Integrated molecular and clinical analysis for understanding human disease relationships

Winston A. Haynes[1,2,3], Rohit Vashisht[1], Francesco Vallania[3], Charles Liu[3,4], Gregory L. Gaskin[1], Erika Bongen[3,5], Shane Lofgren[1,3], Timothy E. Sweeney[1,3], Paul J. Utz[3], Nigam H. Shah[1,*], Purvesh Khatri[1,3,*]

[1]Stanford Center for Biomedical Informatics Research, Department of Medicine, Stanford University, Stanford, California, USA

[2] Stanford Institute for Immunity, Transplantation, and Infection, Stanford University, Stanford, California, USA

[3]Biomedical Informatics Training Program, Stanford University, Stanford, California, USA

[3]Stanford Institutes of Medicine Research Program, Stanford University, Stanford, California, USA

[3]Stanford Immunology, Stanford University, Stanford, California, USA

*To whom correspondence should be addressed; E-mail: pkhatri@stanford.edu, nigam@stanford.edu

We jointly examined gene-expression and electronic health record data for 104 diseases to identify unbiased clusters of molecularly and clinically related diseases. We performed gene expression meta-analysis of 41,000 samples and computed diseases' clinical profile similarity using 2 million patient records. Based on molecular data, we observed autoimmune diseases clustering with their specific infectious triggers and brain disorders clustering by disease class. In contrast, the electronic health records based clinical profiles clustered diseases according to the similarity of their initial manifestation and later complications. Our integrated molecular and clinical analysis identified diseases with under-appreciated, therapeutically actionable relationships, such as between myositis and interstitial cystitis. This global understanding of relationships between diseases has potential to identify disease causing mechanisms and offer novel therapeutic targets.

1 Introduction

Analyses that examine relationships among human diseases at multiple different scales have the potential to significantly improve our ability to understand disease etiology, postulate therapeutic targets, and connect disparate research communities (*1, 2*). Molecular relationships among diseases have primarily been studied in terms of the genetic similarities derived from either genome wide association studies (GWAS) (*3–10*) or, more recently, phenome wide association studies (PheWAS) (*11–13*). By identifying patterns of genetic architecture that are shared by diseases, such studies can answer questions about the genetic origin of diseases and explain the seemingly unrelated phenotypic effects of genetic variation (*14*).

For understanding the molecular relationships among diseases that go beyond genetics, researchers have leveraged the vast public repositories of gene expression microarray data which include more than 2.2 million samples as of September 2017 (*15, 16*). Early studies to characterize disease similarity using gene expression data (*17–20*), used a small number of data sets and focused on a specific disease group such as autoimmune diseases (*19, 20*) or cancers (*18*). For example, one study, which characterized 54 diseases, each disease had a single dataset, meaning that similarities between diseases might instead represent similarities in the sampled patient cohorts, experimental protocols, or treatment regimens (*17*). Subsequently, researchers are increasingly using much larger datasets, and are utilizing similarities based on drug treatment and gene expression profiles for repurposing existing drugs for novel indications (*21–24*). These approaches typically use gene expression data to represent drug effects and disease state to identify therapeutic relationships based on profile complementarity (*25–28*). Collectively, the GWAS, PheWAS, gene expression similarity, and drug repositioning studies have advanced the "science of medicine" by identifying shared patterns of genetic and transcriptomic architecture among diseases (*14*).

On the other hand, the "practice of medicine", as captured in the electronic medical records (EMRs), can also identify diseases with similar manifestations and offer opportunities for enhancing patient care (*25, 27, 29–32*). When examining disease relatedness using clinical data, most research to date has focused on disease co-occurance and the relative frequency of patients diagnosed with both diseases given the prevalence of each individual disease. These studies have resulted in networks of comorbidities, stratification of patients based on phenotypic information, discovery of clusters of comorbid diseases, recognition temporal disease trajectories, and identification of disease subtypes (*33–40*). However, co-occurrence based approaches require that a patient be diagnosed with both diseases and can miss relatedness of other form. For example, diabetes mellitus patients typically have either type 1 or type 2 diabetes, but not both. However, both sets of patients may take insulin, and are at risk for similar complications, such as macular degeneration. Traditional co-occurence based approaches will simply miss such relationships. In contrast, leveraging similarity in the symptoms, treatment, and downstream manifestations of diseases may yield additional insights which are missed by co-occurrence based methods (*41, 42*).

Finally, ontologies such as the Disease Ontology and the International Classification of Dis-

eases 10 Clinical Modification (ICD10-CM) (*43, 44*) embody our current knowledge and beliefs about disease relationships. These knowledge-based representations provide a declaration of our current "understanding of medicine", from the perspectives of biomedical researchers and health care providers. Disease similarity and relatedness declared in such ontologies are complementary to data-driven analyses, which are agnostic to current beliefs about disease relationships (*45–47*). Thus, these ontologies provide a third vantage point to quantify disease similarity.

To date, no study has combined molecular, clinical, and ontological evidence to define disease relationships. We hypothesized that combining these complementary views of disease similarity would yield novel insights about disease relationships and generate hypotheses about new therapeutic uses of existing drugs. We integrated publicly-available transcriptome profiles of over 41,000 samples from 104 diseases (*48*), electronic health records of over 2 million patients, and ontologies declaring disease relationships. Unlike previous studies that used one dataset per disease, we analyzed multiple independent datasets per disease using MetaIntegrator to identify more robust signatures of diseases that account for both biological and technical heterogeneity (*48,49*). MetaIntegrator has been used to identify diagnostic and prognostic gene signatures, drug targets, and repurpose FDA-approved drugs across a broad range of conditions including organ transplantation, autoimmune diseases, cancer, vaccination, and infections (*23,24,50–57*).

Unlike co-occurrence based approaches, we defined a "clinical profile" of a disease based on a summarization of the diagnosis and procedure codes found in the medical records of patients with that disease. These clinical profiles–analogous to the expression profile of a disease–allowed us to quantify similarity amongst diseases based on the initial manifestations and later complications seen in that disease. Our clinical profile approach captured similarity between diseases co-occurrences can not (e.g., type 1 and type 2 diabetes). A meta-analysis of the gene expression and clinical profile based disease similarities identified candidate disease relationships that persist across both molecular and clinical realms. Finally, by incorporating both ontological and drug indication data, we identified under-appreciated and therapeutically promising disease relationships.

2 Results

2.1 Defining Molecular and Clinical Disease Similarity

We hypothesized that comparison of the molecular and clinical dimensions of disease similarity would provide novel insights into disease relationships. To build molecular disease profiles, we used an integrated multi-cohort analysis framework, implemented in MetaIntegrator, to calculate gene expression profiles for 104 diseases based on 36,915 samples from 615 studies (*58*). We utilized these molecular profiles–comprised of an effect size for each gene–to compute pairwise correlation of all 104 diseases based on those genes with an effect size FDR of less than 5% in either disease [Figure 1:"Disease Molecular Correlation", Figure S1a]. To build the clin-

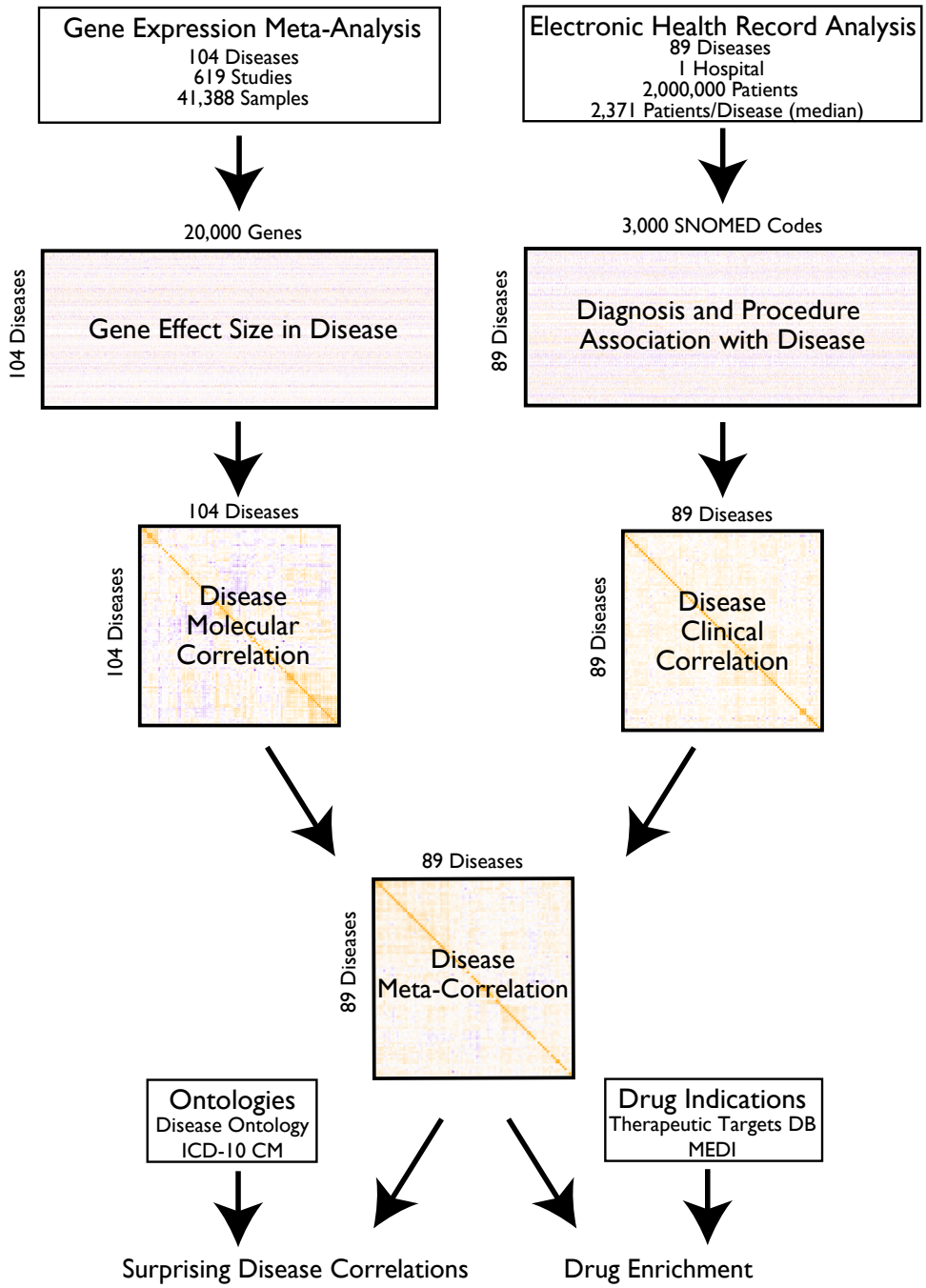


Figure 1: **Integrated analysis across gene expression and electronic health records for novel insight into disease similarity .** Gene expression meta-analysis was performed to calculate effect sizes, a measure of differential expression, of genes across 104 diseases or conditions. In parallel, for 89 of those diseases, we analyzed electronic health records to define "clinical profiles" of diseases based on the association strength of a disease with all other diagnoses and procedure codes occurring in records of patients with that disease. From these effect sizes and clinical profiles, we calculated disease correlations. We performed a meta-analysis of disease correlations to calculate an integrated molecular and clinical disease correlation value.

ical profile of a disease, we used the de-identified medical records of 2 million patients in the Stanford's Clinical Data Warehouse (*59*) [Figure 1:"Electronic Health Record Analysis", Table S2]. We calculated the adjusted odds ratio of that diseases' association with other diagnosis and procedure codes occurring in the records of patients with the disease. In contrast to traditional analyses that compare diseases based on their co-occurrence in patients, we quantify the similarity between diseases based on the pairwise correlation among the clinical profiles of the diseases. Thus, each clinical profile of a disease is a vector of 3,000 diagnosis and procedure codes, with each value representing the association strength of the disease with these codes in patient medical records.

