, 2004.

[86] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[87] Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen, and Alvis Brazma. A global map of human gene expression. *Nature biotechnology*, 28(4):322–324, 2010.

[88] Luca Malcovati and Mario Cazzola. Refractory anemia with ring sideroblasts. *Best Practice & Research Clinical Haematology*, 26(4):377–385, 2013.

[89] Florent M Martin, Gabriela Bydlon, and Jeffrey S Friedman. Sod2-deficiency sideroblastic anemia and red blood cell oxidative stress. *Antioxidants & redox signaling*, 8(7-8):1217–1225, 2006.

[90] Smitha Mathews, Suja Ann Mathew, Pawan Kumar Gupta, Ramesh Bhonde, and Satish Totey. Glycosaminoglycans enhance osteoblast differentiation of bone marrow derived human mesenchymal stem cells. *Journal of tissue engineering and regenerative medicine*, 8(2):143–152, 2014.

[91] Matthew N. McCall, Benjamin M. Bolstad, and Rafael A. Irizarry. Frozen robust multiarray analysis (frma). *Biostatistics*, 11(2):242–253, 2010.

[92] Matthew N. McCall, Karan Uppal, Harris A. Jaffee, Michael J. Zilliox, and Rafael A. Irizarry. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(suppl 1):D1011–D1015, 2011.

[93] Hans-Werner Mewes, Dmitrij Frishman, Christian Gruber, Birgitta Geier, Dirk Haase, Andreas Kaps, Kai Lemcke, Gertrud Mannhaupt, Friedhelm Pfeiffer, C Schüller, et al. Mips: a database for genomes and protein sequences. *Nucleic acids research*, 28(1):37–40, 2000.

[94] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[95] Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, et al. A survey of genetic human cortical gene expression. *Nature genetics*, 39(12):1494–1499, 2007.

[96] Chad L Myers and Olga G Troyanskaya. Context-sensitive data integration and prediction of biological networks. *Bioinformatics*, 23(17):2322–2330, 2007.

[97] T. Nakazawa, K. Ohashi, M. Yamada, S. Shinoda, F. Saji, Y. Murata, and H. Araki. Effect of different concentrations of amino acids in human serum and follicular fluid on the development of one-cell mouse embryos in vitro. *Journal of Reproduction and Fertility*, 111(2):327–332, 1997.

[98] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[99] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[100] Christopher Y Park, David C Hess, Curtis Huttenhower, and Olga G Troyanskaya. Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components. *PLoS Comput Biol*, 6(11):e1001009, 2010.

[101] Christopher Y Park, Aaron K Wong, Casey S Greene, Jessica Rowland, Yuanfang Guan, Lars A Bongo, Rebecca D Burdine, and Olga G Troyanskaya. Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput Biol*, 9(3):e1002957, 2013.

[102] Leo S Payne and Paul H Huang. The pathobiology of collagens in glioma. *Molecular Cancer Research*, 11(10):1129–1140, 2013.

[103] Mark Peifer and Paul Polakis. Wnt signaling in oncogenesis and embryogenesis–a look outside the nucleus. *Science*, 287(5458):1606–1609, 2000.

[104] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999.

[105] TS Keshava Prasad, Kumaran Kandasamy, and Akhilesh Pandey. Human protein reference database and human proteinpedia as discovery tools for systems biology. In *Reverse Chemical Genetics*, pages 67–79. Springer, 2009.

[106] Kim D Pruitt, Garth R Brown, Susan M Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M Farrell, Jennifer Hart, Melissa J Landrum, Kelly M McGarvey, et al. Refseq: an update on mammalian reference sequences. *Nucleic acids research*, 42(D1):D756–D763, 2014.

[107] John Quackenbush. Computational analysis of microarray data. *Nature reviews genetics*, 2(6):418–427, 2001.

[108] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P. Mesirov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S. Lander, and Todd R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

[109] Daniel R Rhodes, Jianjun Yu, Kalyan Shanker, Nandan Deshpande, Radhika Varambally, Debashis Ghosh, Terrence Barrette, Akhilesh Pander, and Arul M Chinnaiyan. Oncomine: a cancer microarray database and integrated data-mining platform. *Neoplasia*, 6(1):1–6, 2004.

