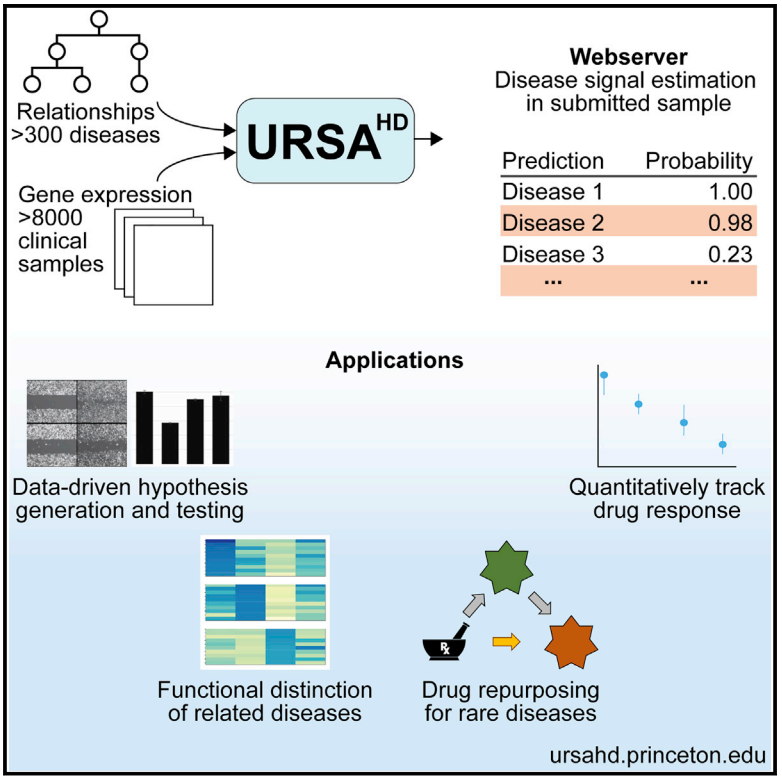
Article

Cell Systems

A Computational Framework for Genome-wide Characterization of the Human Disease Landscape

Graphical Abstract



Highlights

- URSA^{HD} integrates >8,000 clinical gene expression profiles across >300 diseases
- Identifies unique characteristics for each disease in a data-driven manner
- Enables data-driven targeted research even for rare and understudied diseases
- Tracks therapeutic drug response in expression profiles from disease samples

Authors

Young-suk Lee, Arjun Krishnan, Rose Oughtred, ..., Kara Dolinski, Chandra L. Theesfeld, Olga G. Troyanskaya

Correspondence

chandrat@princeton.edu (C.L.T.), ogt@cs.princeton.edu (O.G.T.)

In Brief

Discovering unique properties among diseases is needed to develop targeted treatments, especially for related disorders. To address this, we developed a unified framework, URSA^{HD}, which leverages physiological relationships between diseases and integrates thousands of clinical samples across >300 diseases to identify distinct characteristics that can be used to guide biomedical research. We demonstrate applications of URSA^{HD}, including guiding hypothesis generation and experiments, drug repurposing, and quantitatively tracking drug response.

Lee et al., 2019, Cell Systems *8*, 152–162
February 27, 2019
https://doi.org/10.1016/j.cels.2018.12.010

CellPress

CellPress

Cell Systems
Article

A Computational Framework for Genome-wide Characterization of the Human Disease Landscape

Young-suk Lee,[1,2,3] Arjun Krishnan,[1,4] Rose Oughtred,[1] Jennifer Rust,[1] Christie S. Chang,[1] Joseph Ryu,[1] Vessela N. Kristensen,[5,6,7] Kara Dolinski,[1] Chandra L. Theesfeld,[1,*] and Olga G. Troyanskaya[1,2,8,9,*]

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
[2]Department of Computer Science, Princeton University, Princeton, NJ, USA
[3]School of Biological Sciences, Seoul National University, Seoul, South Korea
[4]Departments of Computational Mathematics, Science, and Engineering and Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA
[5]Department of Genetics, Institute of Cancer Research, Oslo University Hospital, Radiumhospitalet, Oslo, Norway
[6]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[7]Department of Clinical Molecular Biology (EpiGen), Division of Medicine, Akershus University Hospital, Lørenskog, Norway
[8]Flatiron Institute, Simons Foundation, New York, NY, USA
[9]Lead Contact
*Correspondence: chandrat@princeton.edu (C.L.T.), ogt@cs.princeton.edu (O.G.T.)
https://doi.org/10.1016/j.cels.2018.12.010

SUMMARY

A key challenge for the diagnosis and treatment of complex human diseases is identifying their molecular basis. Here, we developed a unified computational framework, URSA^{HD} (Unveiling RNA Sample Annotation for Human Diseases), that leverages machine learning and the hierarchy of anatomical relationships present among diseases to integrate thousands of clinical gene expression profiles and identify molecular characteristics specific to each of the hundreds of complex diseases. URSA^{HD} can distinguish between closely related diseases more accurately than literature-validated genes or traditional differential-expression-based computational approaches and is applicable to any disease, including rare and understudied ones. We demonstrate the utility of URSA^{HD} in classifying related nervous system cancers and experimentally verifying novel neuroblastoma-associated genes identified by URSA^{HD}. We highlight the applications for potential targeted drug-repurposing and for quantitatively assessing the molecular response to clinical therapies. URSA^{HD} is freely available for public use, including the use of underlying models, at ursahd.princeton.edu.

INTRODUCTION

A primary goal of current biomedical research is the detailed characterization of the molecular basis of complex human diseases to enable precise diagnosis and treatment. This requires genome-scale approaches and the ability to distinguish among diseases that fall along a continuous landscape of molecular changes. Many human disease studies have used gene expression profiles to systematically quantify and compare genome-wide changes between healthy/normal samples and patient samples for a number of different diseases (Dunckley et al., 2006; Gomez Ravetti et al., 2010; Hodges et al., 2006; Moran et al., 2006; Das and Rao, 2007; Kawakami et al., 2006; Namløs et al., 2012). The resulting differential mRNA abundance of thousands of genes captures the genome-wide perturbations of genes and pathways that underlie the disease of interest compared to normal tissue. However, complex diseases share other underlying genetic and functional changes, with only some perturbations being unique to a specific disease. For example, in autoimmune diseases Sjögren syndrome (SS) and systemic lupus erythematosus (SLE), blood profiling experiments revealed remarkably similar upregulation of type 1 IFN-inducible genes in clinical samples for both conditions, yet there are clear distinctions in clinical presentations suggesting important differences in etiology undetectable with the traditional approaches (Pascual et al., 2010). From the myriad of observed expression changes, it is impossible to tease apart those unique to a single disease when the analysis of disease gene expression is done in isolation.

Integration of individual genome-wide expression studies offers a promising path toward a better understanding of the distinct characteristics of multiple human diseases. Several efforts have taken a meta-analysis approach and integrated the differential expression analyses of individual studies (Huang et al., 2010; Suthram et al., 2010). However, such approaches neglect the relationships among complex diseases and inadvertently pass over distinctive signatures unique to a single disease in the disease continuum. Early efforts at examining multiple diseases holistically demonstrated the promise of such analyses (Amar et al., 2015; Schmid et al., 2012), but they were limited to a single expression platform (Schmid et al., 2012) in disease coverage (Amar et al., 2015) and scale (Amar et al., 2015; Schmid et al., 2012). Therefore, a unified scalable 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 (i.e., covering a large number of diseases) and data-driven (i.e., taking advantage of thousands of clinical gene expression

datasets) in order to uncover subtle differences between similar diseases and highlight identifiable aspects of even rare diseases.

Here, we present URSA^{HD} (Unveiling RNA Sample Annotation for Human Diseases), a systematic framework that utilizes thousands of clinical samples and explicitly identifies the distinctive molecular characteristics of 335 human diseases. Leveraging the large gene expression data compendium and the hierarchy of anatomical relationships of human diseases, URSA^{HD} uses machine learning to build individual disease-specific models and integrates them in a probabilistic framework to provide hierarchically consistent estimates of disease signals. This approach does not rely on literature-based disease-gene associations and overcomes patient, tissue, dataset, and profiling-technology biases. We demonstrate that URSA^{HD} outperforms other approaches using individual disease genes or traditional normal/disease differential expression analysis in quantifying disease signals. In addition to accurately detecting and characterizing disease in any sample, URSA^{HD} provides interpretable genome-wide models that enable insights into the biological processes, pathways, and tissues underlying each prediction. These can be applied to addressing central challenges in research and drug development, including guiding hypothesis generation and experimentation, associating molecularly similar diseases as the first step for drug repurposing, and quantitively tracking responses to drug therapy in disease samples so as to differentiate responders from non-responders. URSA^{HD} is implemented in a publicly available, user-friendly web server at ursahd.princeton.edu, where biomedical researchers can submit their gene expression data to obtain data-driven quantification of disease signals.

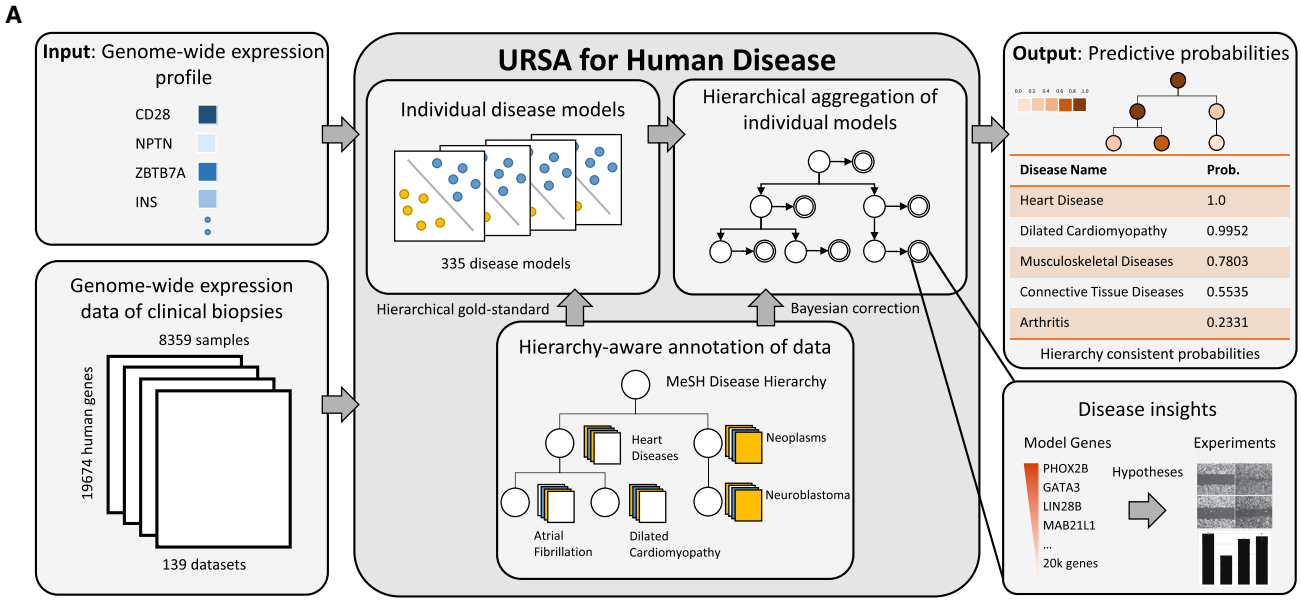
RESULTS

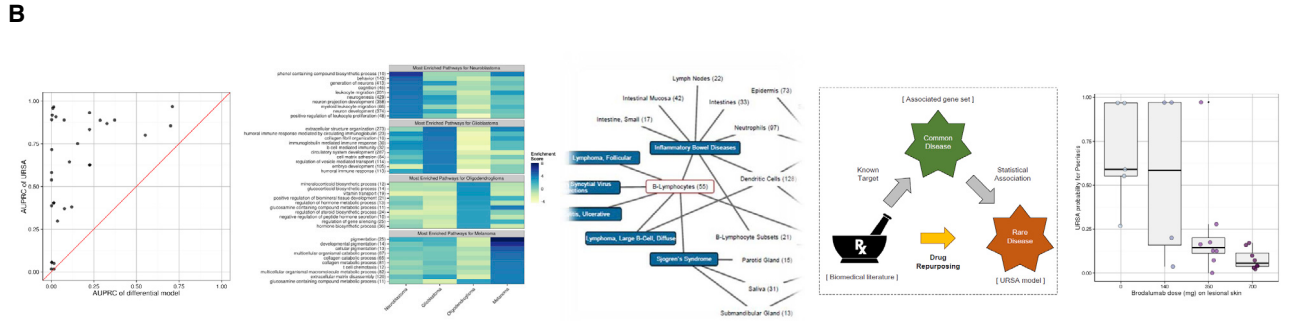
Accurate Data-Driven Characterization of the Human Disease Landscape

To characterize both distinctive and shared molecular processes underlying each human disease (Figure S1A), we developed a data-driven approach leveraging a large compendium of clinical expression data. First, we manually annotated the disease information for 8,359 gene expression experiments of clinical biopsies (both disease and normal samples) from 139 datasets in the Gene Expression Omnibus (GEO) (Barrett et al., 2013). This data compendium spans 335 disease terms after propagation along the MeSH hierarchy (Figure S1B; Table S1; STAR Methods). To characterize each human disease, we first computationally model the distinctive, genome-wide features of each individual disease by contrasting expression profiles of clinical disease samples to that of all matched normal samples, all other non-matched normal samples, and all other disease samples (Figure 1A). URSA^{HD} then integrates these individual disease models into a unified Bayesian network based on the structure of the MeSH hierarchy, which is loosely reflective of the anatomical relationships among diseases, to provide hierarchically consistent estimates of the specific disease signal (Figure 1B). These data-driven models implicitly identify and "up-weight" genes whose expression differentiates the profiles of the disease of interest from all other samples in the compendium (all normal, matched, and other tissues, and all other disease samples). The models, at the same time, shrink weights of the non-discriminative genes (see STAR Methods). Through this data-driven approach, we detect complex disease signals in gene expression profiling assays, explore the landscape of human diseases, and provide tools that can help biomedical researchers address unsolved problems for therapeutic drug development (Figure 1B).