2.2 Comparing Molecular and Clinical Disease Relationships

We performed hierarchical clustering of diseases based on both molecular and clinical correlation values to explore classes of related diseases [Figure 2]. Our molecular analysis identified some interesting clusters that provide illustrative examples of the utility of this approach. First, infectious diseases formed two distinct clusters, one containing all viral infections (maroon) and a neighboring one containing all bacterial infections (green) [Figure 2]. We have previously identified a conserved host response to viral infections (*51*) and shown that host response can distinguish viral and bacterial infections (*56*). These results support the existence of distinct viral and bacterial clusters. More detailed analysis of these infectious disease clusters revealed that the only non infectious disease in the viral infection cluster was systemic lupus erythematosus, a complex autoimmune disease whose pathogenesis is associated with viral infection (*60*). The bacterial infection cluster contained non-infectious diseases, including: Wegener's granulomatosis, sarcoidosis, pulmonary arterial hypertension, scleroderma, vasculitis, chronic myeloid leukemia, and pregnancy. Of these, Wegener's granulomatosis and sarcoidosis are autoimmune diseases that may be triggered by bacterial infection (*61, 62*); vasculitis is associated with both bacterial and viral infections (*63*); and pulmonary arterial hypertension is associated with both sepsis and HIV infection (*64, 65*). Thus, almost all diseases in these two clusters were either infectious or have been repeatedly associated with infection in the literature, which may explain the observed molecular relationships between these diseases.

Second, our analysis also identified surprising and hitherto unknown disease associations. For example, a varied set of brain disorders formed a single cluster (blue) that further grouped into three distinct sub-clusters for neurodegenerative diseases, psychiatric disorders, and brain cancers [Figure 2]. Arguably, this cluster may be confounded by tissue-specific expression profiles because our analysis included data from blood, solid tissue, and sorted cell types. Therefore, we repeated our analysis by using only gene expression profiles from blood samples, and compared these results to those from our combined blood and tissue analysis. Despite using significantly less data, the agreement between the full gene expression analysis and the blood subset was statistically significant in terms of both correlation of the disease correlations (Pearson's correlation = 0.846, bootstrap p-value < 1e-4) and the correlation of the cophenetic distances, a measure of dendrogram similarity (Pearson's correlation =0.519, bootstrap p-value

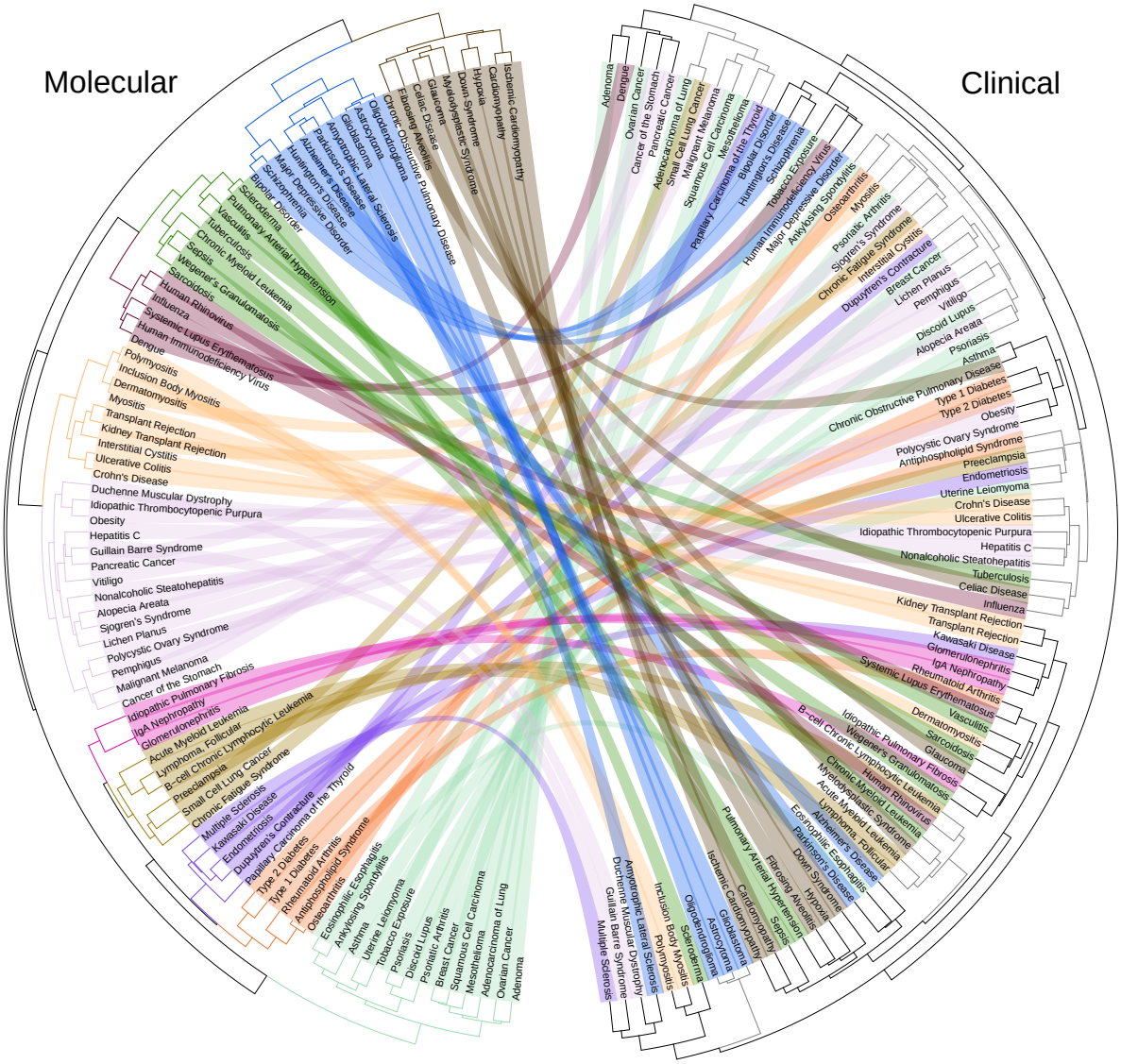


Figure 2: **Comparing molecular and clinical perspectives of disease relationships.** We created a dendrogram for molecular (left) and clinical (right) disease relationships based on pairwise disease correlations from our gene expression meta-analysis and electronic health record analysis, respectively. Arcs connect the same disease in the molecular and clinical dendrograms and are colored based on molecular groupings. See also Figure S1.

< 1e-4) [Figure S1e]. When using only gene expression data from blood, Huntington's disease, Alzheimer's disease, and amyotrophic lateral sclerosis, the only neurological disorders with blood samples, were contained in two neighboring clusters.

Next we compared dendrograms resulting from our clinical and molecular correlation analyses [Figure 2]. The correlation of the cophenetic distance matrices and the correlation matrices were both statistically significant (p < 1e-4), indicating reasonable agreement between molecular and clinical correlations. Despite this agreement, there were important differences between the disease relationships. For instance, in our clinical dendrogram, the brain cluster (blue), with the exception of Huntington's disease and amyotrophic lateral sclerosis, split into three clusters which separately contained the brain cancers, psychiatric disorders, and neurodegenerative diseases. This split in neurological diseases reflects the differences in the clinical practice managing these diseases. On the other hand, the seemingly random dispersal of the bacterial (green) and viral (maroon) clusters was unexpected. It is possible that the wide variation in signs and symptoms of bacterial and viral infections (based on the site of the infection) leads to different clinical profiles for the infectious diseases. This observation further suggests that molecular data are better able to group infectious diseases because the molecular host response for bacterial infections at different sites would still be very similar.

2.3 Integrated Molecular and Clinical Disease Relationships

Disease correlations calculated from gene expression and clinical profiles quantified the molecular and clinical similarities of disease. To combine these two perspectives, and obtain a consensus, we performed a meta-analysis of the molecular and clinical disease correlations [Figure 1:"Integrated Disease Correlation", Figure S2d]. The resulting disease correlations were those that are consistently seen at both the molecular and clinical levels.

In the resulting hierarchical clustering, we observed that many of the disease groups were preserved: (1) bacterial (teal) and viral (blue) infection clusters contained many of the same members and were nearby in the dendrogram; and (2) the brain diseases, except for amyotrophic lateral sclerosis, were entirely contained in two neighboring clusters (purple and grey) [Figure 3]. Further, we observed the emergence of a cluster of diseases that affects the muscles (maroon), including several types of myositis, Duchenne muscular dystrophy, and amyotrophic lateral sclerosis. Most of these clusters were reproducible at a p-value < 0.1 based on permutations of the gene expression and clinical profiles data [Figure S2e]. These results demonstrate that the integrated analysis preserves meaningful groups from the molecular analysis while incorporating the clinical perspective.

To illustrate the clinical utility of our integrated molecular and clinical analysis, we explored whether diseases with indications for the same drug cluster together in our analysis. If the clustering is clinically meaningful, then diseases treated with the same drug will be closer and more highly correlated than the background population. Indeed, we found that diseases treated with the same drugs were significantly closer and more similar to each other than other disease pairs (Wilcox test p-value < 1e-4 in all cases) [Figure S4]. Our results strongly suggest that

Figure 3: **Integrated molecular and clinical disease relationships.** Dendrogram resulting from meta-analysis of molecular and clinical disease correlations. See also Figure S2.

disease pairs that are neighbors in the dendrogram or have highly positive correlation values should (or already do) share therapies.

