A SYSTEMATIC ASSESSMENT OF LINKING GENE EXPRESSION WITH GENETIC VARIANTS FOR PRIORITIZING CANDIDATE TARGETS

HUA FAN-MINOGUE[*]
Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: fminogue@stanford.edu

BIN CHEN[*]
Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: binchen1@stanford.edu

WERONIKA SIKORA-WOHLFELD
Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: wsikora@stanford.edu

MARINA SIROTA
Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: msirota@stanford.edu

ATUL J BUTTE[†]
Division of System Medicine, Department of Pediatrics, Stanford University
Stanford, CA 94305, USA
Email: abutte@stanford.edu

Gene expression and disease-associated variants are often used to prioritize candidate genes for target validation. However, the success of these gene features alone or in combination in the discovery of therapeutic targets is uncertain. Here we evaluated the effectiveness of the differential expression (DE), the disease-associated single nucleotide polymorphisms (SNPs) and the combination of the two in recovering and predicting known therapeutic targets across 56 human diseases. We demonstrate that the performance of each feature varies across diseases and generally the features have more recovery power than predictive power. The combination of the two features, however, has significantly higher predictive power than each feature alone. Our study provides a systematic evaluation of two common gene features, DE and SNPs, for prioritization of candidate targets and identified an improved predictive power of coupling these two features.

[*] co-first author
[†] corresponding author

1. Introduction

A major goal of biomedical research is to identify disease genes to guide drug discovery that aims to improve the disease outcomes (1). Genes are defined as disease genes when they carry disease-causing aberrations (2). To identify an aberration of a gene, or a gene feature, and prove it as a causal link between the gene and a disease involves experimental testing and is time consuming. The advancement in high-throughput experimental techniques has facilitated this process by enabling rapid generation of vast amount of data for disease-associated gene features. Those techniques include the gene expression microarray, which allows the study of differential gene expression (DE) between disease and control samples; and high-throughput genotyping and next generation sequencing, which allows the study of disease-associated single nucleotide polymorphisms (SNPs) by comparing disease and control populations. However, these disease-associated features could be assigned to thousands of candidate genes. Prioritizing genes by incorporating these features for further experimental testing of causal relation is therefore necessary to narrow down the search space and increase the effectiveness of translating these candidates (3).

DE is often considered when prioritizing candidate genes, largely because it has been widely used to discover differentially regulated genes and deregulated molecular mechanisms (4). However, it has also been shown that DE genes might not perform well for specific diseases, where highly differentiated genes were not directly related to diseases (5). Yet, whether it can be generalized for all diseases is not clear and most researchers still use DE genes as their primary choice for seeking molecular explanations of biological phenotypes. SNPs to phenotype associations from genome-wide association studies provide unbiased screens of common variant associations. Using disease-associated SNPs to prioritize candidate genes are on the rise, especially as the sequencing technology is getting cheaper and more comprehensive computational tools have been developed to facilitate the process of the raw sequencing data. However, disease-associated SNPs derived from a defined population could fail in a larger or different population (6) and how SNPs perform across different disease conditions is largely unknown.

Increasing effort has been put to link different types of gene features from different sources to improve the performance of each individual feature. As an example, highly differentially expressed genes were found more likely to harbor disease-associated SNPs (7). However, how this feature combination would affect the candidacy of the gene for target validation has not been studied. More comprehensive integration of genetic variants with other types of genomic and biological data has been performed in individual disease condition (8). Although it showed great promise of using genetics to guide drug discovery, whether this can be generalized for other disease conditions is not clear.

An objective assessment of the performance of DE genes and disease-associated SNPs alone or in combination in different disease conditions will help understand the utility of these features and provide guidance to the application of them for target prioritization. However, that

type of assessment is currently lacking, mainly because it will require multiplex data collection and incorporation between features across disease conditions.

In this study, we integrated gene expression with disease-associated SNPs and therapeutic target data sets across a diverse set of 56 diseases in 12 disease categories (Figure 1). We systematically evaluated how successful DE genes, disease-associated SNPs or the combination of both can recover known disease targets, and how well they can predict the known targets by comparing with random sampling of these features. We demonstrate that the performance of DE genes, disease-associated SNPs or the combination of both varies across diseases. We observe that both DE genes and disease-associated SNPs have more recovery power than predictive power. The combination of the two features, however, has more predictive power than each feature alone. This suggests linking DE genes with disease-associated SNPs improves the accuracy of prioritizing candidate targets.