The disease state of a given clinical sample is often inferred from the expression of a single disease marker gene that is considered representative of that particular disease state. While the marker's expression hints at abnormal molecular changes in the underlying tissue or cell type, many known disease-associated genes are not exclusive to individual diseases. Thus, while effective at distinguishing a disease sample from the corresponding normal tissue, these marker genes often cannot distinguish among related diseases. Tumor necrosis factor (TNF), for example, is a very common marker gene being documented in the literature as associated, in some way, with 96 different human diseases as diverse as psoriasis, non-Hodgkin lymphoma, and obesity (Table S2) (Bonifati and Ameglio, 1999; Leonardi et al., 2003; Rosmond et al., 2001; Rothman et al., 2006). URSA^{HD} models, in contrast, can detect disease signal by taking advantage of all genes in a gene expression profile. We tested the accuracy of URSA^{HD} in classifying gene expression samples to specific diseases or healthy tissues and compared its performance to the top scoring single disease-associated gene for that disease (as defined by gene2mesh; Ade et al., 2007). We performed a stringent evaluation to measure URSA^{HD} accuracy using diseases with at least two independent data sets available (32 diseases), allowing us to hold out a whole dataset during training and demonstrate generalizability of the models. URSA^{HD} outperformed the best documented single genes for 30 of the 32 diseases (Figure 2A; STAR Methods). Furthermore, for 75% of these diseases, URSA^{HD} models were over 10-fold more accurate than the best-performing documented disease gene.

A key advantage of URSA^{HD} models over published individual gene markers is access to the full genome-wide array of expression signals, without relying only on a single gene or previously published markers. Differential expression analysis between disease and matched healthy samples is the traditional genome-wide approach for identifying disease-associated genes. We thus compared the accuracy of URSA^{HD} predictions with that of a differential expression approach. URSA^{HD} outperformed the differential model for all 32 tested diseases on an independent holdout test set (Figure 2B; STAR Methods). This difference in performance is likely due to two factors: first, not all disease-specific genes are differentially expressed in a statistically significant manner, and second, traditional differential expression models are focused on distinguishing between matched normal and disease samples without considering the broader tissue and disease context. For example, both the URSA^{HD} and the differential expression models for dilated cardiomyopathy were enriched for heart-related anatomical MeSH terms such as left ventricular hypertrophy (MESH:D017379), atrial fibrillation (MESH:D001281), and heart atria (MESH:D006325). However, performance of the two models was drastically different: AUPRC of URSA^{HD} = 0.9069, AUPRC of differential model = 0.0752, and background random model AUPRC = 0.0074. This is likely because 30 dilated cardiomyopathy genes documented in the literature were not differentially expressed ($z = -0.155$)





① Accurate detection of human disease signals

② Distinctive functional characterization of related human diseases

③ Anatomical context of human disease landscape

④ Drug-repurposing for rare human diseases

⑤ Tracking drug response

Figure 1. Overview of Method and Applications of URSA^{HD} for Biomedical Research

(A) URSA^{HD} integrates gene expression profiles for 8,359 diseased 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. The resulting models provide accurate and sensitive disease characterization of any microarray or RNA-Seq expression profile and enable disease insights for experimental follow-up. Note that no feature selection method or curated gene sets are used in our approach.

(B) We demonstrate the applications of URSA^{HD} for (1) accurate disease signal detection, (2) specific functional and (3) anatomical characterization of each disease, (4) repurposing known drugs for the treatment of rare human diseases, and (5) tracking therapeutic drug effects using gene expression profiles.

but were enriched in the URSA^{HD} model ($z = 4.326$). Neither approach used any prior literature information. This trend persists across other human diseases, including rare diseases (Figure S2A). Furthermore, without retraining the models, the URSA^{HD} predictions were also consistent and accurate for RNA-seq samples in the Cancer Genome Atlas (TCGA) (TCGA data were not used in training URSA^{HD}), further demonstrating the accuracy and utility of URSA^{HD}, independent of the profiling platform (Figure S2C; Table S3).

Interpretation of Disease Signals in URSA^{HD} Models

To demonstrate the practical use of the model information for uncovering the pathophysiology of human diseases, we examined the biological relevance of the top-weighted genes in the URSA^{HD} model for neuroblastoma (Table S4). Of the top 20 genes, 16 are verified in the literature to be involved in neuroblastoma: two (PHOX2B and LIN28B) are known susceptibility genes for neuroblastoma, 11 are known to be highly expressed and/or serve as biomarkers for diagnosis and prognosis, and three are

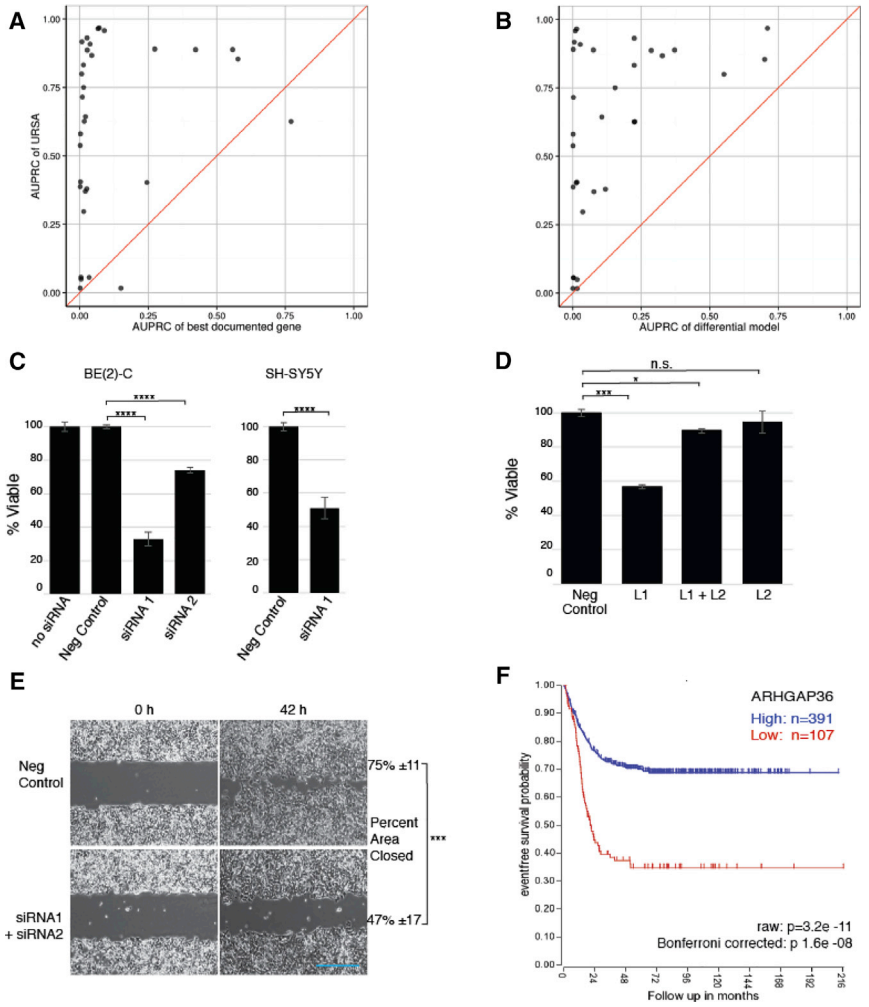


Figure 2. URSA^{HD} Accurately Detects Human Disease Signals in Gene Expression Profiles

(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 performance of these methods for a specific disease. Red line is the identity line; dots above the red line indicate diseases for which URSA^{HD} outperforms the documented disease gene method.

(B) URSA^{HD} outperforms traditional genome-wide differential expression approach (normal/disease). Scatterplot of AUPRC of URSA^{HD} (y-axis) and traditional differential expression approach (x-axis). Each dot represents the performance of these methods for a specific disease. Red line is the identity line; dots above the red line indicate diseases for which URSA^{HD} outperforms the differential expression approach.

(C) MAB21L1 (Genbank: 4081) is required for the viability of human neuroblastoma cells (BE(2)-C, MYCN amplified, and SH-SY5Y, normal MYCN). Error bars are SE from three replicates (unpaired t test:****$p < 0.001$).

(D) MAB21L2 (Genbank: 10586) activity may oppose MAB21L1. The growth defect incurred by knockdown of MAB21L1 is suppressed by further knockdown of MAB21L2. Cells were transfected with siRNA against MAB21L1 (L1, 1.25 pmol/well), both (L1 + L2, 1.25 pmol/well and 5 pmol/well, respectively), or MAB21L2 (L2, 5 pmol/well) and viability was measured. Error bars are SE (unpaired t test:*$p < 0.05$, ***$p = 0.001$).

(E) Loss of ARHGAP36 (Genbank: 158763) leads to a strong defect in migration. Cells were reverse-transfected with Negative Control siRNA (10 pmol/well) or siRNA1+ siRNA2 (5 pmol/well, each) against ARHGAP36, and then Ibidi silicon inserts were pulled 24 h after transfection to initiate gap formation. Average percent open area was used to measure cellular migration (unpaired t test: ***$p = 0.0063$, 95% CI.) Three control wells and six knockdown wells were followed. The experiment was repeated with similar results. Blue scale bar is 500 μm.

(F) Kaplan-Meyer survival analysis: high ARHGAP36 (NM_144967) expression levels are associated with improved survival for neuroblastoma patients (RNA-Seq dataset GEO: GSE49710, n = 498).

associated with embryonic and nervous system development with relevance to neuroblastoma (Table S4). We experimentally investigated the four remaining top-weighted genes, which had no prior association with neuroblastoma: MAB21L1, MAB21L2, ARHGAP36, and LOC105007194. Using RT-PCR, we first confirmed the expression of all four genes in BE(2)-C cells, a human MYCN-amplified neuroblastoma cell line (Figure S3). In these cells, knockdown of MAB21L1 with two different siRNAs led to a growth phenotype: 35% viable (siRNA1) and 77% viable (siRNA2) (Figure 2C, left panel; Figures S3A and S3B). This phenotype was corroborated in a second, non-MYCN-amplified, human neuroblastoma cell line, SH-SY5Y (Figure 2C, right panel). MAB21L2, a MAB21L1 homolog, did not have a growth defect on its own, despite knockdown of mRNA to a similar extent as MAB21L1 (Figure 2D). Subsequently, we tested if the depletion of both MAB21L1 and MAB21L2 would lead to a larger effect than MAB21L1 alone. However, knockdown of MAB21L2 suppressed the growth defect of MAB21L1, suggesting antagonistic functions for this pair of gene products in neuroblastoma proliferation.