2.4 Integrated Molecular and Clinical Analysis Identifies Under-appreciated, Therapeutically Promising Disease Relationships

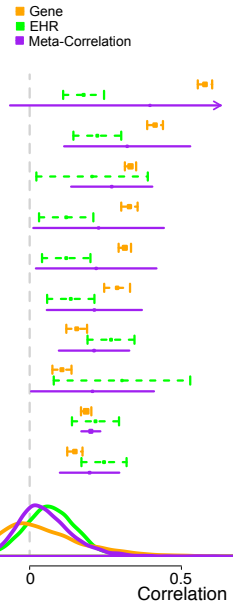
Finally, we compared disease relationships identified in our integrated analysis with ontological representations of disease relationships from the Disease Ontology and ICD10-CM (*43, 44*) to explore the extent to which our analysis reflected current understanding of medicine. We observed statistically significant similarities between them in terms of correlation of the cophenetic distance and correlation matrices ($p < 1e\text{-}4$), demonstrating that our integrated disease similarity recapitulates the current understanding of disease relationships substantially [Figure S2a].

In addition to the general agreement with prior knowledge, we found that the similarity in disease relationships we discovered was also largely preserved when considering specific disease pairs. Disease pairs with the largest integrated correlation values were close to each other in the ontologies, including (1) astrocytoma and oligodendroglioma, (2) Huntington's disease and Parkinson's disease, and (3) malignant melanoma and squamous cell carcinoma [Figure S3a]. Although there is a positive bias of disease correlation values in both the molecular and clinical data, there were few disease pairs with negative integrated correlation values. We observed that diseases with significantly negative integrated correlation values tended to be distant in both the Disease Ontology and ICD10-CM [Figure S3b].

Next, we explored what insights our integrated disease relationships could provide in situations where where one disease has many drugs and the other has few. We focused on novel disease pairs that are absent in current ontological representations, exhibit a disparity in approved drugs, and are highly correlated at the molecular level [Figure 4]. Such disease pairs included (1) psoriasis and squamous cell carcinoma, (2) myositis and interstitial cystitis, (3) and ulcerative colitis and lichen planus. Although it may be possible to connect many of these disease pairs based on retrospective insights, current knowledge-based representations do not these diseases to be related. Further, the majority of these top relationships have been co-mentioned in the same publication thirty times or fewer. Identifying such under-appreciated relationships has potential to benefit both disease research communities (*1, 2*).

Particularly intriguing are associations among orphan diseases with no clinically-useful therapies and diseases with many approved therapies. Consider the example of myositis and interstitial cystitis: myositis has been discussed in 17,503 PubMed publications and has 29 drugs in the MEDI database (*66*). In contrast, interstitial cystitis has been discussed in 1,824 publications and has no drugs in the MEDI database. Despite these diseases having an integrated correlation value of 0.32, stemming from positive correlations at both the molecular and clinical level, a search of PubMed reveals only one paper discussing both diseases. We suggest that researchers pursue synergies between this and other significantly correlated disease pairs.

To demonstrate the potential for the kind of exploratory analyses possible for such disease



Disease 1	Disease 2	Meta−Correlation	Disease Ontology Distance	ICD−10 CM Distance	Disease 1 # Drugs	Disease 2 # Drugs	Number Co-Publications
Psoriasis	Squamous Cell Carcinoma	0.396	9	6	42	0	316
Myositis	Interstitial Cystitis	0.321	9	7	29	0	1
Myositis	Dengue	0.27	9	6	29	0	12
Myositis	Guillain Barre Syndrome	0.227	11	7	29	1	27
Ulcerative Colitis	Lichen Planus	0.219	9	8	16	0	21
Human Immuno-deficiency Virus	Wegener's Granulomatosis	0.213	8	7	50	1	1
Systemic Lupus Erythematosus	Glomerulonephritis	0.212	9	6	26	1	7002
Multiple Sclerosis	Inclusion Body Myositis	0.207	11	6	24	0	2
Adenocarcinoma of Lung	Papillary Carcinoma of the Thyroid	0.202	9	6	25	1	6
Systemic Lupus Erythematosus	IgA Nephropathy	0.197	10	7	26	1	180

Figure 4: **Under-appreciated, therapeutically promising disease correlations.** The figure shows the top 10 disease pairs were distant in both the Disease Ontology and ICD-10 CM, had a disparity in the number of available drugs, and had positive correlation based on molecular, clinical, and integrated analyses. The top 10 pairs are selected by the following criteria: at least 15 drugs in the MEDI database for disease 1, fewer than 3 drugs for disease 2, Disease Ontology distance less than 7, ICD-10 CM distance greater than 5, and strictly positive 95% confidence intervals for both gene and EHR correlations. The density plot shows the background distribution of correlation values. See also Figure S3.

relationships, we examined the genes driving the correlation between myositis and interstitial cystitis. We focused on genes with significant effect sizes in both diseases. The mutually significant genes included: LOC100510692, LGALS9B, TMSB10, APOBEC3A, SNORA14B [Figure S6]. With the exception of APOBEC3A, the other 4 top genes all have one or fewer Gene Ontology annotations, indicating that little is understood about these potentially important genes (*67*). Gene Ontology enrichment analysis of all genes significant in both diseases at an FDR < 5% identified T cell apoptotic processes, microtubule organization, and cytokinesis cytokinesis as significant biological processes (*68*).

3 Discussion

We hypothesized that comparison of the molecular and clinical dimensions of disease similarity would provide novel insights into disease relationships. Therefore, we performed a gene expression meta-analysis of 104 diseases from over 41,000 patient samples to evaluate the molecular similarity of diseases. For 89 of those diseases we constructed clinical profiles that describe a disease as a vector of diagnosis and procedure codes found in the medical records of patients with that disease. Our integrated molecular and clinical analysis represents the first analysis to integrate data across these vastly different scales to assess disease relatedness. As a result, we elucidated disease pairings that were significant in terms of molecular and clinical manifestations. Identification of unexpected pairs of diseases as being significantly associated can connect disparate research communities, provide better understanding of disease causation, and generate possible new therapeutic strategies (*1, 2*). In particular, we identify relationships like myositis and interstitial cystitis where deep knowledge about myositis may strengthen understanding and treatment of interstitial cystitis.

In the molecular analysis, we identified several clusters of diseases that provoke interesting discussion. As an example, infectious diseases separately clustered into bacterial and viral clusters alongside autoimmune diseases, whose onset may be triggered by those infections. Our analysis also found that disease relationships based only on data from blood were similar to relationships discovered when including all data irrespective of sample source. These results agree with prior findings that the effect of disease is stronger than the effect of tissue (*69*). This finding enables opportunities for studying diseases based on circulating blood when affected tissue such as brain, spleen, or pancreas may be technically challenging or dangerous to obtain. We invite researchers to use our gene expression meta-analysis data at [http://metasignature.stanford.edu] to formulate their own new hypotheses about disease pairs to study further.

Our clinical similarity analysis distinguishes itself from prior work because we quantified disease similarity based on the association with all diagnosis and procedure codes seen in patients with a given disease. Our approach captures the overall phenotypic similarity of diseases without the requirement that the same patients be diagnosed with both diseases. Phenotypic profile based correlation will capture diseases which appear similar in terms of their symptoms, clinical lab testing, treatments, and complications. Although this work draws only on

the Stanford electronic health record system, the possibility of integrating multiple electronic health records databases through multi-cohort analysis methods analogous to those common in gene expression data is very exciting. Analyses across multiple clinical centers would lead to a more reproducible and generalizable understanding of the clinical similarity between diseases in terms of how they present to the clinic and how they are managed subsequently.

Since the number of unique pairwise disease comparisons was too large to effectively visualize in a static graphic, we developed an interactive visualization of our data [available at http://metasignature.stanford.edu]. Users can interactively explore all correlation plots from this paper [Figures S1a, S1b, and S2d], including the ability to visualize pairwise-correlation values and focus on subsections of interest.

Since diseases with positive integrated correlation values are more likely to share drugs, we propose exploration of significantly correlated disease pairs for drug repositioning studies. Such exploration is particularly exciting when one disease, such as myositis, has therapeutic options while the other disease, such as interstitial cystitis, lacks effective treatments. Such an investigation into drug repositioning differs from traditional drug repositioning approaches because it is focused on the underlying similarity of diseases instead of the dissimilarity of drug effects and disease state. This perspective is valuable because researchers can explore therapeutic options already available instead of limiting research to a single candidate molecules whose clinical efficacy may not pan out.

We believe that by focusing on the both gene-expression and clinical presentation driving disease similarities, we can increase our understanding of disease relationships in a manner not possible before; leading to novel insights about disease relatedness that can increase our understanding of disease causation as well as provide ideas for therapeutic alternatives.

4 Materials and Methods

4.1 Gene expression data collection and meta-analysis

Gene expression meta-analysis data was compiled from the MetaSignature database (*48*). MetaSignature includes data from manual meta-analysis of over 41,000 samples, 619 studies, and 104 diseases Supplemental Table S1]. Briefly, relevant data were downloaded from Gene Expression Omnibus and ArrayExpress (*15, 16*). Cases and controls were manually labeled for each disease and meta-analysis was performed using the MetaIntegrator package (*48*). We used the Hedges' g summary effect size, standard error, and false discovery rate which the MetaIntegrator package calculates for every gene.

For blood-only gene expression meta-analysis, we manually identified datasets that were derived from whole blood or peripheral blood mononuclear cells. We re-ran the MetaIntegrator analysis pipeline using only these datasets. Similarly, for the tissue and cell-type specific meta-analysis, we identified datasets derived from either solid tissue or sorted cell populations and ran MetaIntegrator using only these datasets. In total, 35 and 89 diseases had at least one dataset

derived from blood and solid tissue/sorted cell, respectively.

4.2 Data collection for disease-gene publications, SNP data, and Gene Ontology annotations

We downloaded the number of publications for each disease-gene relationship from PubPular and HuGE Navigator for as many of the 104 disease in MetaSignature as were present in the databases (102 in PubPular and 81 in HuGE) (*70–72*). PubPular gave the top 261 gene associations, and HuGE gave all known associations. For all correlations, we only considered disease-gene associations with at least 10 publications to limit false positive associations.