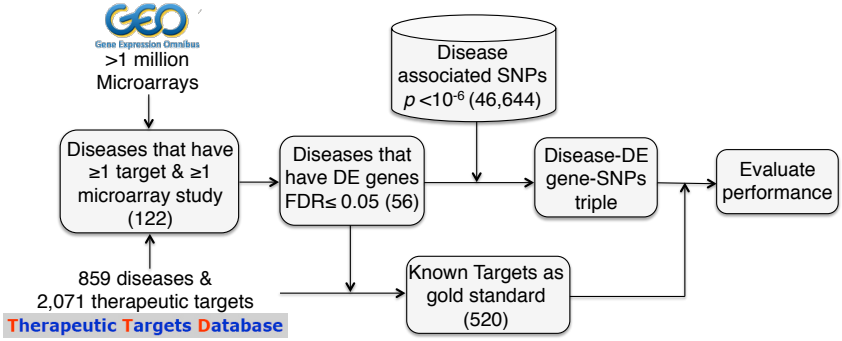

Figure 1. The schematic diagram of the work flow.

2. Methods

2.1 *Selection of diseases*

To examine the relation between gene expression and disease targets, we focused our study on diseases that have at least one gene expression microarray study and one known target.

To identify diseases and their associated microarray data, we utilized a text mining approach using previously published methods (9,10). Briefly, Gene Expression Omnibus (GEO) experiments that are relevant to human diseases and measure both normal and disease states were collected by an automated annotation and mapping between the Medical Subject Heading (MeSH) terms of the experiment associated publications and the disease concepts in the Unified Medical Language System (UMLS). Disease annotations and the associated microarray datasets were manually reviewed in a post-processing step to ensure accuracy. The resulting datasets included 238 disease concepts and 8,435 microarray samples.

To identify diseases that have at least one known target, we used the Therapeutic Target Database (TTD) (11), which provides manually annotated information about known therapeutic targets, their targeted disease conditions and corresponding drugs. It had 897 disease conditions

and 2,071 therapeutic targets (accessed in Aug. 2013), which include targets that are successful, in clinical trials, in pre-clinical research or discontinued. We extracted the UniProt IDs of the targets and mapped them to human Gene IDs. Then we converted the disease conditions of successfully mapped targets to UMLS concept IDs using the MetaMap (12). The maximum confidence score of 1000 was used as a cutoff for the successful mapping. The resulting dataset consists of disease-known target pairs that are represented by the disease concept ID and the target gene ID across 859 diseases and 2,071 therapeutic targets.

Next, we mapped the disease concept IDs between the disease-microarray and the disease-target datasets, which resulted in 122 diseases (Figure 1). These diseases that have at least one known therapeutic target and one gene expression microarray study were further analyzed for their DE genes.

2.2 *Determination of DE genes and disease-associated SNPs*

We used the method of significance analysis of microarray (SAM) (13) and its Bioconductor R package (siggenes) to identify DE genes for each of the 122 diseases. For diseases that have multiple associated studies, the study that has the largest sample size was chosen. For genes with multiple probes, the expression level of the probe that had the highest absolute value was used. The raw data of the microarrays were processed and normalized as described in our previous publication (14). With a false discovery rate (FDR) < 0.05, 56 diseases were found to have at least one DE gene, which includes a total of 17,409 unique DE genes across all the diseases.

To identify the associated SNPs of these 56 diseases, we utilized a human disease-SNP association database (VARIMED) (15,16). In a recent release (Sep. 2013), we have manually-curated over 466,000 disease-associated SNPs across about 6,600 associated diseases and related phenotypes from 17,088 publications. To evaluate the performance of using SNPs for recovering known targets, we used a cutoff $p<10^{-6}$ and obtained 46,644 disease-associated SNPs from VARIMED. The SNP associated disease names were then mapped to concept IDs of the 56 diseases that have at least one DE genes and at least one known target. Thirty-eight diseases were assigned with at lease one SNPs. Unassigned diseases were marked as having 0 SNPs in Table 1. SNPs associated genes were obtained from the dbSNP138 database. Linkage disequilibrium (LD) effect was not counted for selecting disease-associated SNPs to obtain a general pool of SNPs.

By combining the disease-DE genes dataset with the disease-SNPs dataset, we built 129,905 triples between the 56 diseases, their DE genes and associated SNPs. The resulted dataset was mapped with the gold standard of disease targets, which allowed us to examine how often DE genes and associated SNPs alone or in combination can recover and predict the known targets of each disease.

2.3 *Determination of the gold standard for disease targets*

Targets of the 56 diseases that have at least one DE genes were extracted from the disease-target dataset derived from TTD. Total 520 targets were selected and used as the gold standard for the

evaluation. These targets are primary targets, which are directly responsible for the efficacies of the corresponding drugs that were confirmed by strong experimental evidence (11).