ARHGAP36 is a putative Rho GTPase-activating protein (GAP) that activates the expression of Hippo pathway transcription factors but its pathophysiological role in neuroblastoma remains to be elucidated (Rack et al., 2014). Since Rho-GAP proteins are known to regulate actin organization and cell migration (Jaffe and Hall, 2005), we tested whether cell migration might be perturbed when ARHGAP36 expression was disrupted. To this end, we used Ibidi silicon inserts to generate uniform gaps, and simultaneously plated and transfected BE(2)-C cells with two siRNAs at low concentrations to knockdown all isoforms of ARHGAP36 (Figure S3C). 42 h following the removal of the inserts, we found a 45%±17 deficit in gap closure in ARHGAP36 depleted cells (Figure 2E, n = 6, p value = 0.0063). Even after 72 h, the gaps were not closed in the ARHGAP36 knockdown wells, whereas all control-transfected gaps were completely filled, indicating that ARHGAP36 promotes migration

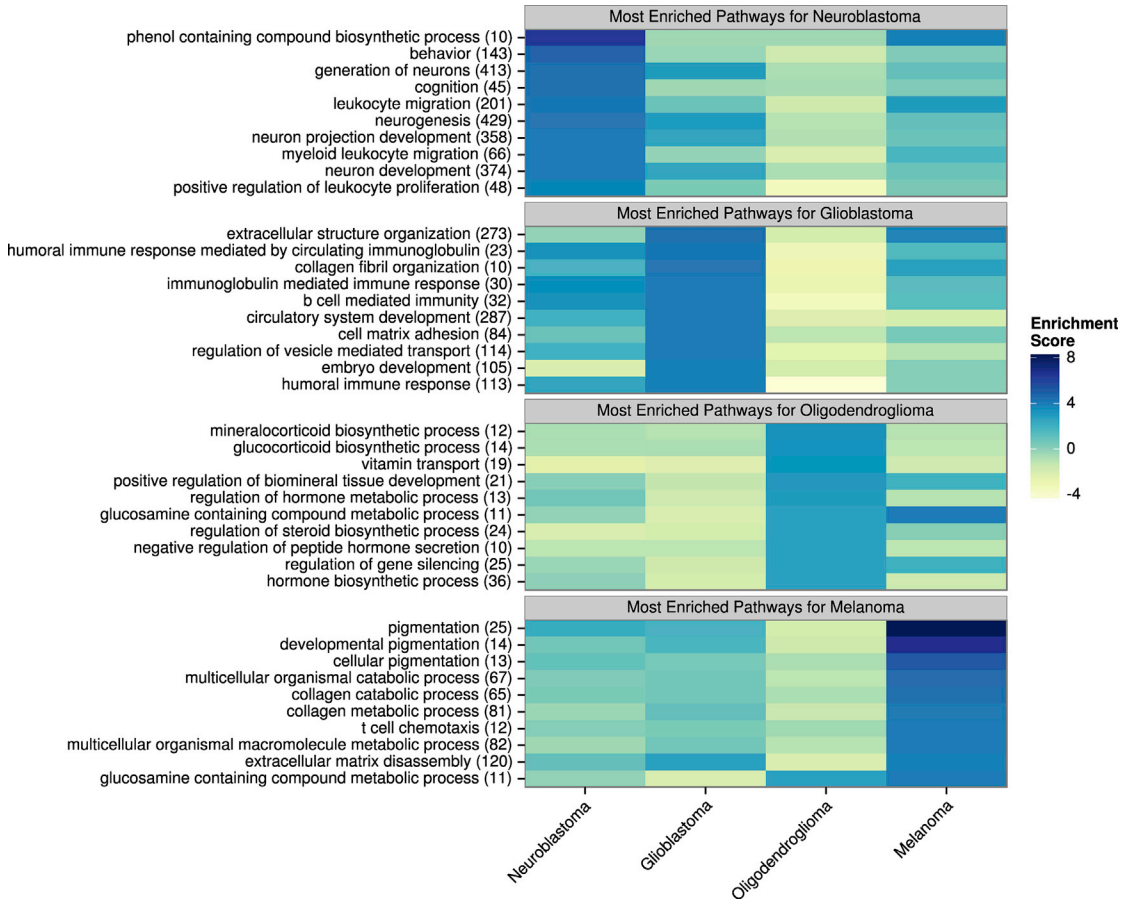


Figure 3. Distinctive Functional Characterization of Neuroblastoma and Related Diseases of Ectodermal Origin

Each heatmap summarizes the top functional (GO biological process) enrichments in the URSA^{HD} models for neuroblastoma, glioblastoma, oligodendroglioma, and melanoma. Each ectodermal disease is characterized by their distinctive functional associations. See also Table S5.

in neuroblastoma cells. This function may be protective *in vivo*, as high expression of ARHGAP36 was associated with increased survival among neuroblastoma patients (GEO:GSE49710) (Bonferoni-corrected p value = 1.6e^{-05}; Figure 2F). These preliminary experiments are consistent with possible roles for these genes in neuroblastoma—genes that were not previously linked to neuroblastoma but are among the top-weighted URSA^{HD} model genes for neuroblastoma. The URSA^{HD} model gene weights are available on the website for use by researchers in studying their disease of interest.

We next examined how well URSA^{HD} models resolve similar diseases. Neuroblastoma, glioblastoma, oligodendroglioma, and melanoma all originate from ectodermal tissues and exhibit similar characteristics of tumor-host interaction at the molecular level (Somasundaram and Herlyn, 2009). Nonetheless, each ectodermal disease exhibits unique signs and symptoms used for diagnosis, targeted treatment, and prognosis. URSA^{HD} models for the four ectodermal diseases were enriched with functional characteristics specific to each individual disease, consistent with the known literature as indicated by the associated disease MeSH term (Figure 3). For example, the URSA^{HD} model for neuroblastoma was enriched with neuron-development-related processes and leukocyte migration-related processes, recapitulating its known neural-crest-derived origins and lymphocytic infiltration (Lauder and Aherne, 1972; Yang et al., 1994). The glioblastoma model was specifically enriched with distinctive pathways relevant to its strong dysregulation of circulating immunoglobulin, extracellular matrix structure, and angiogenesis to aggressively invade the brain parenchyma (Godard et al., 2003; Kaufman et al., 2005; Payne and Huang, 2013; Zhou et al., 2010). Note that a biological process with low enrichment does not mean the process is not operational or that there is a complete lack of related gene expression: it does indicate the gene or process is less discriminative for the disease state (e.g., housekeeping genes). Interestingly, glucocorticoid metabolism-related biological processes were most enriched in the oligodendroglioma models. Such association encourages the further investigation of glucocorticoid metabolism and its role in oligodendroglioma, especially in the context of MYOC (myocilin, trabecular meshwork inducible glucocorticoid response, also known as TIGR), which is a known mediator of oligodendrocyte differentiation (Clark et al., 2001; Kwon et al., 2014). We provide the functional enrichments for the URSA^{HD} models in Table S5.

Understanding the anatomical context of each disease is crucial for accurate diagnosis and treatment of the disease.

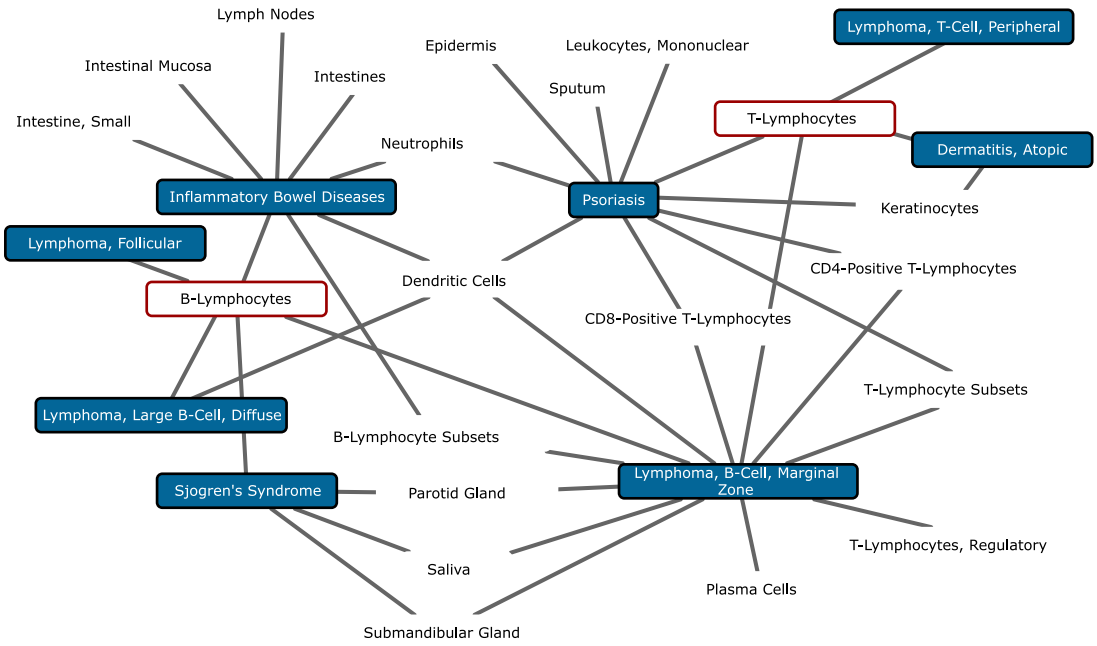


Figure 4. Anatomical Context of Diseases Associated with T Lymphocytes and B Lymphocytes (Red Borders)

Bipartite graph of disease terms (blue squares) and anatomical MeSH terms (black font). Associations based on enrichment score over 5 are shown. See also Table S5.

A unified human disease framework must account for such anatomical characteristics while focusing on specific disease signals rather than picking up general characteristics of the tissue of origin. In order to control for such bias, URSA^{HD} used corresponding normal tissue samples as negative examples to discourage any discrimination derived only from the tissue-specific differences between an unrelated disease and the disease of interest (see STAR Methods). Thus, URSA^{HD} provides a data-driven view of disease-specific tissue and cell-type association (Figure S4; Table S5). To illustrate, we show the disease model associations with B- and T-lymphocyte-specific genes (Figure 4). Mycosis fungoides, a common form of cutaneous T cell lymphoma and peripheral T cell lymphoma were exclusively and appropriately associated with T lymphocytes and not B lymphocytes. The anatomical context from inclusion body myositis (IBM)—a type of inflammatory myopathy characterized by the invasion of T cells to muscle fiber tissue—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 was used to construct the URSA^{HD} model. B lymphocyte genes were instead over-represented with B cell lymphomas such as follicular lymphoma and diffuse large B cell lymphoma (Figure 4). Autoimmune or immune-mediated pathogen diseases associated with B lymphocytes were SS, inflammatory bowel diseases, and respiratory syncytial virus infections. This separate clustering among immune-related diseases shows the distinctive, anatomical context set by the data-driven models of URSA^{HD}. See Table S5 for the complete list of the anatomical enrichment scores for all disease models.

Targeted Drug Repurposing for Rare Diseases Using the Distinctive Models of URSA^{HD}

The data-driven approach of URSA^{HD} provides a promising avenue for general drug repurposing, potentially beneficial for over 5,400 rare (or orphan) diseases (Aymé et al., 2015). Drug development for these diseases is difficult as their underlying mechanisms are not well-understood and detailed study is limited by sample numbers and financial considerations. Here, we leveraged the URSA^{HD} models, which only require expression data and thus are robust to lack of prior knowledge, to associate each rare disease to a well-studied disease with documented drug therapies. Specifically, we calculated the PAGE enrichment of the well-studied disease genes in the rare disease URSA^{HD} models (see STAR Methods) and significant associations were used to generate candidates for potential drug repurposing (Figure 5A; Table S6).

To evaluate the relevance of these associations for drug repurposing, we used pairs of diseases that have known shared therapeutic drugs documented in Comparative Toxicogenomics Database (CTD). We found diseases that share a therapeutic drug were significantly more associated than random pairs of diseases that may or may not share a drug (paired ranked-sum test, p value = 10^{-23}). These results indicate that the URSA^{HD}-based approach can provide data-driven drug repurposing hypotheses for rare diseases without relying on prior knowledge about either disease biomarkers or known drug targets.