We downloaded disease-SNP relationships, including gene mappings, odds ratios, and p-values, from the GWAS Catalog and HuGE Navigator for 61 and 54, respectively, of the 103 diseases in MetaSignature (*73, 74*). From Gene Ontology, we calculated the counts of non-Inferred from Electronic Annotation annotations for all the genes in the MetaSignature database (*67*). The Spearman rank correlation was used for all correlations.

Our plots show the top 10,0000 gene associations for each disease by effect size FDR rank. Correlation calculations do not include a similar limit.

4.3 EHR data collection and analysis

Patient level data was obtained from Stanford Clinical Data Warehouse (CDW), which represents electronic medical records (EMRs) of patients from both Lucile Packard Children hospital and Stanford Hospitals and clinics (*59*). The CDW contains data on about 2 million adult and pediatric patients consisting of demographic information, 25 million clinical encounters, 48 million inpatient and outpatient ICD9-CM diagnosis codes, 157 million laboratory test results along with various types of pharmacy orders, clinical text, surgical reports and radiology reports. The data were converted into OMOP common data model v5 (*75*). The common data model allows consistent representation of the data in using common terminologies such as SNOMED-CT, RxNorms, and LOINC. Of the 104 diseases from the gene expression meta-analysis, we focused on 89, which clearly mapped to diagnosis codes and had at least 100 patients in STRIDE.

We calculated clinical profiles of diseases as follows: The ICD-9 CM diagnosis codes corresponding to each of the 89 diseases were manually curated using literature searches as well as prior published studies on those diseases. For example, for Type-2 diabetes we selected the codes 250.00, 250.02, 250.10, 250.12, 250.20, 250.22, 250.30, 250.32, 250.40, 250.42, 250.50, 250.52, 250.60, 250.62, 250.70, 250.72, 250.80, 250.82, 250.90, 250.92. Each code was then mapped to its corresponding SNOMED-CT code based on the mapping available from Unified Medical Language System (*76*). Cohorts of patients with and without disease were identified based on the presence or absence of the SNOMED-CT codes in a patient medical record [Figure S5].

The first occurrence of a SNOMED-CT code in the longitudinal record of a patient defined an index date. A patient was included in the cohort if at least 90 days of medical data was

available. We also included demographic information such as age and gender. The control group of patients were identified by 1:1 matching of patients using a propensity score model adjusting for age, gender, and total length of record [Figure S5b]. The set of patients with the disease, and their matching controls, were then used to build a time agnostic, binary patient feature matrix capturing the occurrence of all other diseases that are found in the medical records of the individuals with the disease or in the records of the matching controls [Figure S5d]. For each disease, we obtained a matrix in which patients were in rows, and columns were binary indicators for the presence of absence of SNOMED-CT codes in the medical records of a given patient. To obtain a clinical profile of the disease from this matrix, we fit a ridge regression with presence of the disease of interest as the outcome. The resulting clinical profile of a disease is comprised of coefficients for each of the binary indicators (i.e. the SNOMED-CT codes). The exponent of the coefficient can be interpreted as an adjusted odds ratio, quantifying the importance (and association strength) of a specific indicator in the profile of the given disease.

We used 10-fold cross validation while fitting the regression, and coefficients were selected for most regularized lambda. For each cohort, we bootstrapped the ridge regression for the fixed value of lambda (obtained from 10-fold cross validation) to compute the variance associated with the coefficients. Clinical profiles were constructed for each of the 89 diseases under study, and then were used to calculate pair-wise correlations between each disease.

Earlier approaches to infer associations between diseases relied on co-frequency (*7, 33*), which does not control for confounding comorbidities that might drive an observed association. For example, consider Diabetes Mellitus Type 2 (T2DM), Retinopathy and, Renal failure. An association between Retinopathy and Renal failure based on co-frequency will not account for the fact that an observed association between the two could be due to the high association between T2DM and Retinopathy as well as the strong association between T2DM and Renal failure. Using a regression "adjusts" for such effects. To illustrate the effect of such adjustment, we computed both the un-adjusted and adjusted odds ratios for all the disease pairs. 2351 disease pairs lost their positive association they had among them based on co-frequency, thus highlighting the importance of such adjusting when computing associations.

The R packages matching (*77*) and glmnet (*78*) were utilized to perform patient-level matching and ridge regression respectively, while custom SQL queries were written to extract patient-level data from the CDW.

4.4 Correlation calculation

We followed the same process for gene expression and electronic health record data, so we generically refer to each of these matrices as a matrix of features. For each pair of diseases, we calculated the Spearman correlation of all features that were significant at a FDR of less than 5% in either disease. We calculated confidence intervals and p-values for these correlations. We adjusted p-values using the Bonferroni correction.

4.5 Integrated molecular and clinical correlation analysis

We used meta-analysis of correlation values to combine disease correlations from clinical and molecular datasets. We converted correlation values from gene expression and EHR to Fisher's z statistic and applied an inverse variance, random effects meta-analysis model to calculate our integrated correlation value (*79*). To calculate empirical p-values for the correlations, we performed 10,000 bootstrap resamplings of the gene expression and EHR data.

4.6 Clustering calculation

We converted correlation values to a distance as: $1 - \text{cor}$. We performed hierarchical clustering of correlation distances using Ward's clustering criterion (*80, 81*). To assess the robustness of our clustering, we adapted the pvclust R package and performed 10,000 bootstrap resamplings of the terms used for calculating correlation (*82*). For visualization, we leveraged the circlize R package (*83*).

4.7 Ontology analysis

We downloaded OWL formatted data from the Disease Ontology and ICD-10 CM (*43, 44*). We mapped all MetaSignature diseases to SNOMED CUI, which map to 103 diseases in ICD-10 CM and 97 diseases in the Disease Ontology. Using the rdflib library in Python, we calculated the number of steps in the shortest path between pairs of terms by traversing "subClassOf" relationships. From these distances, we calculated a hierarchical clustering of terms using Ward's clustering criterion (*80, 81*). The number of co-publications was determined through a PubMed search using both disease names connected with the 'AND' operator.

4.8 Dendrogram similarity

To numerically compare pairs of dendrograms, we pruned them to include only diseases present in both dendrograms, and calculated both the Pearson correlation of the underlying correlation matrices and the correlation of the cophenetic distance matrices (*84*). We used the R package dendextend for calculating the correlation of the cophenetic distance matrices (*85*). To empirically calculate the significance of both of these correlations, we performed 10,000 random permutations of the dendrogram and correlation matrix labels. For visualization, we used chord diagrams from the R package circlize and tanglegrams from the dendextend package (*83, 85*).

4.9 Drug analysis

We downloaded drug indication data from Therapeutics Target DB and MEDI (*66, 86*). We limited our analysis to drugs with less than ten distinct disease indications matching to the Metasignature database. From Therapeutics Target DB and MEDI we identified 67 and 153

pairs of diseases from MetaSignature with indications for the same drug. We compared the similarity of diseases which share drugs in these databases to all other disease pairs.

4.10 Example disease pair

We explored the relationship of myositis and interstitial cystitis. We identified all genes with an FDR of less than 5% in both diseases. We highlighted those genes with an effect size of greater than 1 or less than 1 in both diseases. For the significant genes, we performed Gene Ontology enrichment analysis using the PANTHER Gene Ontology enrichment analysis tool (Analysis Type: PANTHER Overrepresentation Test (release 20170413), Annotation Version and Release Date: GO Ontology database Released 2017-04-24, Reference List: Homo sapiens (all genes in database), Annotation Data Set: GO biological process complete) (*68*).

5 References

References and Notes

1. S. M. Leach, *et al.*, *PLoS Comput Biol* **5**, 1 (2009).
2. R. B. Altman, *et al.*, *IEEE Intelligent Systems* **14**, 68 (1999).
3. K.-I. Goh, *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **104**, 8685 (2007).
4. B. Bulik-Sullivan, *et al.*, *Nature genetics* **advance on** (2015).
5. J. Menche, *et al.*, *Science* **347**, 1257601 (2015).
6. K. Lage, *et al.*, *Nature biotechnology* **25**, 309 (2007).
7. S. C. Bagley, M. Sirota, R. Chen, A. J. Butte, R. B. Altman, *PLoS computational biology* **12**, e1004885 (2016).
8. J. K. Pickrell, *et al.*, *Nature Genetics* **advance on** (2016).
9. D. Gligorijevic, *et al.*, *Scientific reports* **6**, 32404 (2016).
10. D. R. Blair, *et al.*, *Cell* **155**, 70 (2013).
11. A. Verma, *et al.*, *PLOS ONE* **11**, e0160573 (2016).
12. D. M. Roden, *The Journal of Physiology* (2017).
13. A. Verma, *et al.*, *BMC Medical Genomics* **9**, 32 (2016).

14. N. H. Shah, *Nature biotechnology* **31**, 1095 (2013).

15. R. Edgar, *Nucleic Acids Research* **30**, 207 (2002).

16. A. Brazma, *et al.*, *Nucleic Acids Research* **31**, 68 (2003).

17. S. Suthram, *et al.*, *PLoS computational biology* **6**, e1000662 (2010).

18. D. R. Rhodes, *et al.*, *Neoplasia* **6**, 1 (2004).

19. J. Chen, J. Cerise, A. Jabbari, R. Clynes, A. Christiano, *Cell Systems* **1**, 326 (2015).

20. T. Tuller, S. Atar, E. Ruppin, M. Gurevich, A. Achiron, *Genes and Immunity* **14**, 67 (2013).

21. M. Sirota, *et al.*, *Science translational medicine* **3**, 96ra77 (2011).

22. J. T. Dudley, *et al.*, *Science translational medicine* **3**, 96ra76 (2011).

23. R. Chen, *et al.*, *Cancer Research* **74**, 2892 (2014).

24. P. Khatri, *et al.*, *The Journal of experimental medicine* **210**, 2205 (2013).

25. T. T. Ashburn, K. B. Thor, *Nature Reviews Drug Discovery* **3**, 673 (2004).

26. J. Li, *et al.*, *Briefings in bioinformatics* **17**, 2 (2016).

27. D. Sardana, *et al.*, *Briefings in bioinformatics* **12**, 346 (2011).

28. G. Hu, P. Agarwal, *PLoS ONE* **4**, e6536 (2009).

29. P. B. Jensen, L. J. Jensen, S. Brunak, *Nature Reviews Genetics* **13**, 395 (2012).

30. S. J. Shah, *et al.*, *Circulation* **131**, 269 (2015).

31. Y. S. Low, *et al.*, *Journal of the American Medical Informatics Association* **24**, ocw161 (2016).

32. H. Xu, *et al.*, *Journal of the American Medical Informatics Association* **22**, 179 (2014).

33. C. A. Hidalgo, N. Blumm, A.-L. Barabási, N. A. Christakis, *PLoS Computational Biology* **5**, e1000353 (2009).

34. F. S. Roque, *et al.*, *PLoS Computational Biology* **7**, e1002141 (2011).

35. E. Capobianco, P. LioâĂŹ, *Trends in Molecular Medicine* **19**, 515 (2013).

36. A. B. Jensen, *et al.*, *Nature Communications* **5**, 1769 (2014).

37. E. Capobianco, P. Liò, *Journal of Complex Networks* **3**, 319 (2015).

38. J. X. Hu, C. E. Thomas, S. Brunak, *Nature Reviews Genetics* **17** (2016).

39. L. Li, *et al.*, *Science Translational Medicine* **7** (2015).

40. A. Rzhetsky, D. Wajngurt, N. Park, T. Zheng, *Proceedings of the National Academy of Sciences* **104**, 11694 (2007).