2.4 *Evaluation*

To evaluate how often the DE genes and disease-associated SNPs can recover and predict the known targets in each disease, we calculated the percentage of targets that have each feature (recall) and the percentage of each feature that are associated with targets (precision). We also calculated the percentage of targets that have both features and the percentage of having both features and being targets for each disease. This allowed us to evaluate the combinatory effect of differential expression and genetic variants on recovering and predicting known targets. To obtain the expectation of the performance of these features, we randomly sampled (1,000 times) the same amount of genes and SNPs against the total gene sets in the microarray and the whole dbSNP138 pool, respectively. The precision and recall of the random samples were then calculated the same way as above. The q value was calculated as the percentage of the precision or recall of random sampling that is better than the original. The known targets of each disease were used as the gold standards. For comparing the performance between features, the precision and recall of each feature for all diseases were plotted (Figure 3).

2.4.1 *Precision*

For each disease, the precision of DE genes, disease-associated SNPs and both are calculated using the following formulas:

$$\textit{Precision}\ (\text{DE genes}) = \frac{\text{Number of DE genes that are targets}}{\text{Number of DE genes}} \quad (1)$$

$$\textit{Precision}\ (\text{SNPs}) = \frac{\text{Number of disease-associated SNPs in targets}}{\text{Number of disease-associated SNPs}} \quad (2)$$

$$\textit{Precision}\ (\text{DE genes \& SNPs}) = \frac{\text{Number of DE genes harboring SNPs that are targets}}{\text{Number of DE genes harboring SNPs}} \quad (3)$$

2.4.2 *Recall*

For each disease, the recall of DE genes, disease-associated SNPs and both are calculated using the following formulas:

$$\textit{Recall}\ (\text{DE genes}) = \frac{\text{Number of targets that are DE genes}}{\text{Total number of targets}} \quad (4)$$

$$\textit{Recall}\ (\text{SNPs}) = \frac{\text{Number of targets that harbor SNPs}}{\text{Total number of targets}} \quad (5)$$

$$\textit{Recall}\ (\text{DE genes \& SNPs}) = \frac{\text{Number of targets that are DE genes \& harbor SNPs}}{\text{Total number of targets}} \quad (6)$$

3. Results

3.1 *Statistics of the diseases studied*

Overall we studied 56 diseases (Figure 2). According to Human Disease Ontology, they consisted of 16 cancers, 7 nervous system diseases, 6 metabolic diseases, 5 gastrointestinal system diseases,

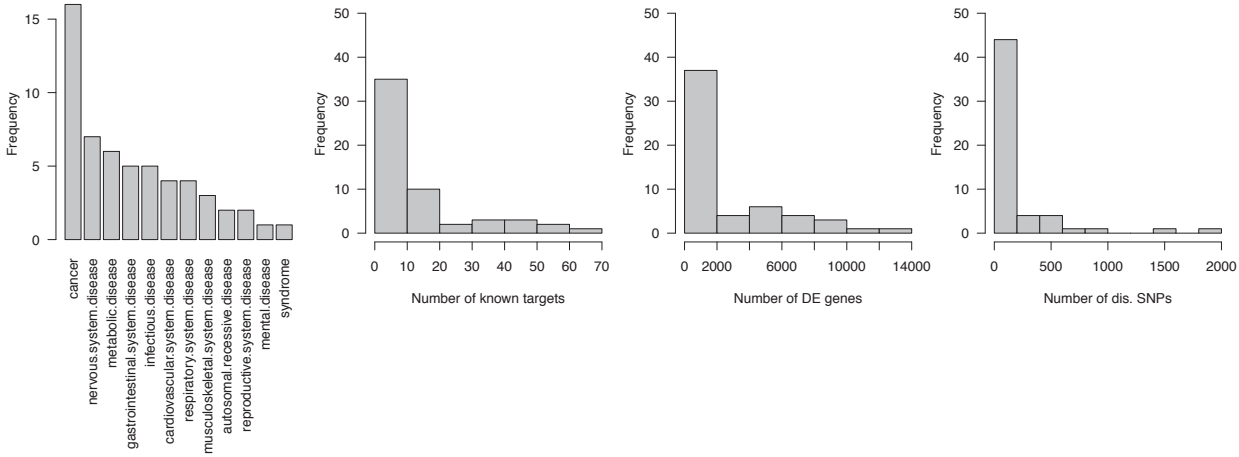

Figure 2. Histogram of disease categories, known targets, DE genes, and disease SNPs of the 56 diseases studied.