[110] Roland Rosmond, Monique Chagnon, Claude Bouchard, and Per Bjorntorp. G-308a polymorphism of the tumor necrosis factor α gene promoter and salivary cortisol secretion 1. *The Journal of Clinical Endocrinology & Metabolism*, 86(5):2178–2180, 2001.

[111] Richard B Roth, Peter Hevezi, Jerry Lee, Dorian Willhite, Sandra M Lechner, Alan C Foster, and Albert Zlotnik. Gene expression analyses reveal molecular relationships among 20 regions of the human cns. *Neurogenetics*, 7(2):67–80, 2006.

[112] Nathaniel Rothman, Christine F Skibola, Sophia S Wang, Gareth Morgan, Qing Lan, Martyn T Smith, John J Spinelli, Eleanor Willett, Silvia De Sanjose, Pierluigi Cocco, et al. Genetic variation in tnf and il10 and risk of non-hodgkin lymphoma: a report from the interlymph consortium. *The lancet oncology*, 7(1):27–38, 2006.

[113] Johan Rung and Alvis Brazma. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 14(2):89–99, 2013.

[114] Chris B Russell, Hugh Rand, Jeannette Bigler, Keith Kerkof, Martin Timour, Edgar Bautista, James G Krueger, David H Salinger, Andrew A Welcher, and David A Martin. Gene expression profiles normalized in psoriatic skin by treatment with brodalumab, a human anti–il-17 receptor monoclonal antibody. *The Journal of Immunology*, 192(8):3828–3836, 2014.

[115] Gabriella Rustici, Nikolay Kolesnikov, Marco Brandizi, Tony Burdett, Miroslaw Dylag, Ibrahim Emam, Anna Farne, Emma Hastings, Jon Ison, Maria Keays, Natalja Kurbatova, James Malone, Roby Mani, Annalisa Mupo, Rui Pedro Pereira, Ekaterina Pilicheva, Johan Rung, Anjan Sharma, Y. Amy Tang, Tobias Ternent, Andrew Tikhonov, Danielle Welter, Eleanor Williams, Alvis Brazma, Helen Parkinson, and Ugis Sarkans. Arrayexpress updatetrends in database growth and links to data analysis tools. *Nucleic Acids Research*, 41(D1):D987–D990, 2013.

[116] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl 1):D91–D94, 2004.

[117] Vijay G Sankaran and Mitchell J Weiss. Anemia: progress in molecular mechanisms and therapies. *Nature medicine*, 21(3):221–230, 2015.

[118] Noboru Sato, Laurent Meijer, Leandros Skaltsounis, Paul Greengard, and Ali H Brivanlou. Maintenance of pluripotency in human and mouse embryonic stem cells through activation of wnt signaling by a pharmacological gsk-3-specific inhibitor. *Nature medicine*, 10(1):55–63, 2004.

[119] Michael C Schatz, Ben Langmead, and Steven L Salzberg. Cloud computing and the dna data race. *Nature biotechnology*, 28(7):691, 2010.

[120] Patrick R. Schmid, Nathan P. Palmer, Isaac S. Kohane, and Bonnie Berger. Making sense out of massive data by going beyond differential expression. *Proceedings of the National Academy of Sciences*, 109(15):5594–5599, 2012.

[121] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[122] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[123] Alex D Sheftel, Des R Richardson, Josef Prchal, and Prem Ponka. Mitochondrial iron metabolism and sideroblastic anemia. *Acta haematologica*, 122(2-3):120–133, 2009.

[124] Yasuaki Shirayoshi, TS Okada, and Masatoshi Takeichi. The calcium-dependent cell-cell adhesion system regulates inner cell mass formation and cell surface polarization in early mouse development. *Cell*, 35(3):631–638, 1983.

[125] Radha Shyamsundar, Young H Kim, John P Higgins, Kelli Montgomery, Michelle Jorden, Anand Sethuraman, Matt van de Rijn, David Botstein, Patrick O Brown, and Jonathan R Pollack. A dna microarray survey of gene expression in normal human tissues. *Genome biology*, 6(3):R22, 2005.

[126] Noah Simon and Robert Tibshirani. Comment on" detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. *arXiv preprint arXiv:1401.7645*, 2014.