Successful application of this approach is demonstrated by the URSA^{HD} drug repurposing predictions for two different anemias: sideroblastic anemia (SA) and refractory anemia with excess blasts (RAEB). SA and RAEB are both conditions in which the blood does not contain enough healthy red blood cells to

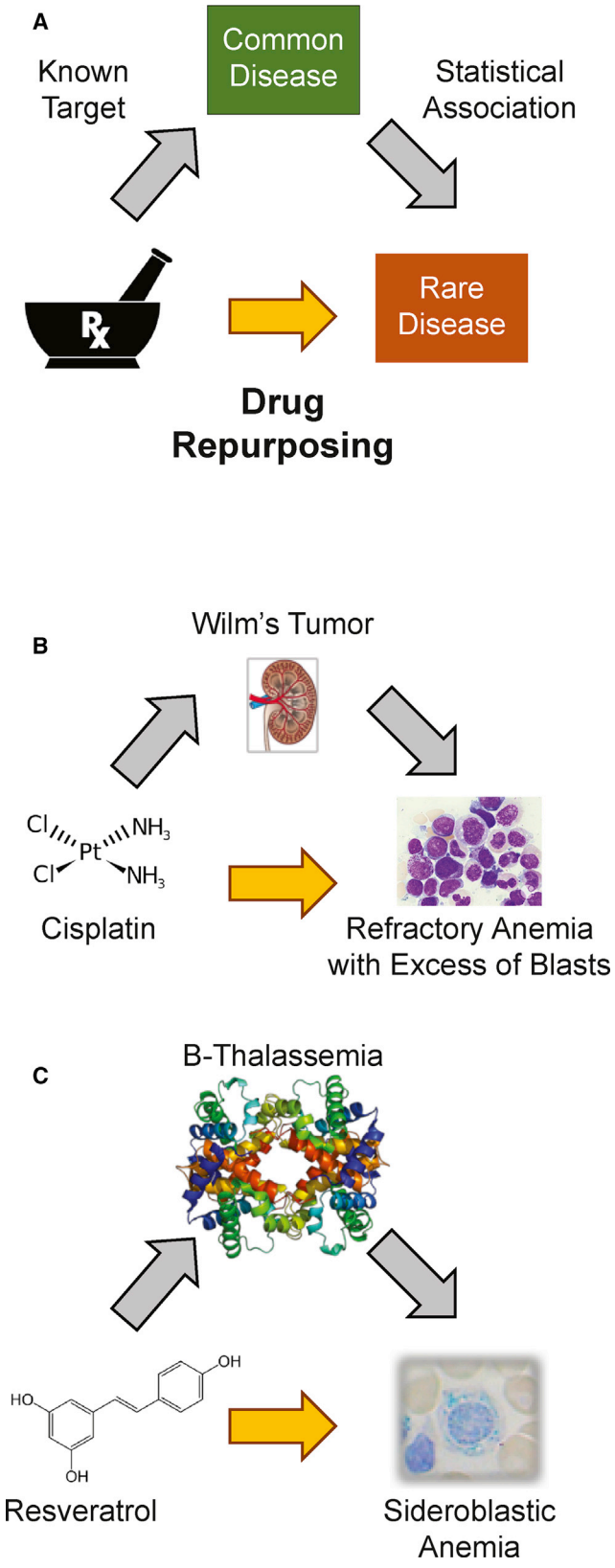


Figure 5. Targeted Drug Repurposing from Data-Driven Disease-Disease Associations

(A) Drug repurposing schematic. Known targets were taken from curated associations in the Comparative Toxicogenomics Database. Associations between well-studied diseases and rare diseases were calculated based on enrichment of common disease genes in the URSA^{HD} model for a rare disease (STAR Methods).

(B and C) Representative drug predictions for two different anemias. Both "sideroblastic anemia" and "refractory anemia with excess blasts" have very similar gene expression profiles, yet URSA^{HD} models identified distinctive biological signals and led to specific and mechanistically appropriate drug predictions for each disease. See Figure S5 for enriched processes in SA and RAEB and Table S6 for all drug predictions. The RAEB histology photo is reprinted with permission from the American Society of Hematology. The SA photo is from the SA Wikipedia page where no attribution was indicated.

carry sufficient oxygen (Sankaran and Weiss, 2015), consequently, both anemias exhibit similar gene expression profiles (median sample correlation within biological replicates for each disease is 0.891 (SA) and 0.875 (RAEB), and correlation between the two diseases is 0.875). However, this similarity discounts the mechanistic differences between SA and RAEB that lead to the anemia and necessitate different treatment approaches: SA is an iron storage disease and RAEB manifests due to defects in cell differentiation.

URSA^{HD} models detected appropriate hallmark processes for each disease that mechanistically differentiate these two anemias leading to appropriate candidates for repurposing. Specifically, we found a strong statistical association between the URSA^{HD} RAEB model and Wilms tumors (MESHID:D009396) due in part to up-weighting genes related to aberrant hypermethylation and megakaryocyte differentiation (Figures 5B and S5, p value = 2.3×10^{-4}). Based on this association, URSA^{HD} predicted cancer chemotherapy drugs such as cisplatin, etoposide, melphalan, tretinoin, and vincristine to be effective for RAEB. Unlike for RAEB, heme-biosynthesis-related biological processes were enriched in the URSA^{HD} model for SA (Figure S5). Based on its association with β-Thalassemia (MESH:D017086) and Iron Overload (MESH:D019190), iron chelators, hematopoiesis stimulants, and antioxidants (such as deferiprone, resveratrol, and mangiferin) were among the drug treatment predictions for SA (Figure 5C, p value < 1.0×10^{-5}). Indeed, chemotherapy drugs were shown to bring remission for RAEB patients (Itoh et al., 1992; Kikuchi et al., 2012; Kuendgen et al., 2005; Whittle et al., 2013) and iron chelators were successfully used for SA patients (Bottomley and Fleming, 2014; Martin et al., 2006). Note that these treatments were not recorded in the CTD drug database. Therefore, URSA^{HD} models accurately distinguish the mechanistic differences between seemingly similar diseases, and this distinction can be leveraged to propose mechanistically appropriate drugs for targeted treatments. The targets for 97% of drug predictions (190 out of 196) were genes that are in fact expressed in the rare diseases. See Table S6 for the complete list of therapeutic drug predictions for rare diseases.

URSA^{HD} Detects Molecular Changes in Samples after Therapeutic Treatment

Accurately quantifying the efficacy of treatment is crucial for understanding drug-specific resistance and developing more effective therapies. However, the effects of drugs are often poorly understood at the molecular level, and clinical measures

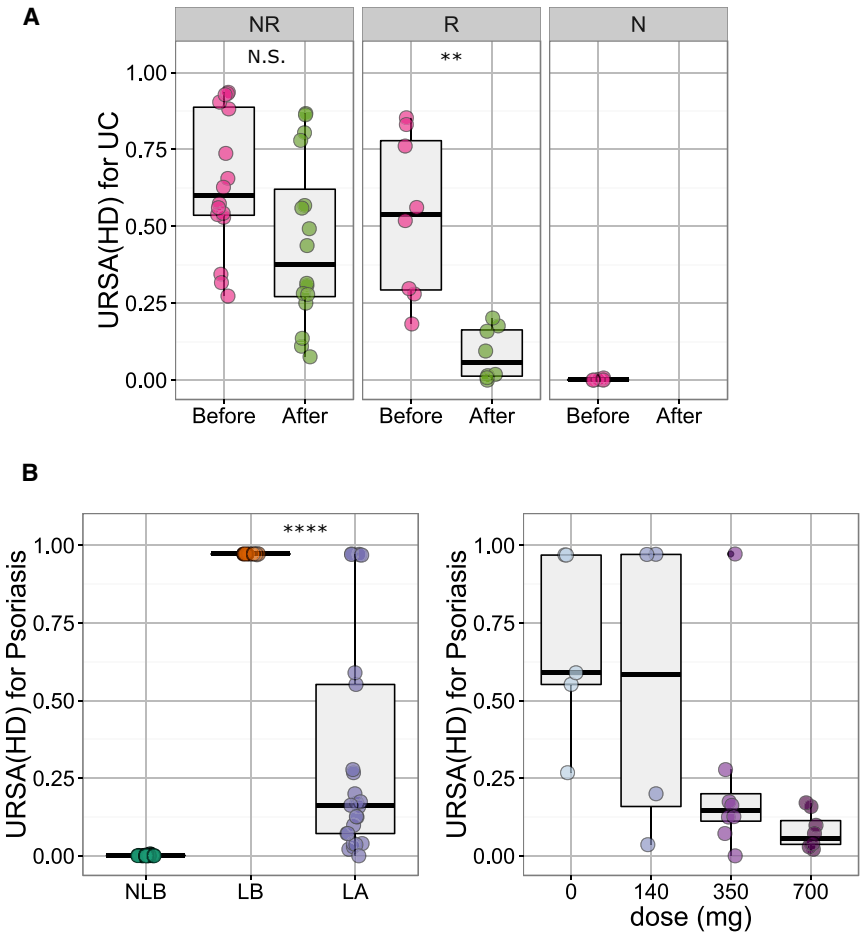


Figure 6. URSA^{HD} Quantification of Efficacy Following Therapeutic Drug Treatment

(A) The URSA^{HD} predictions for ulcerative colitis (UC) samples before and after treatment (GEO: GSE16879, dataset not included in URSA^{HD} training). Each panel plots non-responsive (NR), responsive (R), and control samples (N), respectively. Before treatment with infliximab, URSA^{HD} estimated ulcerative colitis with high probabilities for both responsive and non-responsive samples. However, probabilities for ulcerative colitis decreased after treatment only for the responsive samples (Wilcoxon rank-sum test, N.S. = 0.021, ** = 0.00031).

(B) (left panel) The URSA^{HD} predictions for psoriasis of skin biopsy samples before and after treatment with brodalumab (GEO: GSE53552, dataset not included in URSA^{HD} training). (left panel) URSA^{HD} estimated zero disease probability in control non-lesional samples (NLB), high predictive probabilities of psoriasis for lesional samples before treatment (LB), and low probabilities of disease for samples after treatment (LA), demonstrating a decrease in disease signal in response to treatment. Wilcoxon rank-sum test, **** = 2.4×10^{-9} (right panel). The URSA^{HD}-estimated psoriasis probabilities negatively correlated with dose of brodalumab treatment, indicating a dose-dependent decrease in disease signal.

of patient-specific outcomes are mostly qualitative, making it difficult to systematically capture meaningful drug actions. Here, we tested the ability of URSA^{HD} to recognize the genome-wide disruption caused by drugs in gene expression profiles of post-treatment patient samples. URSA^{HD} models do not use any post-treatment samples for training but instead rely on changes in disease-specific molecular signals to predict the effectiveness of drugs for each patient (STAR Methods). We first applied URSA^{HD} to pre- and post-treatment ulcerative colitis patient samples (GEO: GSE16879) (Arijs et al., 2009). Note that this entire dataset was excluded from training URSA^{HD}. For samples taken prior to treatment with infliximab, URSA^{HD} estimated high probabilities for ulcerative colitis for all patients—both responders and non-responders. For post-treatment samples, URSA^{HD} estimates were different for responders and non-responders and were consistent with the degree of mucosal healing, which was assessed independently in the original study (Figure 6A).

As a second test, we applied URSA^{HD} to a set of samples from 25 psoriasis patients treated with brodalumab (GEO: GSE53552) (Russell et al., 2014). The estimate of URSA^{HD} for psoriasis signal was low (essentially 0) for non-lesional samples and high for pre-treatment lesional samples (Figure 6B, left). URSA^{HD} was never trained to recognize the effects of brodalumab treatments or any other treatment effect (STAR Methods), but nonetheless, URSA^{HD} found a significant dose-dependent decrease of the URSA^{HD} psoriasis signal in post-treatment samples (Figure 6B). Taken together, we found that URSA^{HD} tracks molecular changes in response to drug treatment and even accurately predicts the decreases of disease signal in patient samples that are responsive to drug treatment. Thus, URSA^{HD} could provide a quantitative way for researchers to define responders from non-responders and assess the correspondence between molecular response and clinical endpoints.

DISCUSSION

Here, we describe URSA^{HD}, a computational framework that integrates the massive amount of gene expression data in the public compendium guided by the anatomical relationships inherent in the MeSH disease hierarchy and uses machine learning to generate individual disease models based on a particular disease's distinctive genome-wide traits. We show that URSA^{HD} can accurately detect the molecular signals unique to a disease, even distinguishing among closely related diseases, using ectodermal diseases as an example. We also demonstrate that URSA^{HD} outperforms traditional disease markers and differential gene expression approaches in predicting disease states and classifies disease samples accurately while identifying many understudied genes associated with human disease (Figure S6). Furthermore, we provide experimental corroboration of the predictions that result from the URSA^{HD} model for neuroblastoma. The URSA^{HD} approach relies on large dataset collections,

underscoring the need for publicly available disease-specific and normal tissue (i.e., GTEx) expression data.

Conceptually, URSA^{HD} brings two key advantages: a focus on differentiating between diseases (not just between normal and disease states) and a genome-wide data-driven approach that obviates biases toward well-studied genes and biomarkers. These aspects of URSA^{HD} are especially important for generating hypotheses for disease and treatment research, particularly for diseases that are rare or under-studied on the molecular level.

The URSA^{HD} framework is extendable, requiring only training samples and expert expansion of the disease hierarchy. To demonstrate this, our expert curator expanded the MeSH breast cancer node to include terms representing the PAM50 molecular subtypes (normal-like, luminalA, luminalB, Her2+, and basal), which are not part of MeSH. We then retrained URSA^{HD} with appropriate gene expression samples. For each subtype, URSA^{HD} provided high scores to the appropriate subtype; thus, demonstrating the extensibility of the URSA^{HD} framework (Figure S2B). Requests to include additional disease models are accepted on the URSA^{HD} website.