41. K. Jung, *et al.*, *PloS one* **9**, e89324 (2014).

42. G. Wang, K. Jung, R. Winnenburg, N. H. Shah, *Journal of the American Medical Informatics Association : JAMIA* **22**, 1196 (2015).

43. L. M. Schriml, *et al.*, *Nucleic Acids Research* **40** (2012).

44. *National Center for Health Statistic* **1** (2015).

45. L. Cheng, *et al.*, *Scientific Reports* **6**, 30024 (2016).

46. R. Pivovarov, N. Elhadad, *Journal of Biomedical Informatics* **45**, 471 (2012).

47. J. Li, *et al.*, *BMC Bioinformatics* **12**, 266 (2011).

48. W. A. Haynes, *et al.*, *Pac Symp Biocomput* **Web** (2017).

49. T. E. Sweeney, W. A. Haynes, F. Vallania, J. P. Ioannidis, P. Khatri, *Nucleic acids research* **Web**, gkw797 (2016).

50. T. E. Sweeney, A. Shidham, H. R. Wong, P. Khatri, *Science Translational Medicine* **7**, 287ra71 (2015).

51. M. Andres-Terre, *et al.*, *Immunity* **43**, 1199 (2015).

52. M. D. Li, T. C. Burns, A. A. Morgan, P. Khatri, *Acta neuropathologica communications* **2**, 93 (2014).

53. P. K. Mazur, *et al.*, *Nature* **advance on** (2014).

54. P. K. Mazur, *et al.*, *Nature Medicine* **21**, 1163 (2015).

55. T. E. Sweeney, L. Braviak, C. M. Tato, P. Khatri, *The Lancet Respiratory Medicine* **4**, 213 (2016).

56. T. E. Sweeney, H. R. Wong, P. Khatri, *Science translational medicine* **8**, 346ra91 (2016).

57. H.-C. S. P. HIPC-CHI Signatures Project Team, H.-I. HIPC-I Consortium, *Science immunology* **2**, eaal4656 (2017).

58. W. Haynes, A. Tomczak, P. Khatri, *Pacific Symposium on Biocomputing* (2017).

59. H. J. Lowe, T. A. Ferris, P. M. Hernandez, S. C. Weber, *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* **2009**, 391 (2009).

60. J. A. James, *et al.*, *Arthritis and Rheumatism* **44**, 1122 (2001).

61. H. Tadema, P. Heeringa, C. G. M. Kallenberg, *Current opinion in rheumatology* **23**, 366 (2011).

62. O. M. Kon, R. M. duBois, *Thorax* **52**, S47 (1997).

63. C. Pagnoux, P. Cohen, L. Guillevin, *Clinical and experimental rheumatology* **24**, 71 (2006).

64. K. M. W. Tauil, E. Gaio, C. A. Melo-Silva, R. S. Carvalho, V. M. Amado, *Medical hypotheses* **83**, 290 (2014).

65. S. Janda, B. S. Quon, J. Swiston, *HIV medicine* **11**, 620 (2010).

66. W.-Q. Wei, *et al.*, *Journal of the American Medical Informatics Association : JAMIA* **20**, 954 (2013).

67. M. Ashburner, *et al.*, *Nature genetics* **25**, 25 (2000).

68. H. Mi, A. Muruganujan, P. D. Thomas, *Nucleic acids research* **41**, 377 (2013).

69. J. T. Dudley, R. Tibshirani, T. Deshpande, A. J. Butte, *Molecular systems biology* **5**, 307 (2009).

70. W. Yu, M. Clyne, M. J. Khoury, M. Gwinn, *Bioinformatics* **26**, 145 (2010).

71. Maggie Lam, PubPular: Identifying the focus of biomedical research.

72. M. P. Y. Lam, *et al.*, *Journal of proteome research* **Web** (2016).

73. W. Yu, *et al.*, *European Journal of Human Genetics* **19**, 1095 (2011).

74. D. Welter, *et al.*, *Nucleic acids research* **42**, 1001 (2014).

75. F. FitzHenry, *et al.*, *Applied clinical informatics* **6**, 536 (2015).

76. O. Bodenreider, *Nucleic acids research* **32**, 267 (2004).

77. J. S. Sekhon, *Journal of Statistical Software* **55**, 1 (2011).

78. J. Friedman, T. Hastie, R. Tibshirani, *Journal of Statistical Software* **33**, 1 (2010).

79. M. Borenstein, L. V. Hedges, J. P. T. Higgins, H. R. Rothstein, *Introduction to Meta-Analysis* (2009).

80. J. H. Ward, Hierarchical grouping to optimize an objective function (1963).

81. F. Murtagh, P. Legendre, *Journal of Classification* **31**, 274 (2014).

82. R. Suzuki, H. Shimodaira, *Bioinformatics* **22**, 1540 (2006).

83. Z. Gu, L. Gu, R. Eils, M. Schlesner, B. Brors, *Bioinformatics* **30**, 2811 (2014).

84. R. R. Sokal, F. J. Rohlf, *Taxon* **11**, 33 (1962).

85. T. Galili, *Bioinformatics* **31**, 3718 (2015).

86. F. Zhu, *et al.*, *Nucleic acids research* **40**, 1128 (2012).

6 Author Contributions

Conceptualization, W.A.H., R.V., P.J.U., N.H.S., and P.K; Methodology, W.A.H., R.V., N.H.S., and P.K; Software, W.A.H. and R.V.; Investigation, W.A.H. and R.V.; Data Curation, W.A.H., R.V., F.V., C.L, G.L.G, E.B., S.L., T.E.S.; Writing- Original Draft, W.A.H., R.V., P.J.U. N.H.S., and P.K.; Writing- Reviewing and Editing, W.A.H., R.V., F.V., C.L, G.L.G, E.B., S.L., T.E.S., P.J.U. N.H.S., and P.K.; Visualization, W.H.; Funding Acquisition, P.J.U., N.H.S., and P.K.

7 Acknowledgements

We thank Alex Schrenchuk for computer support. WAH is funded by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747. FV is funded by the National Institute of Health K12 Career Award 5K12HL120001-02. EB is funded by Gabilan Fellowship. PK is funded by the National Institute of Allergy and Infectious Diseases grants 1U19AI109662, U19AI057229, U54I117925, and U01AI089859. PJU is the recipient of a gift from Gustav Floren Family Trust, and a gift from The Ben May Charitable Trust of Mobile, Alabama. PJU is supported by NIH grants NIAID U19-AI1110491, Stanford Autoimmunity Center of Excellence (ACE); NIAID 1 UM1AI110498-01, ACE Collaborative Project; 1 U19-AI090019 ; NIAID 1 R01 AI125197-01; and Alliance for Lupus Research, Grant Number 21858. NHS is supported in part by NLM grant R01 LM011369, NIGMS grant R01 GM101430, and a research award from Janssen R&D LLC.

8 Supplemental Information

Table S1: Sample size and number of studies for all 103 diseases in MetaSignature.

Disease Name	Sample Size	Number of Studies
Acute Myeloid Leukemia	64	2
Acute Promyelocytic Leukemia	10	1
Adenocarcinoma of Lung	1524	17
Adenoma	38	2
Alopecia Areata	17	2
Alzheimer's Disease	997	9
Amyotrophic Lateral Sclerosis	38	2
Ankylosing Spondylitis	56	3
Antiphospholipid Syndrome	6	1
Asthma	130	2
Astrocytoma	83	2
B-cell Chronic Lymphocytic Leukemia	124	3
Bacterial Infection	82	3
Bipolar Disorder	219	6
Breast Cancer	1840	15
Cancer of the Stomach	43	2
Cardiomyopathy	706	13
Celiac Disease	1372	10
Chronic Fatigue Syndrome	148	5
Chronic Myeloid Leukemia	56	2
Chronic Obstructive Pulmonary Disease	58	4
Crohn's Disease	160	5
Dengue	687	4
Dermatomyositis	294	11
Discoid Lupus	38	1
Down Syndrome	25	2
Duchenne Muscular Dystrophy	137	3
Dupuytren's Contracture	34	4
Endometriosis	241	9
Eosinophilic Esophagitis	10	1
Familial Combined Hyperlipidemia	33	2
Fibrosing Alveolitis	6	1
Glaucoma	78	2
Glioblastoma	83	2
Glomerulonephritis	102	4
Guillain Barre Syndrome	14	1
Hepatitis C	106	5
Human Immunodeficiency Virus	133	3

Disease Name	Sample Size	Number of Studies
Human Rhinovirus	140	3
Huntington's Disease	587	6
Hypoxia	10	2
Idiopathic Fibrosing Alveolitis	294	9
Idiopathic Pulmonary Fibrosis	340	8
Idiopathic Thrombocytopenic Purpura	4	1
IgA Nephropathy	195	7
Inclusion Body Myositis	17	1
Influenza	1346	13
Interstitial Cystitis	57	4
Ischemic Cardiomyopathy	38	2
Kawasaki Disease	89	2
Kidney Transplant Rejection	1486	11
Lichen Planus	47	3
Major Depressive Disorder	62	2
Malignant Melanoma	61	2
Mesothelioma	108	4
Meta-Bacterial	726	8
Meta-Virus	437	9
Multiple Sclerosis	1668	19
Myelodysplastic Syndrome	89	2
Myositis	273	11
Narcolepsy	20	1
Non-small Cell Lung Cancer	1339	9
Nonalcoholic Steatohepatitis	92	3
Obesity	281	4
Oligodendroglioma	83	2
Osteoarthritis	46	2
Ovarian Cancer	386	7
Pancreatic Cancer	294	9
Papillary Carcinoma of the Thyroid	32	2
Parkinson's Disease	236	11
Pemphigus	24	1
Polycystic Ovary Syndrome	70	3
Polymyositis	17	1
Preeclampsia	346	12
Pregnancy	171	3
Primary Open Angle Glaucoma	9	2
Psoriasis	488	12
Psoriatic Arthritis	38	2

Disease Name	Sample Size	Number of Studies
Pulmonary Arterial Hypertension	227	4
Respiratory Syncytial Virus	251	6
Rheumatoid Arthritis	680	20
Sarcoidosis	487	10
Schizophrenia	305	8
Scleroderma	562	14
Senescence	101	5
Sepsis	788	9
Sjogren's syndrome	35	1
Small Cell Lung Cancer	581	6
Squamous Cell Carcinoma	36	2
Systemic Lupus Erythematosus	3748	21
Tobacco Exposure	603	8
Transplant Rejection	251	8
Type 1 Diabetes	1196	19
Type 2 Diabetes	122	2
Ulcerative Colitis	467	20
Uterine Leiomyoma	38	3
Vasculitis	29	1
Vitiligo	10	1
Wegener's Granulomatosis	127	1

Table S2: Number of patients for each of the 89 diseases we examined from the Stanford EHR.