5 infectious diseases, 4 cardiovascular system diseases, 4 respiratory system diseases, 3 musculoskeletal system diseases, 3 autosomal recessive diseases, 2 reproductive system diseases, 1 mental disease, and 1 syndrome (Figure 2). These diseases had total 520 unique known targets, 17,409 unique DE genes and 8,235 unique disease-associated SNPs. About 2/3 of them had fewer than 10 targets; 2/3 of them had under 2,000 DE genes; and about 80% of them had fewer than 200 associated SNPs (Figure 2). On average, these diseases had 13.6 known targets, with obesity (64), prostate cancer (59) and breast cancer (51) having the largest number of known targets (Table 1). With a FDR<0.05, the average DE genes these diseases had was 2320. Spinal muscular atrophy and breast cancer had the largest number of DE genes, which were 12,648 and 10,314 respectively. Given the 10^{-6} p-value cutoff, the average number of disease-associated SNPs was 163. Rheumatoid arthritis (RA) and type 1 diabetes mellitus had the largest number of disease-associated SNPs, which were 1,826 and 1,456 respectively. However, 18 out of the 56 diseases did not have any associated SNPs with $p<10^{-6}$ (# of dis. SNPs=0, Table 1).

3.2 *Recovering known targets by DE genes and disease-associated SNPs*

Next, we asked for each disease how often the known targets were differentially expressed, harbored disease-associated SNPs or had both gene features. Those are essentially the true positive rates (recall) of using DE genes, disease-associated SNPs or both to recover known targets. We also calculated how often DE genes, disease-associated SNPs or both were associated with known targets, which are the positive predictive values (precision) of using these gene features to predict targets (Table 1).

Table 1. Statistics of the 56 diseases in the study

				Recall			Precision		
Disease Name	**# of known targets**	**# of DE genes** (FDR<0.05)	**# of dis. SNPs** ($p<10^{-6}$)	**% of targets being DE genes**	**% of targets harboring dis. SNPs**	**% of targets being DE genes & harboring dis. SNPs**	**% of DE genes being targets**	**% of dis SNPs in targets**	**% of DE genes harboring SNPs that are targets**
Obesity	64	1	507	0	3.1*	0	0	0.4*	NA
Prostate Cancer	59	1030	407	23.7**	1.7	0	1.4**	0.5*	0
Breast Cancer	51	10314	189	82.4**	2.0	2.0	0.4*	0.5	3.6**
Asthma	48	2754	348	12.5	4.2	0	0.2	1.7**	0
Rheumatoid Arthritis	43	858	1826	11.6	7.0*	4.7**	0.6	0.5**	15.4**
Type 2 Diabetes Mellitus	41	26	647	0	9.8**	0	0	1.4**	0
Alzheimer's Disease	39	4682	416	30.8	7.7**	0	0.3	2.2**	0
Atherosclerosis	35	93	0	0	0	0	0	NA	NA
Hypertension	32	71	161	0	3.1*	0	0	0.6*	NA
Parkinson's Disease	25	1235	899	8.0	0	0	0.2	0	0
Multiple Sclerosis	22	131	435	9.1*	4.6	4.6**	1.5*	0.2	25.0**
Inflammatory Bowel Disease	18	4682	39	50.0*	0	0	0.2	0	0
Non-small Cell Lung Cancer	17	7834	4	52.9	0	0	0.1	0	0
Hypercholesteremia	17	7179	1	64.7	0	0	0.2	0	NA
Malignant Melanoma	17	7901	78	82.4	5.9*	5.9*	0.2	3.9**	6.2**
Myocardial Infarction	17	4	132	0	0	0	0	0	NA
Osteoarthritis	15	131	59	6.7	0	0	0.8	0	0
Lymphoma	12	888	14	33.3	0	0	0.5	0	NA
Crohn's disease	11	5590	352	54.5	9.1*	0	0.1	0.3*	0
Glaucoma	11	140	28	0	0	0	0	0	NA
Chronic Obstructive Pulmonary Disease	11	23	42	0	0	0	0	0	NA
Acute Myeloid Leukemia	10	2097	0	40.0	0	0	0.2	NA	NA
Malaria	10	194	27	0	0	0	0	0	NA
Erectile Dysfunction	10	1	3	0	0	0	0	0	NA
Sepsis	10	29	0	0	0	0	0	NA	NA
Colon Cancer	9	2547	235	11.1	0	0	0	0	0
Irritable Bowel Syndrome	8	69	0	0	0	0	0	NA	NA
Ulcerative Colitis	7	2587	119	0	0	0	0	0	0
Cystic Fibrosis	7	4	0	0	0	0	0	NA	NA
Type 1 Diabetes Mellitus	7	35	1456	0	0	0	0	0	NA
Small Cell Carcinoma of Lung	7	8118	0	28.6	0	0	0	NA	NA
Bacterial Infection	6	233	0	0	0	0	0	NA	NA
HIV	6	356	350	16.7	16.7*	0	0.3	0.3*	NA
Chronic Lymphocytic Leukemia	6	5996	40	66.7	0	0	0.1	0	0
Amyotrophic Lateral Sclerosis	5	2	84	0	0	0	0	0	NA
Skin Squamous Cell Carcinoma	5	877	2	0	0	0	0	0	NA
Cancer of the Stomach	5	1846	55	0	0	0	0	0	0
Gastro-esophageal Reflux Disease	4	2	0	0	0	0	0	NA	NA
Huntington's Disease	4	8100	15	75.0	0	0	0	0	0
Pulmonary Hypertension	4	10	0	0	0	0	0	NA	NA
Endometriosis	3	9	33	0	0	0	0	0	NA
Acute Promyelocytic Leukaemia	3	2	0	0	0	0	0	NA	NA
Macular Degeneration	3	721	0	0	0	0	0	NA	NA
Pulmonary Fibrosis	3	14	1	0	0	0	0	0	NA
Cervial Cancer	3	66	0	0	0	0	0	NA	NA
Myelodysplastic Syndrome	2	760	0	0	0	0	0	NA	NA
Alpha-1 Anti-trypsin Deficiency	2	5	0	0	0	0	0	NA	NA
Sickle Cell Anemia	1	6429	7	100.0	0	0	0	0	0
Urothelial Carcinoma	1	8767	0	100.0	0	0	0	NA	NA
Cardiomyopathy, Dilated	1	5728	19	100.0	0	0	0	0	0
Hepatic Cirrhosis	1	118	6	0	0	0	0	0	NA
Spinal Muscular Atrophy	1	12648	0	100.0	0	0	0	NA	NA
Vitamin A Deficiency	1	88	0	0	0	0	0	NA	NA
Idiopathic Fibrosing Alveolitis	1	241	18	0	0	0	0	0	NA
Testis Cancer	1	5280	61	0	0	0	0	0	0
Severe Acute Respiratory Syndrome	1	359	0	0	0	0	0	NA	NA
Average	13.6	2319.7	162.8	20.7	1.3	0.3	0.1	0.3	2.4