Understanding the molecular basis of the human disease landscape is paramount for therapeutic drug development and proper repurposing of existing drugs that have been clinically proven to be safe. URSA^{HD} is not fine-tuned for a specific drug treatment but is general in that it captures the distinct disease signatures in gene expression data. We provided proof of concept that disease signatures in URSA^{HD} models can be used to detect disease signal reduction in response to drug treatment and to suggest appropriate drugs for repurposing. Extending URSA^{HD} to other rare diseases is straightforward, as our approach sidesteps the need for prior knowledge of causal genes or tissue of origin. Altogether, this computational framework can be used by biomedical scientists to gain mechanistic insight into rare disease etiologies and repurpose drugs for the thousands of rare human diseases.

The URSA^{HD} interactive user interface provides both estimates of disease signal for researcher's datasets as well as links to the interpretable disease models, including biological processes, associated tissue and anatomical information, and weighted gene lists that are directly usable by the biomedical research community: ursahd.princeton.edu.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and six tables and can be found with this article online at https://doi.org/10.1016/j.cels.2018.12.010.

ACKNOWLEDGMENTS

We thank Raphael Moscher, MD, PhD; Lukas Tanner, PhD; and Joshua Rabinowitz, PhD for valuable technical assistance and discussions regarding the neuroblastoma cell line experiments. We greatly appreciate members of the Troyanskaya laboratory for discussions and suggestions on the manuscript. O.G.T. is a senior fellow of the Genetic Networks program of the Canadian Institute for Advanced Research. This work was supported by the NIH R01 GM071966 to O.G.T. and R24OD011194 awarded to K.D. and O.G.T.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.L., A.K., C.L.T., and O.G.T.; Methodology, Y.L. and A.K.; Software, Y.L.; Validation, Y.L.; Data Curation, Y.L., C.L.T., R.O., J.R., and C.C.; Investigation, Y.L.J.R., and C.LT.; Visualization, Y.L., J.R., C.T., and K.D.; A.K., and V.K. provided expert feedback; Y.L. developed the web server and interface; Writing – Original Draft, Y.L. and C.L.T.; Writing – Review and Editing, Y.L., C.L.T., R.O., J.R., C.C., K.D., V.K., and O.G.T.; Funding Acquisition, K.D. and O.G.T.; Supervision, C.L.T., K.D., V.K., and O.G.T.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 8, 2018
Revised: October 16, 2018
Accepted: December 20, 2018
Published: January 23, 2019

WEB RESOURCES

URSA^{HD}, ursahd.princeton.edu

REFERENCES

Ade, A., Wright, Z., and States, D. (2007). Gene2MeSH. http://www.ncibi.org/gene2mesh.html.

Amar, D., Hait, T., Izraeli, S., and Shamir, R. (2015). Integrated analysis of numerous heterogeneous gene expression profiles for detecting robust disease-specific biomarkers and proposing drug targets. Nucleic Acids Res. *43*, 7779–7789.

Arijs, I., De Hertogh, G., Lemaire, K., Quintens, R., Van Lommel, L., Van Steen, K., Leemans, P., Cleynen, I., Van Assche, G., Vermeire, S., et al. (2009). Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. PLoS One *4*, e7984.

Aymé, S., Bellet, B., and Rath, A. (2015). Rare diseases in ICD11: making rare diseases visible in health information systems through appropriate coding. Orphanet J. Rare Dis. *10*, 35.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995.

Barutcuoglu, Z., Schapire, R.E., and Troyanskaya, O.G. (2006). Hierarchical multi-label prediction of gene function. Bioinformatics *22*, 830–836.

Bonifati, C., and Ameglio, F. (1999). Cytokines in psoriasis. Int. J. Dermatol. *38*, 241–251.

Bottomley, S.S., and Fleming, M.D. (2014). Sideroblastic anemia: diagnosis and management. Hematol. Oncol. Clin. North Am. *28*, 653–670.

Burges, C.J.C. (1998). A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. *2*, 121–167.

Clark, A.F., Steely, H.T., Dickerson, J.E., Jr., English-Wright, S., Stropki, K., McCartney, M.D., Jacobson, N., Shepard, A.R., Clark, J.I., Matsushima, H., et al. (2001). Glucocorticoid induction of the glaucoma gene MYOC in human and monkey trabecular meshwork cells and tissues. Invest. Ophthalmol. Vis. Sci. *42*, 1769–1780.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H., et al. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. *33*, e175.

Das, U.N., and Rao, A.A. (2007). Gene expression profile in obesity and type 2 diabetes mellitus. Lipids Health Dis. *6*, 35.

Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., King, B.L., McMorran, R., Wiegers, J., Wiegers, T.C., and Mattingly, C.J. (2017). The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Res. *45*, D972–D978.

Druzdzel, M.J. (1999). SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: a development environment for graphical decision-theoretic models. In Proceedings of the sixteenth national conference on artificial intelligence and the eleventh innovative applications of artificial intelligence conference innovative applications of artificial intelligence, pp. 902–903.

Dunckley, T., Beach, T.G., Ramsey, K.E., Grover, A., Mastroeni, D., Walker, D.G., LaFleur, B.J., Coon, K.D., Brown, K.M., Caselli, R., et al. (2006). Gene expression correlates of neurofibrillary tangles in Alzheimer's disease. Neurobiol. Aging *27*, 1359–1371.

Godard, S., Getz, G., Delorenzi, M., Farmer, P., Kobayashi, H., Desbaillets, I., Nozaki, M., Diserens, A.C., Hamou, M.F., Dietrich, P.Y., et al. (2003). 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 Res. *63*, 6613–6625.

Gómez Ravetti, M., Rosso, O.A., Berretta, R., and Moscato, P. (2010). Uncovering molecular biomarkers that correlate cognitive decline with the changes of hippocampus' gene expression profiles in Alzheimer's disease. PLoS One *5*, e10153.

Hodges, A., Strand, A.D., Aragaki, A.K., Kuhn, A., Sengstag, T., Hughes, G., Elliston, L.A., Hartog, C., Goldstein, D.R., Thu, D., et al. (2006). Regional and cellular gene expression changes in human Huntington's disease brain. Hum. Mol. Genet. *15*, 965–977.

Huang, H., Liu, C.C., and Zhou, X.J. (2010). Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. Proc. Natl. Acad. Sci. U S A *107*, 6823–6828.

Hubbell, E., Liu, W.M., and Mei, R. (2002). Robust estimators for expression analysis. Bioinformatics *18*, 1585–1592.

International Cancer Genome Consortium, Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. Nature *464*, 993–998.

Itoh, K., Igarashi, T., and Wakita, H. (1992). Successful treatment with vincristine by slow infusion in a patient with refractory anemia and excess of blasts. Am. J. Hematol. *39*, 73–74.

Jaffe, A.B., and Hall, A. (2005). Rho GTPases: biochemistry and biology. Annu. Rev. Cell Dev. Biol. *21*, 247–269.

Joachims, T. (2006). Training linear SVMs in linear time. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Philadelphia, PA: ACM), pp. 217–226.

Kaufman, L.J., Brangwynne, C.P., Kasza, K.E., Filippidi, E., Gordon, V.D., Deisboeck, T.S., and Weitz, D.A. (2005). Glioma expansion in collagen I matrices: analyzing collagen concentration-dependent growth and motility patterns. Biophys. J. *89*, 635–650.

Kawakami, K., Enokida, H., Tachiwada, T., Gotanda, T., Tsuneyoshi, K., Kubo, H., Nishiyama, K., Takiguchi, M., Nakagawa, M., and Seki, N. (2006). Identification of differentially expressed genes in human bladder cancer through genome-wide gene expression profiling. Oncol. Rep. *16*, 521–531.

Kikuchi, A., Hasegawa, D., Ohtsuka, Y., Hamamoto, K., Kojima, S., Okamura, J., Nakahata, T., and Manabe, A.; Japanese Paediatric Myelodysplastic Syndrome Study (2012). Outcome of children with refractory anaemia with excess of blast (RAEB) and RAEB in transformation (RAEB-T) in the Japanese MDS99 study. Br. J. Haematol. *158*, 657–661.

Kim, S.Y., and Volsky, D.J. (2005). PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics *6*, 144.

Kuendgen, A., Knipp, S., Fox, F., Strupp, C., Hildebrandt, B., Steidl, C., Germing, U., Haas, R., and Gattermann, N. (2005). 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. Ann. Hematol. *84 Suppl. 1*, 61–66.

Kwon, H.S., Nakaya, N., Abu-Asab, M., Kim, H.S., and Tomarev, S.I. (2014). Myocilin is involved in NgR1/Lingo-1-mediated oligodendrocyte differentiation and myelination of the optic nerve. J. Neurosci. *34*, 5539–5551.

Lauder, I., and Aherne, W. (1972). The significance of lymphocytic infiltration in neuroblastoma. Br. J. Cancer *26*, 321–330.

Lee, Y.S., Krishnan, A., Zhu, Q., and Troyanskaya, O.G. (2013). Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. Bioinformatics *29*, 3036–3044.

Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. *11*, 733–739.

Leonardi, C.L., Powers, J.L., Matheson, R.T., Goffe, B.S., Zitnik, R., Wang, A., and Gottlieb, A.B.; Etanercept Psoriasis Study (2003). Etanercept as monotherapy in patients with psoriasis. N. Engl. J. Med. *349*, 2014–2022.

Martin, F.M., Bydlon, G., and Friedman, J.S. (2006). SOD2-deficiency sideroblastic anemia and red blood cell oxidative stress. Antioxid. Redox Signal. *8*, 1217–1225.

Moran, L.B., Duke, D.C., Deprez, M., Dexter, D.T., Pearce, R.K., and Graeber, M.B. (2006). Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics *7*, 1–11.

Namløs, H.M., Kresse, S.H., Müller, C.R., Henriksen, J., Holdhus, R., Sæter, G., Bruland, O.S., Bjerkehagen, B., Steen, V.M., and Myklebost, O. (2012). Global gene expression profiling of human osteosarcomas reveals metastasis-associated chemokine pattern. Sarcoma *2012*, 639038.

Pascual, V., Chaussabel, D., and Banchereau, J. (2010). A genomic approach to human autoimmune diseases. Annu. Rev. Immunol. *28*, 535–571.

Payne, L.S., and Huang, P.H. (2013). The pathobiology of collagens in glioma. Mol. Cancer Res. *11*, 1129–1140.

Rack, P.G., Ni, J., Payumo, A.Y., Nguyen, V., Crapster, J.A., Hovestadt, V., Kool, M., Jones, D.T., Mich, J.K., Firestone, A.J., et al. (2014). Arhgap36-dependent activation of Gli transcription factors. Proc. Natl. Acad. Sci. U S A *111*, 11061–11066.

Rosmond, R., Chagnon, M., Bouchard, C., and Björntorp, P. (2001). G-308A polymorphism of the tumor necrosis factor alpha gene promoter and salivary cortisol secretion. J. Clin. Endocrinol. Metab. *86*, 2178–2180.

Rothman, N., Skibola, C.F., Wang, S.S., Morgan, G., Lan, Q., Smith, M.T., Spinelli, J.J., Willett, E., De Sanjose, S., Cocco, P., et al. (2006). Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. Lancet Oncol. *7*, 27–38.

Rung, J., and Brazma, A. (2013). Reuse of public genome-wide gene expression data. Nat. Rev. Genet. *14*, 89–99.

Russell, C.B., Rand, H., Bigler, J., Kerkof, K., Timour, M., Bautista, E., Krueger, J.G., Salinger, D.H., Welcher, A.A., and Martin, D.A. (2014). Gene expression profiles normalized in psoriatic skin by treatment with Brodalumab, a human anti-IL-17 receptor monoclonal antibody. J. Immunol. *192*, 3828–3836.

Sankaran, V.G., and Weiss, M.J. (2015). Anemia: progress in molecular mechanisms and therapies. Nat. Med. *21*, 221–230.

Schmid, P.R., Palmer, N.P., Kohane, I.S., and Berger, B. (2012). Making sense out of massive data by going beyond differential expression. Proc. Natl. Acad. Sci. U S A *109*, 5594–5599.

Somasundaram, R., and Herlyn, D. (2009). Chemokines and the microenvironment in neuroectodermal tumor-host interaction. Semin. Cancer Biol. *19*, 92–96.

Suthram, S., Dudley, J.T., Chiang, A.P., Chen, R., Hastie, T.J., and Butte, A.J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput. Biol. *6*, e1000662.

Whittle, A.M., Feyler, S., and Bowen, D.T. (2013). 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. Leuk. Res. Rep. *2*, 9–11.