Disease Name	Number of Patients
Acute Myeloid Leukemia	3172
Adenocarcinoma of Lung	6214
Adenoma	3785
Alopecia areata	743
Alzheimers disease	2339
Amyotrophic Lateral Sclerosis	379
Ankylosing spondylitis	842
Antiphospholipid syndrome	1446
Asthma	38580
Astrocytoma	2565
B-cell Chronic Lymphocytic Leukemia	4263
Bipolar Disorder	5934
Breast cancer	12718
Cancer of the Stomach	1850
Cardiomyopathy	11448

Disease Name	Number of Patients
Celiac disease	1412
Chronic fatigue syndrome	1605
Chronic myeloid leukemia	1135
Chronic obstructive pulmonary disease	55856
Crohns disease	2996
Dengue	29
Dermatomyositis	506
Discoid lupus	524
Down syndrome	987
Duchenne Muscular Dystrophy	436
Dupuytrens contracture	1441
Endometriosis	7435
Eosinophilic gastritis	569
Fibrosing alveolitis	928
Glaucoma	12057
Glioblastoma	2565
Glomerulonephritis	2091
Guillain Barre Syndrome	562
Hepatitis C	9644
Human immunodeficiency virus	1646
Human Rhinovirus	144
Huntingtons Disease	65
Hypoxia	5711
Idiopathic Thrombocytopenic Purpura	651
IgA nephropathy	4275
Inclusion body myositis	26
Influenza	2246
Interstitial Cystitis	2000
Ischemic Cardiomyopathy	5714
Kawasaki Disease	267
Kidney Transplant Rejection	1462
Lichen planus	770
Lymphoma Follicular	1814
Major depressive disorder	8571
Malignant melanoma	3015
Mesothelioma	6965
Multiple sclerosis	2564
Myelodysplastic Syndrome	1454
Myositis	16397
Nonalcoholic Steatohepatitis	22142

Disease Name	Number of Patients
Obesity	33551
Oligodendroglioma	2565
Osteoarthritis	37864
Ovarian Cancer	2758
Pancreatic cancer	2929
Papillary Carcinoma of the Thyroid	3214
Parkinsons disease	4055
Pemphigus	262
Polycystic ovary syndrome	2140
Polymyositis	240
Preeclampsia	2371
Psoriasis	4623
Psoriatic Arthritis	1060
Pulmonary arterial hypertension	7467
Pulmonary fibrosis	6524
Rheumatoid Arthritis	1031
Sarcoidosis	1477
Schizophrenia	2613
Scleroderma	1073
Sepsis	9127
Sjogrens syndrome	1141
Small Cell Lung Cancer	6214
Squamous Cell Carcinoma	1668
Systemic Lupus Erythematosus	3869
Tobacco Exposure	19568
Transplant Rejection	3528
Tuberculosis	1513
Type 1 Diabetes	7882
Type 2 Diabetes	45544
Ulcerative colitis	3573
Uterine Leiomyoma	13325
Vasculitis	1255
Vitiligo	953
Wegener's granulomatosis	300

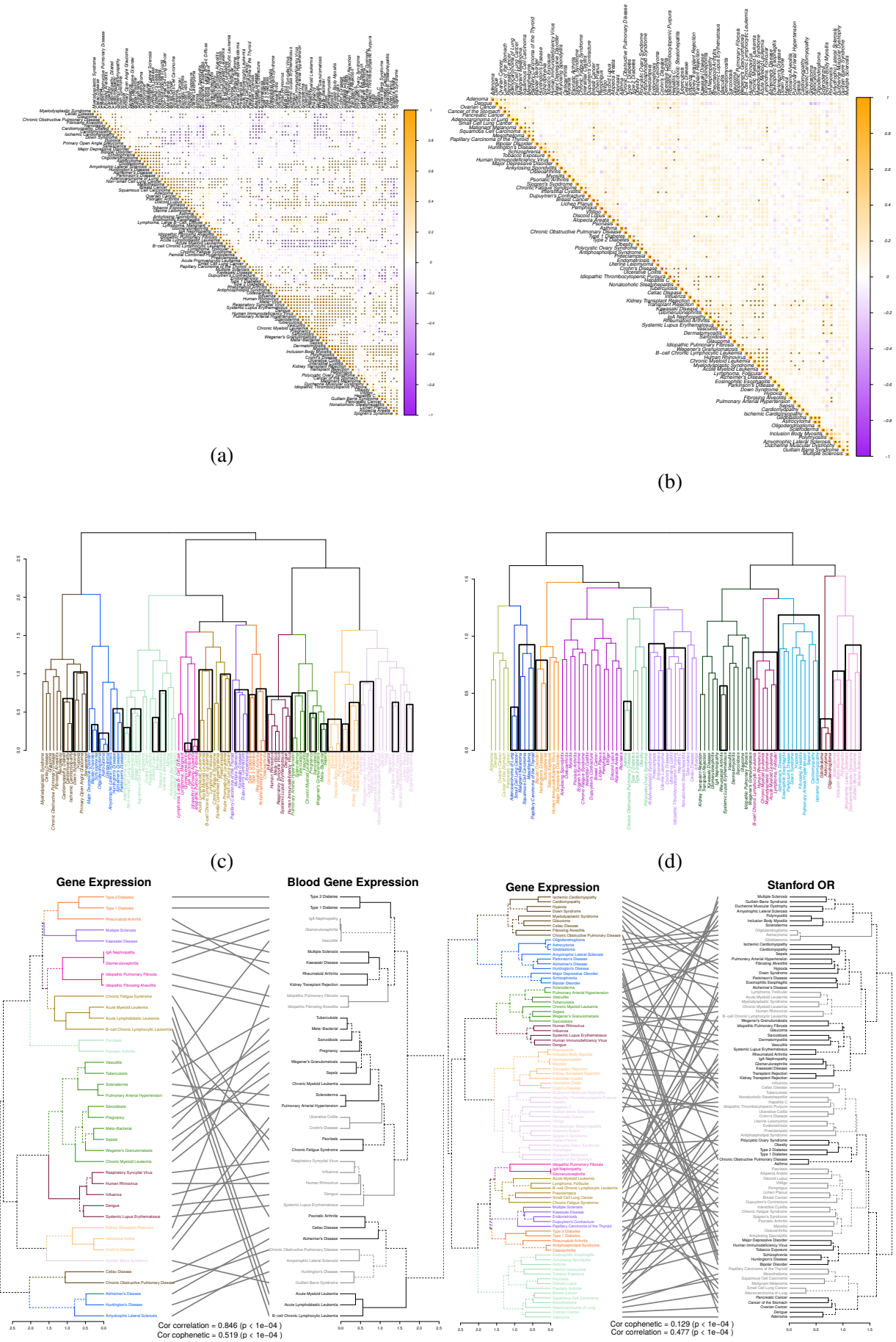
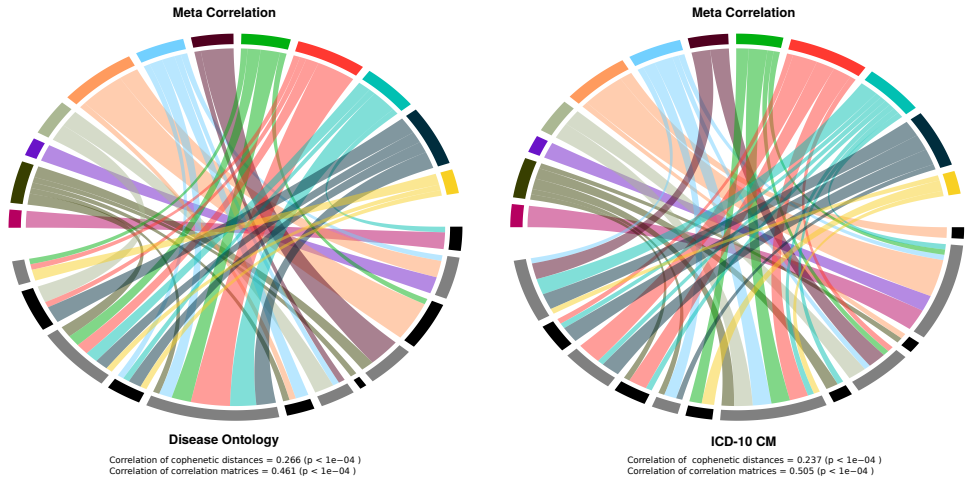
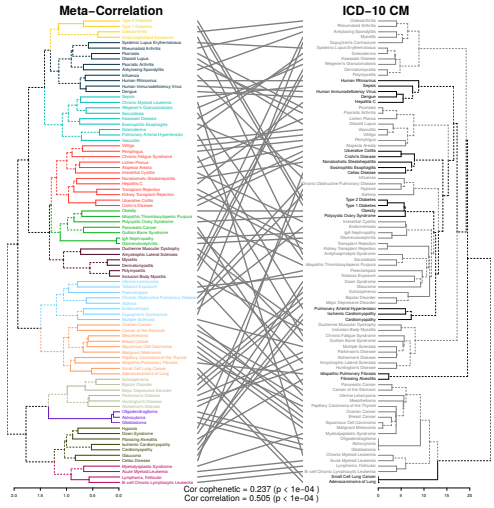


Figure S1: (caption on next page)

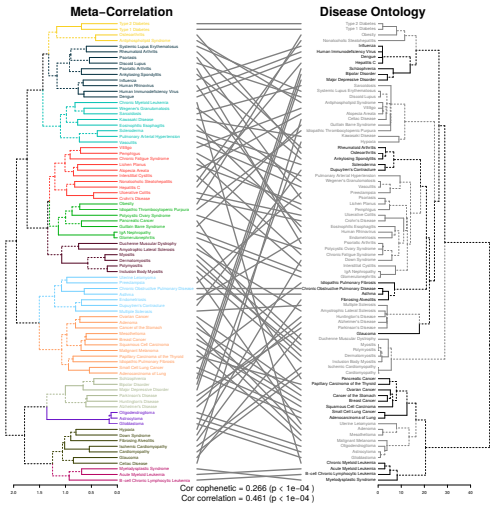
Figure S1: Related to Figure 2. (a) Pairwise correlations of diseases based on gene expression meta-analysis. Stars indicate bonferroni corrected p-values < 1e-10. (b)(c) Pairwise correlations of diseases based on Stanford electronic health records. Stars indicate bonferroni corrected p-values < 1e-5. (c) Bootstrap reproducibility of gene expression dendrogram. Outlined branches are reproducible at approximately unbiased p-value < 0.1. (d) Bootstrap reproducibility of electronic health record dendrogram. Outlined branches are reproducible at approximately unbiased p-value < 0.1. (e) Gene expression clustering based on all samples vs. blood only samples. (f) Clustering based on gene expression vs. electronic health records.