[127] Rajasekharan Somasundaram and Dorothee Herlyn. Chemokines and the microenvironment in neuroectodermal tumor–host interaction. In *Seminars in cancer biology*, volume 19, pages 92–96. Elsevier, 2009.

[128] Swati Sood and Radhika Srinivasan. Alterations in gene promoter methylation and transcript expression induced by cisplatin in comparison to 5-azacytidine in hela and siha cervical cancer cell lines. *Molecular and cellular biochemistry*, 404(1-2):181–191, 2015.

[129] Paul T Spellman, Gavin Sherlock, Michael Q Zhang, Vishwanath R Iyer, Kirk Anders, Michael B Eisen, Patrick O Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 1998.

[130] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, et al. The biogrid interaction database: 2011 update. *Nucleic acids research*, 39(suppl 1):D698–D704, 2011.

[131] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[132] Silpa Suthram, Joel T Dudley, Annie P Chiang, Rong Chen, Trevor J Hastie, and Atul J Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput Biol*, 6(2):e1000662, 2010.

[133] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[134] Paul C Tang and David Lansky. The missing link: bridging the patient–provider health information gap. *Health Affairs*, 24(5):1290–1295, 2005.

[135] Ronald C Taylor. An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. *BMC bioinformatics*, 11(Suppl 12):S1, 2010.

[136] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[137] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[138] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[139] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[140] Bruna T Tuan, Marília B Visacri, Laís S Amaral, Daniele Baldini, Graziele B Ferrari, Júlia CF Quintanilha, Eder C Pincinato, Priscila G Mazzola, Carmen SP Lima, and Patricia Moriel. Effects of high-dose cisplatin chemotherapy and conventional radiotherapy on urinary oxidative and nitrosative stress biomarkers in patients with head and neck cancer. *Basic & clinical pharmacology & toxicology*, 118(1):83–86, 2016.

[141] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[142] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, et al. Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250, 2010.

[143] Laura Vera-Ramirez, Pedro Sanchez-Rovira, Cesar L Ramirez-Tortosa, Jose L Quiles, M Ramirez-Tortosa, and Jose A Lorente. Transcriptional shift identifies a set of genes driving breast cancer chemoresistance. *PloS one*, 8(1):e53983, 2013.

[144] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[145] Annika M Whittle, Sylvia Feyler, and David T Bowen. Durable second complete remissions with oral melphalan in hypocellular acute myeloid leukemia and refractory anemia with excess blast with normal karyotype relapsing after intensive chemotherapy. *Leukemia research reports*, 2(1):9–11, 2013.

[146] Aaron K Wong, Christopher Y Park, Casey S Greene, Lars A Bongo, Yuanfang Guan, and Olga G Troyanskaya. Imp: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic acids research*, 40(W1):W484–W490, 2012.

[147] Kai Xia, Andrey A Shabalin, Shunping Huang, Vered Madar, Yi-Hui Zhou, Wei Wang, Fei Zou, Wei Sun, Patrick F Sullivan, and Fred A Wright. seeqtl: a searchable database for human eqtls. *Bioinformatics*, 28(3):451–452, 2012.

[148] Kuender D Yang, Men-Fang Shaio, Chih-Lu Wang, Nai-Cheng Wu, and Richard M Stone. Neuroblastoma cell-mediated leukocyte chemotaxis: lineage-specific differentiation of interleukin-8 expression. *Experimental cell research*, 211(1):1–5, 1994.

[149] Mi Zhou, Joseph L Wiemels, Paige M Bracci, Margaret R Wrensch, Lucie S Mccoy, Terri Rice, Jennette D Sison, Joseph S Patoka, and John K Wiencke. Circulating levels of the innate and humoral immune regulators cd14 and cd23 are associated with adult glioma. *Cancer research*, 70(19):7534–7542, 2010.

[150] Qian Zhu, Aaron K Wong, Arjun Krishnan, Miriam R Aure, Alicja Tadych, Ran Zhang, David C Corney, Casey S Greene, Lars A Bongo, Vessela N Kristensen, et al. Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nature methods*, 12(3):211–214, 2015.

[151] Michael J Zilliox and Rafael A Irizarry. A gene expression bar code for microarray data. *Nature methods*, 4(11):911–913, 2007.