Yang, K.D., Shaio, M.F., Wang, C.L., Wu, N.C., and Stone, R.M. (1994). Neuroblastoma cell-mediated leukocyte chemotaxis: lineage-specific differentiation of interleukin-8 expression. Exp. Cell Res. *211*, 1–5.

Zhou, M., Wiemels, J.L., Bracci, P.M., Wrensch, M.R., McCoy, L.S., Rice, T., Sison, J.D., Patoka, J.S., and Wiencke, J.K. (2010). Circulating levels of the innate and humoral immune regulators CD14 and CD23 are associated with adult glioma. Cancer Res. *70*, 7534–7542.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-MAB21L1, 1:1000	ThermoFisher	Cat#PA5-31335; RRID:AB_2548809
Critical Commercial Assays		
Lipofectamine RNAiMax	ThermoFisher	Cat#13778-150
SuperScript III	ThermoFisher	Cat#11752-050
RNeasy PLUS Mini Kit	Qiagen	Cat#74134
Experimental Models: Cell Lines		
BE(2)-C	ATCC	Cat# CRL-2268, RRID:CVCL_0529
SH-SY5Y	ATCC	Cat# CRL-2266, RRID:CVCL_0019
Oligonucleotides		
Silencer Select siRNA MAB21L1 (#1)	ThermoFisher	Cat# 8384
Silencer Select siRNA MAB21L1 (#2)	ThermoFisher	Cat# s8383
Silencer Select siRNA ARHGAP36 (#1)	ThermoFisher	Cat# s46108
Silencer Select siRNA ARHGAP36 (#2)	ThermoFisher	Cat# s46110
Silencer Select siRNA LOC100507194	ThermoFisher	Cat# n506417
Silencer Select siRNA Negative Control #1	ThermoFisher	Cat# 4390846
MAB21L1	ThermoFisher	Cat# Hs00366575
MAB21L2	ThermoFisher	Cat# Hs00740710_s1
ARHGAP36	ThermoFisher	Cat# Hs01557499_m1
LOC100507194	ThermoFisher	Cat# Hs04274314_m1
GAPDH	ThermoFisher	Cat# 4333764f
Software and Algorithms		
URSA^{HD}	This paper	http://ursahd.princeton.edu/
T-Scratch	Tobias Gebäck and Martin Schulz, ETH Zürich	http://www.cse-lab.ethz.ch/software/
R2: Genomics Analysis and Visualization Platform	Dr. Jan Koster	http://r2.amc.nl
Other		
GEO training datasets – see Table S1	N/A	N/A
GEO ARHGAP36 Survival Analysis (Figure 2)	N/A	GEO: GSE49710

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Olga G. Troyanskaya (ogt@cs.princeton.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Culture

Cell lines BE(2)-C (CRL-2268, RRID:CVCL_0529, male) and SH-SY5Y (CRL-2266, RRID:CVCL_0019, female) were acquired from the ATCC immediately before experiments and considered authenticated. Cells were cultured according to ATCC guidelines in EMEM:Ham's F12 (1:1) with 10% FBS at 37°C and 5% CO2. Experiments were performed with cells passaged fewer than 25 times.

METHOD DETAILS

URSA^{HD}: Hierarchy-Aware Characterization and Classification of Human Diseases

The URSA^{HD} framework for disease prediction is set up as a hierarchical multi-label classification problem (Barutcuoglu et al., 2006; Lee et al., 2013). Each individual disease classifier characterizes the distinctive features of expression profiles of the disease compared to that of all "other" (i.e., control) normal tissue samples and unrelated disease samples (Table S1). The Bayesian network then models the probabilistic relationship between these classifiers based on the MeSH disease hierarchy to correct the individual predictions and thus provides an interpretable list of disease predictions for a given clinical gene expression profile.

Hierarchy-Aware Individual Disease Characterization

The set of positive and negative samples defines the learning criteria for the classifier, and so systematically setting the context of a particular disease is important for a unified framework. We use the MeSH disease hierarchy to set this context, enabling the final classifiers to 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. For example, both neuroblastoma and oligodendroglioma samples are considered positive examples for the neuroepithelial neoplasms (MESH:D018302) model because both neuroblastoma (MESH:D009447) and oligodendroglioma (MESH:D009837) are descendant terms. Samples annotated to only the ancestor terms of the disease term in question are excluded from training; and those samples annotated to its sibling terms are considered negative. That is, oligodendroglioma samples are considered negative examples for the neuroblastoma model because neuroblastoma (MESH:D009447) and oligodendroglioma (MESH:D009837) are sibling terms according to the MeSH hierarchy. To account for the tissue-specific signals present in the samples (beyond the disease signals the models are built to capture), every negative training set also includes all control (i.e., normal/non-diseased) samples.

Specifically, given l pairs (sample data and their labels) of gene expression profile (vector) x_i and its disease or control label y_i for $i \in l$, we build each individual disease model using the L2 linear SVM (with cost parameter $c = 20$) as follows (Joachims, 2006):

$$min_w \frac{1}{2} w^T w + c \sum_{i \in l} max\left(1 - y_i w^T x_i, 0\right)^2$$

Hierarchy-Aware Probabilistic Aggregation of Distinctive Classification Models

To hierarchically integrate and calibrate individual models, URSA^{HD} used the structure of the MeSH disease hierarchy and defined these relationships in a Bayesian network (Barutcuoglu et al., 2006; Lee et al., 2013). We modeled the unthresholded SVM output of each term 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 $\hat{y}$ established conditional independence of an SVM prediction $\hat{y}_i$ to all other SVM predictions $\hat{y}_j$ ($i \neq j$)given its true label y_i. As such, the likelihood was computed as:

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

Through 2-fold cross-validation, these conditional probability tables, $P(\hat{y}_i|y_i)$, were estimated by counting the number of negative samples with SVM outputs smaller than that of a positive SVM output. Laplace smoothing was applied for robustness.

The parent-child conditional probability tables were defined such that a label was true when any one of its children was true (Barutcuoglu et al., 2006). The prior was computed as:

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

where $ch(y_i)$ is the children terms of y_i. When none of its children were true (including when it has no children), a constant prior of 0.1 was assigned. We used the loopy belief propagation algorithm implemented in the SMILE library to infer the posterior probabilities $P(y_i|\hat{y}_1, ..., \hat{y}_N)$ for each disease term (Druzdzel, 1999). These posterior probabilities were used to annotate gene expression samples to a disease.

Comparison to Individual Disease Prediction Methods

Literature-Documented Gene Prediction Method (Figure 2A)

A common method to predict disease signals is from 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 the Area-Under-the-Precision-Recall-Curve (AUPRC) ranking accuracy. Positive samples are 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' samples were excluded. This method represents a typical single gene-based approach that relies only on documented disease genes.

Normal/Disease Differential Model (Figure 2B)

Genome-wide differential analysis between normal and clinical samples was performed by Support Vector Machine (SVM) trained on disease samples from direct sample annotations (as positives) and normal samples from the same dataset (as negatives), as per traditional differential expression setup with cost parameter c = 20 (Burges, 1998).

Training and Testing Setup

Method evaluations are often done with a random sample holdout. However, genome-wide experiments are prone to laboratory and dataset biases, and so a simple random sample holdout can overestimate the performance of these methods (Leek et al., 2010; Rung and Brazma, 2013). To control for this bias, the series/datasets of the manually annotated samples were randomly partitioned into training and testing datasets for each term as done previously (Lee et al., 2013). Only disease terms with at least 2 positive and negative samples in both the training dataset and the testing dataset were evaluated.

Determining the Relevant Biological Associations of URSA^{HD} Models

To extract biological information from URSA^{HD} models representing biological processes, anatomical context, and disease terms, we calculated the statistical association of defined gene sets using the PAGE enrichment algorithm (Kim and Volsky, 2005). Specifically, to detect biological process and pathway signals in URSA^{HD} models, we used the GO term gene sets; to detect anatomical signals, we used MeSH anatomy terms; to detect associations between diseases for drug repurposing, we used gene2mesh disease gene sets (Tables S5 and S6).

Specifically, each individual URSA^{HD} disease model is a vector of length m, containing weights for each gene: $w=\{w_1, w_2, ..., w_m\}$ where m is the number of genes covered by the gene expression profile assay. Given a disease model w_d (the gene weight vector) and gene set S_t representing term t, the enrichment score z_{td} representing the association between term t and disease d is:

$$z_{td} = \frac{\underline{x}_{td} - \mu_d}{\sigma_d}$$

$$\underline{x}_{td} = \frac{\sum_{g \in S_t} w_g}{n}$$

where n is the size of S_t and μ_d and σ_d is the population mean and standard deviation of w_d, respectively. The statistical significance (p values) of the association between term t and disease d were obtained by computing the probability of obtaining a z-score greater than z_{td} from the standard normal distribution.

Documented Disease and Anatomy Genes

Gene2mesh uses curated MeSH annotations of PubMed articles to find genes that are statistically significantly studied with a particular MeSH term (Ade et al., 2007). 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.

Drug Repurposing Evaluation Based on Disease Associations

URSA^{HD} disease models for rare diseases were associated with well-studied diseases based on the statistical association score of each model with disease genesets (as described in section above).

To first systematically evaluate the approach of using URSA^{HD}-based disease associations to generate drug repurposing hypotheses, we tested our approach on known therapeutic targets (curated in the Comparative Toxicogenomics Database (CTD) (Davis et al., 2017)) using all URSA^{HD} models. Intuitively, if URSA^{HD}-based drug repurposing hypotheses are informative, then the significance score (as defined above) for URSA^{HD}-associated disease pairs will be higher for those pairs that share a drug vs. all other pairs. This evaluation is robust to the lack of true "negative" chemical disease associations (where the chemical is known to have no therapeutic effects on the disease).

Specifically, let C be the set of all disease pairs that share at least one therapeutic chemical according to CTD. A paired Wilcoxon signed rank significance test was used to identify significant associations between disease pair (t, d), testing the following condition:

$$E_d[z_{td} | (d,t) \in C] > E_d[z_{td}]$$

where

$$E_d[z_{td} | (d,t) \in C] = \frac{1}{|C|} \sum_{(d,t) \in C} z_{td} \quad and$$

$$E_d[z_{td}] = \frac{1}{|T|}\sum_{t \in T} z_{td}$$

where E_d is the probabilistic expectation for disease d, and T is the set of all disease terms in gene2mesh with at least 10 documented genes. The association score between diseases that share a therapeutic drug was significantly greater than random pairs of diseases (paired ranked-sum test, p value = 10^{-23}).

To predict drugs that could be used to treat rare diseases, we used significant associations between each rare disease and well-studied disease. Specifically, for each rare disease $t \in T$, the disease enrichment scores z_{td} for all gene2mesh disease terms were computed for URSA^{HD} model of disease t and then, the annotated therapeutic drug for disease d with enrichment score greater than 5 was considered a therapeutic drug candidate for disease t.

Therapeutic Chemical:Disease Associations

Chemical disease associations were downloaded from the Comparative Toxicogenomics Database (CTD) on Mar 2, 2015 (Davis et al., 2017). CTD contains both curated and inferred chemical-disease interactions. Only curated associations with direct therapeutic evidence were used, a total of 27,571 associations with 5,852 unique chemicals to 2,290 diseases. Hypertension (MESH: D006973) had the most associated therapeutic chemicals (n = 343).

Rare Diseases

List of rare diseases were downloaded from OrphaData V 0.9 on Nov 17, 2014 at http://www.orphadata.org (Aymé et al., 2015). 20 of our models were for rare diseases (Table S6).

Expression Data Processing

The Human Genome U133 Plus 2.0 Array (hgu133plus2) raw CEL files were downloaded from Gene Expression Omnibus (GEO) (Barrett et al., 2013). Their 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 (Dai et al., 2005; Hubbell et al., 2002). Therapeutic treatment datasets (GEO: GSE16879 and GEO: GSE53552) used only for analysis were also pre-processed and normalized using the same pipeline (Arijs et al., 2009; Russell et al., 2014). Clinical information (patient id, diagnosis, treatment type, and response type) were from the author-provided sample description in GEO.

Gold Standard Construction by Manual 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'' (GSM404013) and ''surrounding noncancerous cells'' (GSM490997) were annotated as ''other''. A total of 1 996 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.

TCGA's RNA-Seq Sample Predictions

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 (Lee et al., 2013). 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 RNA-Seq Version 2 IlluminaHiSeq normalized gene expression data (Data level 3) was downloaded on July 18 2014 (International Cancer Genome Consortium et al., 2010). Fifteen different cancer-types were covered by both TCGA's RNA-Seq Version 2 and the current disease models at the time (available on the URSA^{HD} website). Predictions were made for a total of 6,172 RNA-Seq samples.