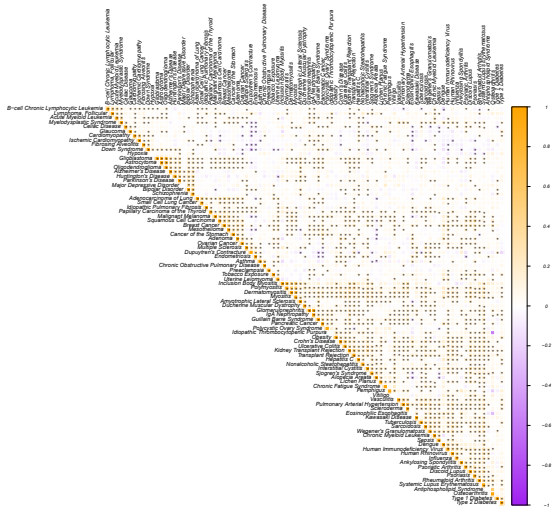
(a)



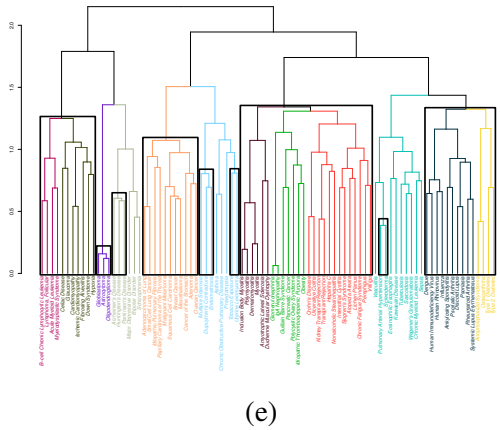
(b)



(c)



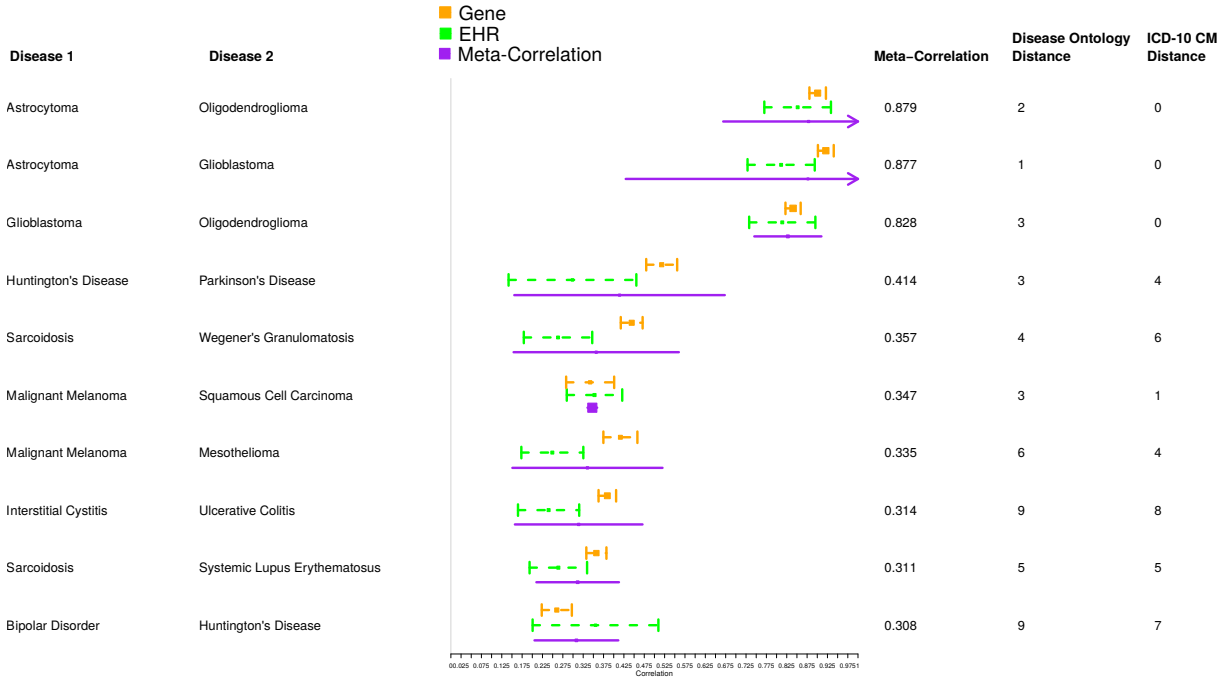
(d)



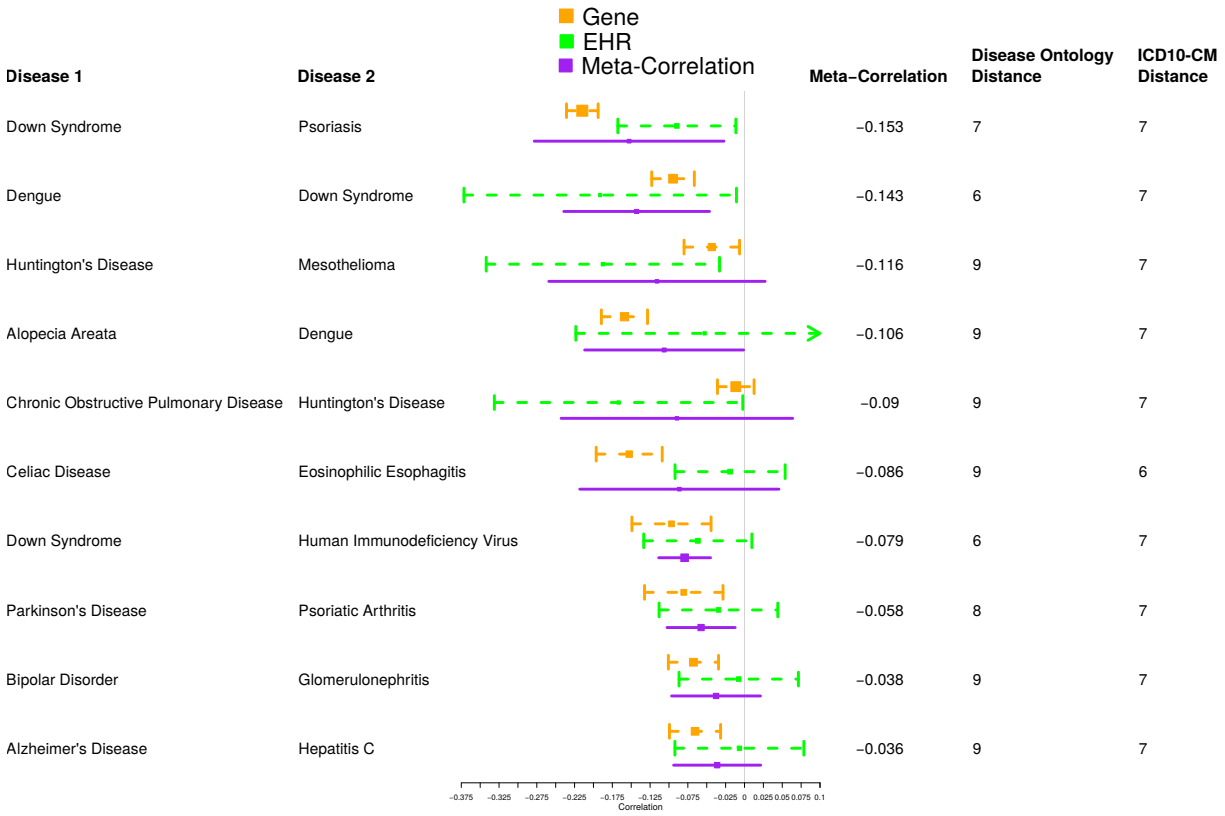
(e)

Figure S2: (caption on next page)

Figure S2: Related to Figure 3. (a) Comparison of the clustering from the meta-correlation analysis to the Disease Ontology and ICD10 Clinical Modification clusterings, respectively (*43, 44*). Colored arcs correspond to the clusters from Figure 3. Flows between arcs represent correspondence of disease cluster members. (b) Clustering based on meta-correlation vs. ICD-10 Clinical Modification. (c) Clustering based on meta-correlation vs. Disease Ontology. (d) Pairwise correlations of diseases based on integrated molecular and clinical analysis. Stars indicate bootstrap p-values < 1e-3. (e) Bootstrap reproducibility of integrated molecular and clinical correlation dendrogram. Outlined branches are reproducible at approximately unbiased p-value < 0.1.