PubMed Article Gene Annotations

Human gene annotations to PubMed articles were downloaded from the National Center for Biotechnology Information (NCBI) on Oct 31, 2014. 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.

Cell Culture and Transfection

BE(2)-C (CRL-2268) and SH-SY5Y (CRL-2266) were acquired from the ATCC and cultured in EMEM:Ham's F12 (1:1) with 10% FBS at 37°C and 5% CO2, according to ATCC guidelines. Silencer Select siRNAs were used for gene product knockdowns and were from ThermoFisher ((MAB21L1 (s8383 (#2), s8384 (#1)), ARHGAP36 (s46108(#1), s46110(#2)), LOC100507194 (n506417), and Negative Control #1 (4390846)).

Cell Viability Assay

Cells were transfected with 20–25 pmol of siRNA/well in triplicate in 6-well plates using RNAiMAX transfection reagent (ThermoFisher, 13778150) and accompanying protocol. RNA, protein, and viability phenotypes were assayed 72 h post-transfection. Viability was assessed by cell counting, and/or resazurin fluorescence assay. Both gave highly similar results. The assay was done in triplicate for each siRNA; three biological replicates were analyzed.

Quantitative RT-PCR

Expression was assessed by RT-qPCR. RNA was isolated from cells with Qiagen RNeasy Plus kit (74134); first strand cDNA was synthesized from 400ng of RNA using with Invitrogen's Superscript III polymerase kit (11752-050) according to the manufacturer's protocol. Quantitative RT-PCR was performed using 1ul of cDNA with ThermoFisher TaqMan assays (MAB21L1 (Hs00366575), MAB21L2 (Hs00740710_s1), ARHGAP36 (Hs01557499_m1), LOC100507194 (Hs04274314_m1) at 1X, and TaqMan Master Mix (4369016) in triplicate or 5 replicates each. GAPDH was the endogenous control (4333764f). Assays were run on an ABI ViiA7 Real Time PCR system. Efficiency curves were assessed and found to be nearly equal across assays and end products sizes were verified on 2% agarose gels. Quantification of relative gene expression fold changes were calculated by the delta delta Ct method.

Western Blotting

To assess MAB21L1 protein levels, protein lysates were prepared from cells using RIPA lysis buffer and 1X Halt Protease cocktail mix. 10 ug of protein lysate was loaded on to 4–20% BioRAD TGX-Stain Free gels. After resolution of samples on the gel, it was activated by UV light and protein loading assessed. Following semi-dry transfer to PVDF, blots were probed with antibodies to MAB21L1 (rabbit polyclonal anti-MAB21L1, ThermoFisher, PA5-31335, 1:1000) overnight at 4°C.

Cell Migration Assay

Cell migration assays were performed with BE(2)-C cells using Ibidi 2-well silicon culture inserts (catalog #81176) to create uniform 500 m gaps with minimal cell disruption. A minimum of 3 gaps were created for each experiment, imaged, and quantified for each condition, and the assay was performed 3 separate times. Briefly, a 2 mL aliquot of cells (8.0 x 10^5 cells/mL) in culture medium were reverse transfected with RNAiMAx transfection reagent (Thermofisher) and a total of 10 pmol of siRNAs. Immediately, 70l of cells were plated into each well of the inserts and 24–48 h later when confluency could be verified, inserts were pulled up to create gaps. Images were taken every 2–4 h for 48 h. The remaining 2 mL transfection was plated and used for viability assays at the immediate conclusion of the migration assay. Using T-scratch software (Tobias Gebäck and Martin Schulz, ETH Zürich; http://cse-lab.ethz.ch/software/), percent of area closed was calculated at 42 h, a timepoint before closure of control gaps. Cell doubling time is not less than 42 h for BE(2)-C cells. Five isoforms of ARHGAP36 are documented; we used low amounts of two different siRNAs to assess migration phenotype. Alone, neither siRNA1 nor siRNA2 gave a phenotype. The migration defect using both siRNAs was strong, reproducible, and possibly linked to sensing and/or cell shape changes, as cells failed to extend toward the opposing margins, even at close range (CLT unpublished observations). Viability was >90% for the single and double knockdown cultures.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests used for computational tests are explained in each subsection of the STAR Methods. All experiments were performed on at least 2 biological replicates (cells taken from separate cell culture passages for replicate experiments, e.g., for migration assays and Western blots) but usually 3–5 times (qRT-PCR experiments, and viability assays). Samples were not excluded from any experimental analyses. Sample size estimation power analyses were not performed before experiments. Researchers were not blinded to samples for experiments. Unpaired t-test was used to test for significance of different gap sizes in the technical replicates of the migration assay.

Kaplan-Meyer Survival Analysis

Survival data in a neuroblastoma RNA-Seq dataset (GEO: GSE49710, n = 498 RNA-seq profiles from 498 primary neuroblastomas) were analyzed using the R2: Genomics Analysis and Visualization Platform (http://r2.amc.nl). High and low expression cutoff were scanned with a minimal group size of 8. Samples were not filtered by sample variables. Event free survival was used to assess prognostic value.

DATA AND SOFTWARE AVAILABILITY

All data used in the paper are published previously and publicly available at GEO. Datasets used are listed in Table S1 and the Key Resources Table.

ADDITIONAL RESOURCES

Description: ursahd.princeton.edu

The URSA webserver is a public, user-friendly, dynamic web server for researchers to identify disease signals in any gene expression experiment as well as examine and use the underlying URSA^{HD} disease models to further guide research efforts. Researchers may suggest new models from the interface.

Cell Systems, Volume *8*

Supplemental Information

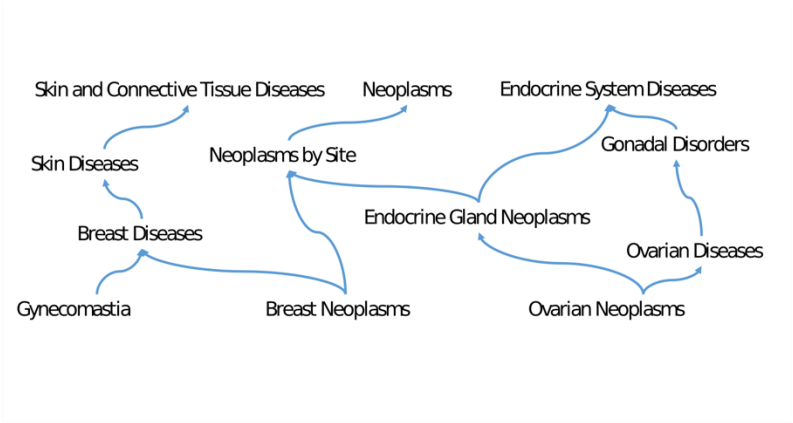
A Computational Framework for Genome-wide Characterization of the Human Disease Landscape

Young-suk Lee, Arjun Krishnan, Rose Oughtred, Jennifer Rust, Christie S. Chang, Joseph Ryu, Vessela N. Kristensen, Kara Dolinski, Chandra L. Theesfeld, and Olga G. Troyanskaya

Supplemental Information titles and legends

Table S1. Related to Figure 1: Manual annotation of gene expression profiles to human diseases
Table S2. Related to Figure 2 and S2a. Popular disease genes
Table S3. Related to Figure 2 and S2c. Mapping of TCGA tumor types to MeSH terms
Table S4. Related to Figure 2. Literature Curation for select URSA^{HD} top model genes
Table S5. Related to Figure 3 and Figure 4. Functional and Anatomical Enrichments of URSA^{HD} models
Table S6. Related to Figure 5. Rare Disease UIDs and therapeutic drug predictions
Supplementary Note. Related to Figure 2.

a



b

Figure S1. Related to Figure 1. Relationships between diseases represented in URSA^{HD} (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 for visualization.

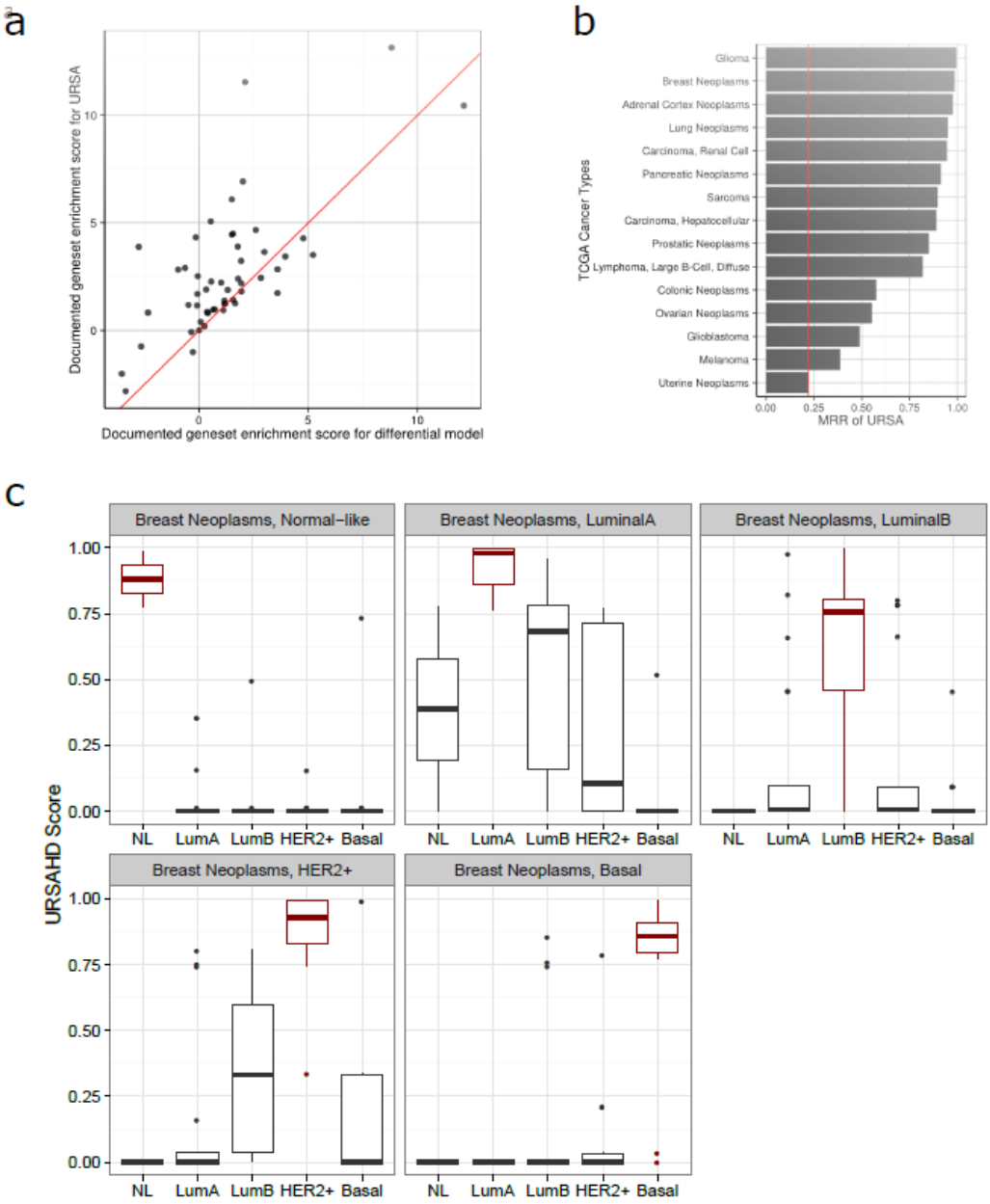


Figure S2. Related to Figure 2. (a) Documented disease genes (i.e. gene2mesh genes) are more enriched in URSA^{HD} models over traditional normal/disease models based on differential gene expression. Scatterplot comparison of documented disease gene set enrichment in URSA^{HD} models (y-axis) and normal/disease differential models (x-axis). Red line is the identity line and so dots above the red line indicate diseases with greater enrichment score. (b) URSA^{HD} accurate prediction for breast subtypes. In each panel, we query URSA^{HD} with samples not used in training belonging to each subtype, and plot the URSA^{HD} probability scores for each of the possible subtypes (Normal-like, LumA, LumB, Her2+, Basal). (c) Without retraining, URSA^{HD-} (trained only on microarray data) accurately predicts cancer samples from TCGA's RNA-Seq 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.

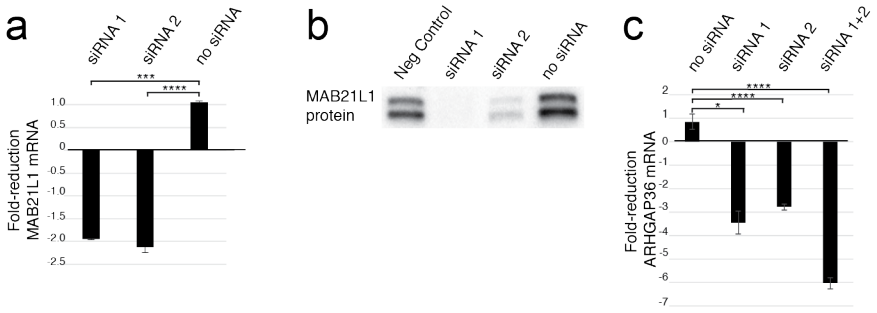


Figure S3. Related to Figure 2. Expression of top neuroblastoma URSA^{HD} model genes: MAB21L1, MAB21L2, and ARHGAP36 in BE(2)-C human neuroblastoma cells (a) Expression and efficient knockdown of MAB21L1 expression. qPCR of MAB21L1 expression levels in BE(2)-C cells transfected with negative control siRNA or MAB21L1 siRNA1 or siRNA2 (20 pmol/well; unpaired t-test: ***p=0.001, ****p <0.0001). Error bars represent standard error in three replicates. (b) corresponding protein levels following MAB21L1 knockdown. (c) Expression and efficient knockdown of ARHGAP36. BE(2)-C cells were transfected in triplicate with negative control siRNA (10pmol/well) or with ARHGAP36 siRNA1 (10 pmol/well), siRNA2 (10 pmol/well) or siRNA1+2 (5 pmol/well, each)(*p<.05, ****p<0.0001). Error bars are SE.

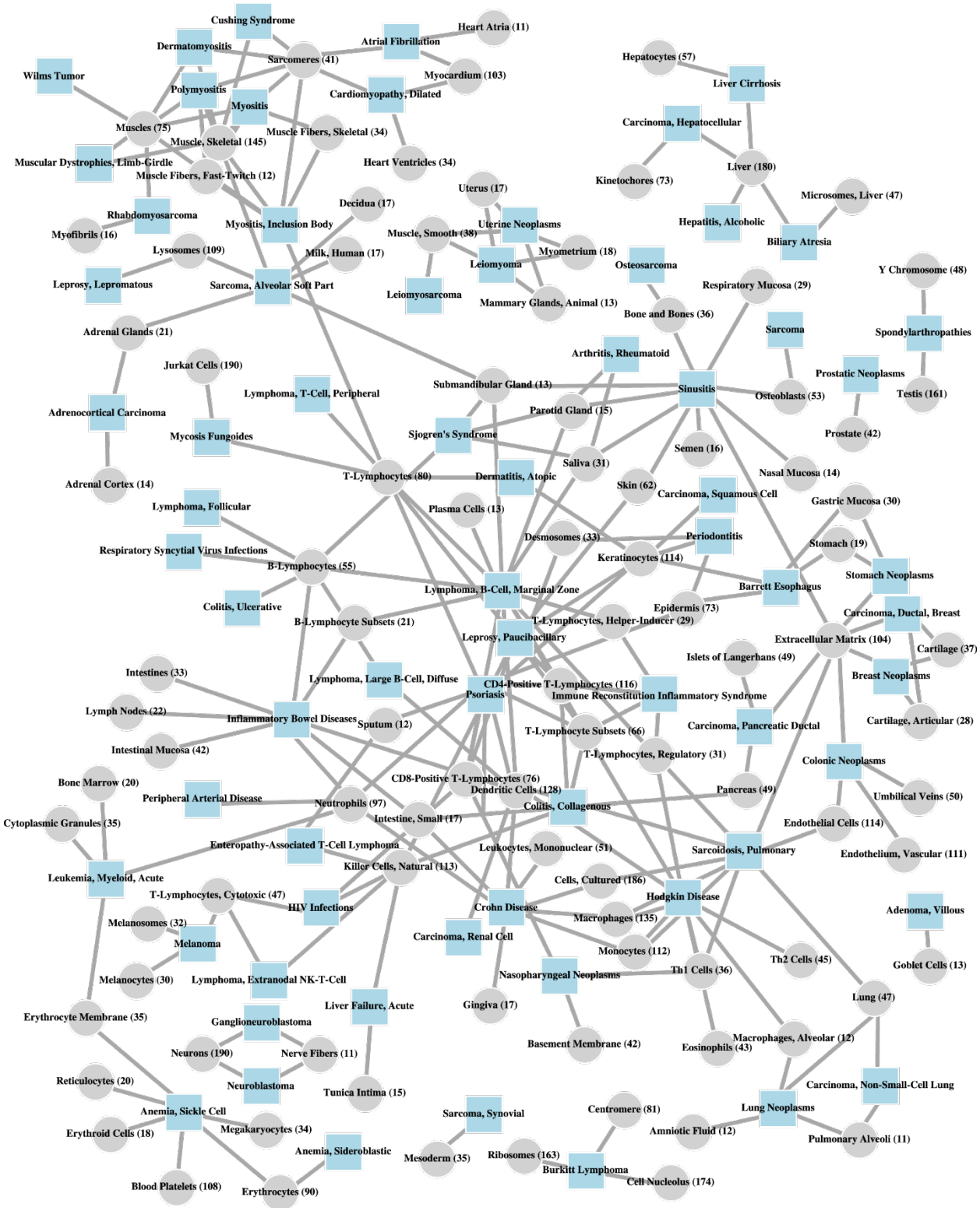


Figure S4. Related to Figure 4. Anatomical context defined by the URSA^{HD-}'s disease models. Bipartite graph of disease terms (blue squares) and anatomical MeSH terms (grey circles). Associations based on enrichment score over 5 were shown. Examples include heart diseases connected to heart-related tissues/cell-types and tissue-specific cancers connected to their appropriate tissue of origin.

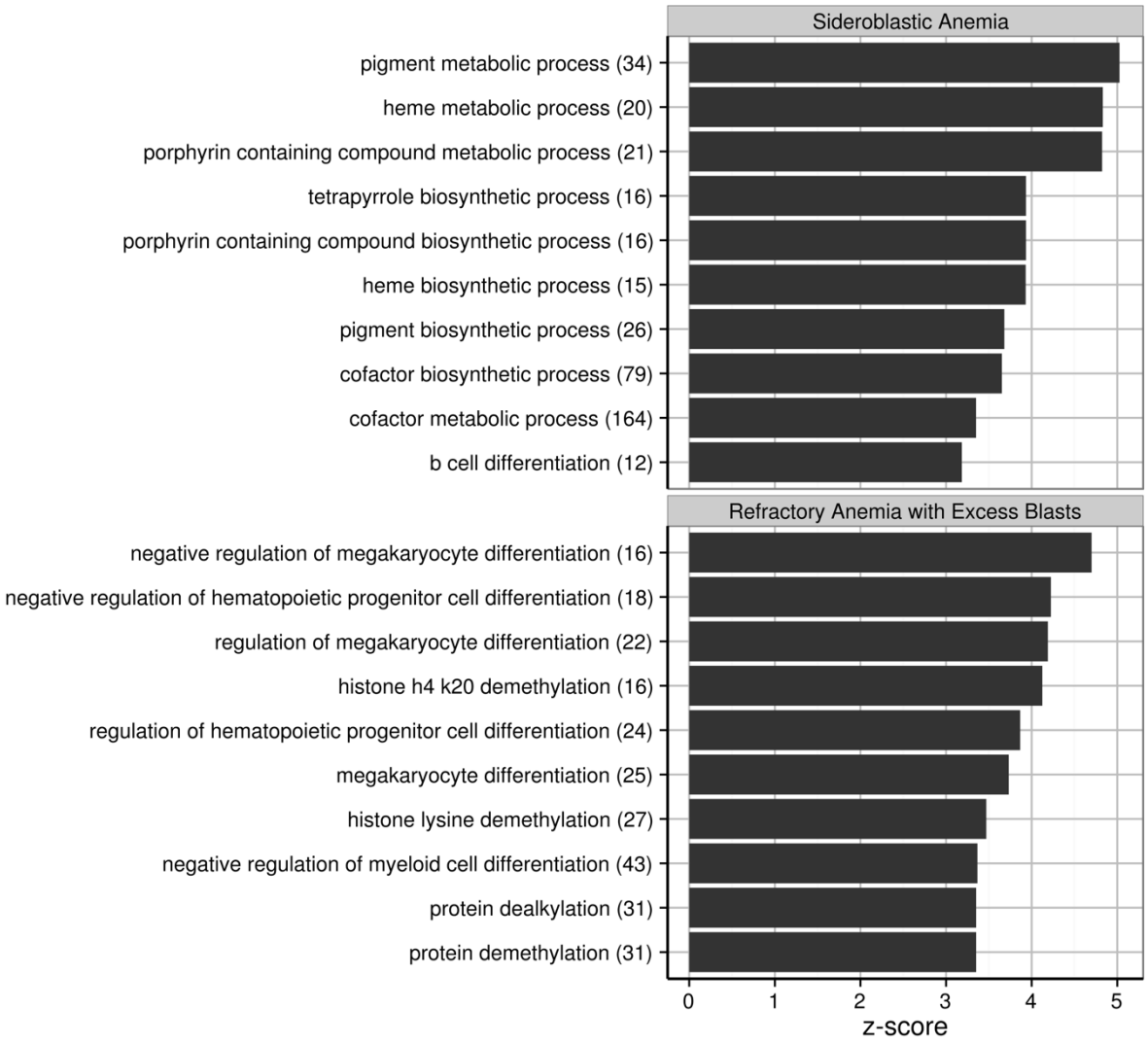


Figure S5. Related to Figure 5. Top 20 GO term enrichments for URSA^{HD}'s Sideroblastic Anemia model and Refractory Anemia with Excess Blasts (RAEB) model. Enrichments for URSA^{HD}'s models accurately describe the inefficient binding and/or transportation of the heme molecule in Sideroblastic Anemia, and the misregulation of hematopoiesis in RAEB.

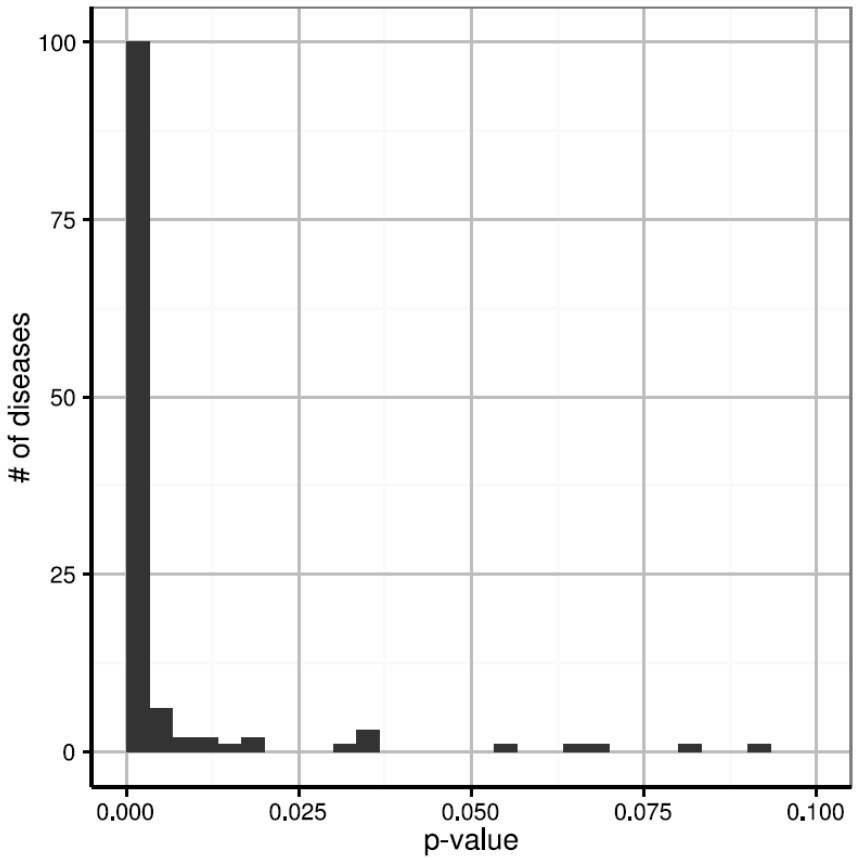


Figure S6. Related to Figure 2 Significant differences between the number of published studies associated with disease genes versus URSA^{HD} top model genes. For each disease, the number of PubMed articles associated with the top model genes was compared to the number of PubMed articles associated with the disease-documented genes and found to be significantly less (unpaired Mann–Whitney U test), implying URSA^{HD} model genes are understudied compared to documented disease genes. The distribution of the p-values for 122 diseases with at least 10 disease-documented genes are shown.