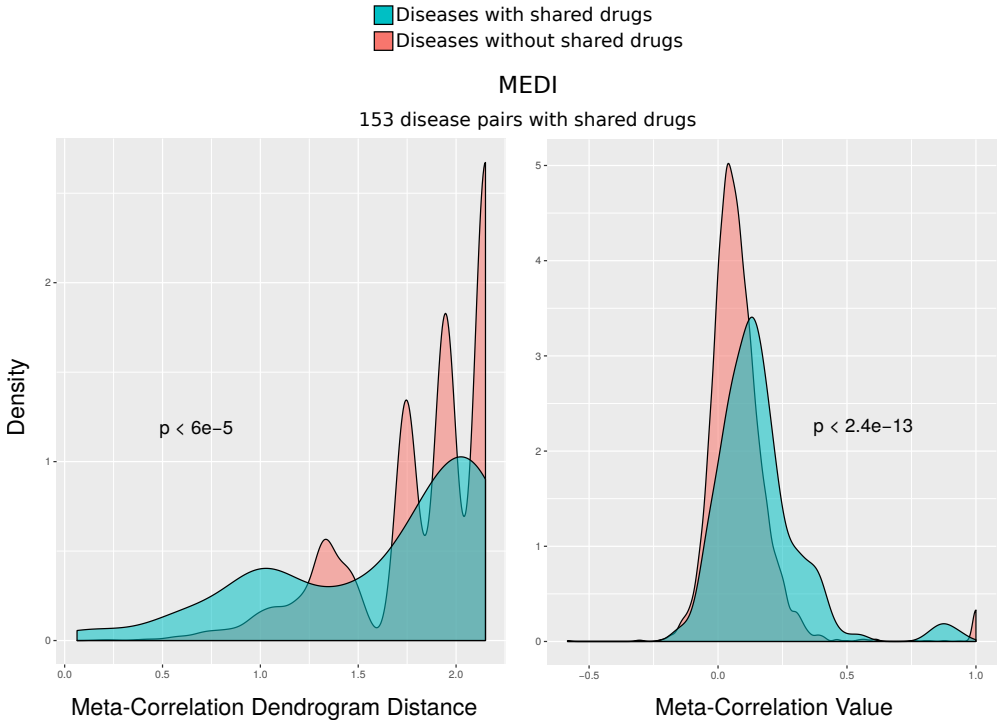
(a)



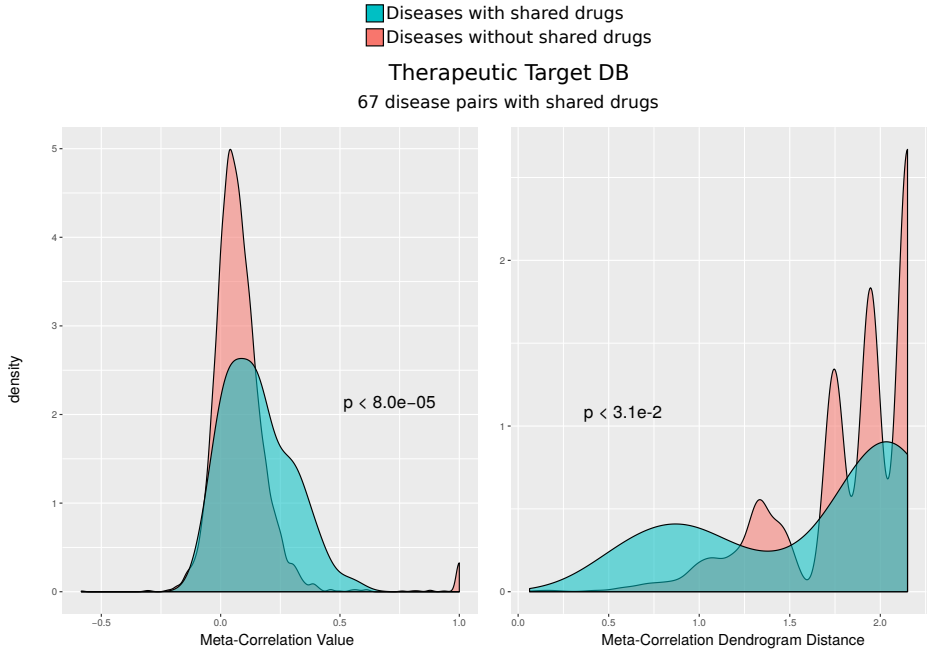
(b)

Figure S3: (caption on next page)

Figure S3: Related to Figure 4. (a) Sampling of significant pairwise meta-correlations. Displayed meta-correlations had a p-value < 1e-3, gene expression and EHR correlations with the same directionality, and mappings to the Disease Ontology. The top 10 pairs of diseases in terms of meta-correlation that passed these criterion are displayed. (b) Meta-correlations with largest negative values. Displayed meta-correlations had negative meta-correlation values, gene expression and EHR correlations with the same directionality, and mappings to the Disease Ontology. The top 10 pairs of diseases in terms of meta-correlation that passed these criterion are displayed.



(a)



(b)

Figure S4: (caption on next page)

Figure S4: **Disease pairs which share drug treatments are enriched in integrated molecular and clinical correlation analysis.** We identified pairs of diseases for which the same drug is indicated in MEDI (*66*) and Therapeutics Target DB (*86*). We compared these disease pairs to the diseases which did not share indications in terms of the distance in the integrated molecular and clinical correlation dendrogram and the integrated correlation values. Across all comparisons, the diseases which shared drugs were significantly more similar than the background population by the Wilcox rank-sum test (p-values indicated on plots).

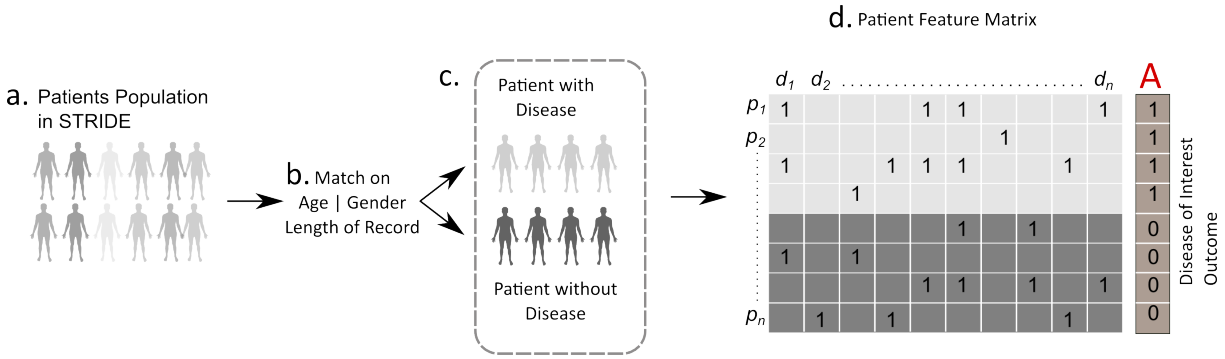


Figure S5: Procedure to compute disease odds ratios a) Patient population in STRIDE b) Patient with disease matched with other patients in the population based on age, gender and length of record to find similar patients without disease c) A disease cohort comprising patients with disease and matched patients without disease d) Patient feature matrix of a given disease cohort where rows are the patients in cohort and column are other diseases seen in the record of a given patient.

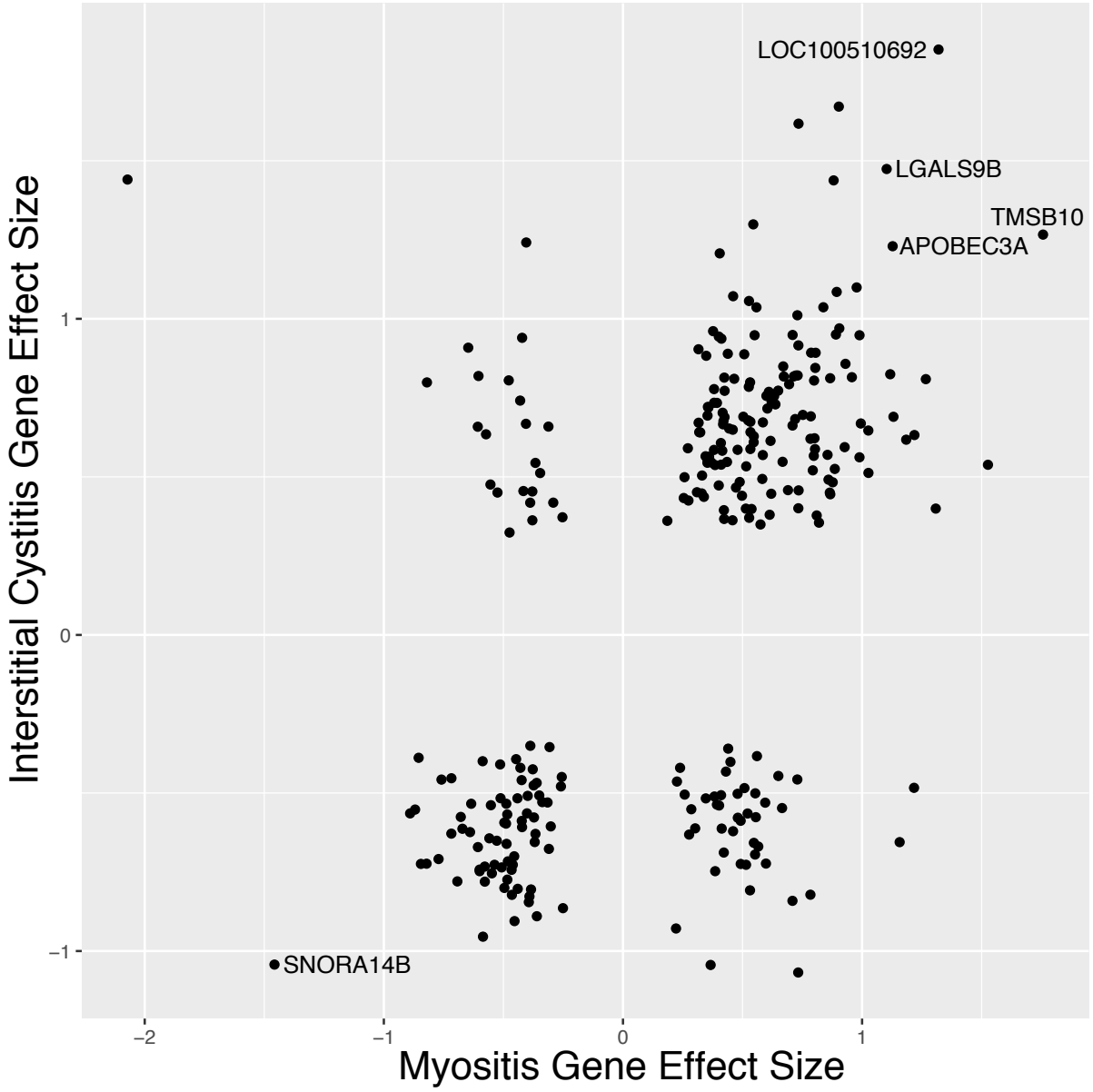


Figure S6: Genes that were significant at an FDR of less than 5% in both myositis and interstitial cystitis. Text labels identify the genes that were most consistently expressed in both diseases.