NA: either the total # of dis. SNPs is 0 or the total # of DE genes harboring SNPs is 0. *: $q < 0.1$, **: $q < 0.05$

The average recall by DE genes was 20.7%. Thirty-two of the 56 diseases (57.1%) had no targets that were DE genes, or 0% recall. Four diseases (urothelial carcinoma, spinal muscular atrophy, sickle cell anemia and dilated cardiomyopathy) had 100% recall, because they had only one known target and that target was a DE gene. Compared to random sampling, DE genes did not perform better in most of the diseases, except for prostate cancer, breast cancer, multiple sclerosis, and inflammatory bowel disease ($q < 0.1$ or $q < 0.05$). The average recall by disease-associated SNPs was 1.3% and 44 of them (78.6%) were 0%, where no targets of those diseases harbored disease-associated SNPs. HIV and type 2 diabetes mellitus had the highest recalls, 16.7% and 9.8% respectively. For diseases with non-zero recall, SNPs of most of them performed better than

random sampling, except prostate cancer, breast cancer, asthma, and multiple sclerosis. The average recall by both DE genes and disease-associated SNPs was 0.3% and 52 of them (92.9%) were 0%. Only 4 diseases (malignant melanoma, rheumatoid arthritis, multiple sclerosis and breast cancer) had targets that were DE genes and harbored disease-associated SNPs, which had a recall of 5.9%, 4.7%, 4.6% and 2%, respectively. Compared to random sampling, the combination performed better in all four diseases, except breast cancer.

On the other hand, the average precision by DE genes was 0.1%, where 39 diseases (69.6%) had no DE genes that were targets, or 0% precision. Multiple sclerosis and prostate cancer had the best precision, 1.5% and 1.4% respectively. Similarly, DE genes did not predict better than random sampling in most of the diseases. Since 18 of the 56 diseases had no associated SNPs (NA in the second to the last column, Table 1), the precision by disease-associated SNPs was calculated for 38 diseases. The average precision by disease-associated SNPs was 0.3% and 26 of them (68.4%) were 0%, where no SNPs of those diseases occurred in targets. Malignant melanoma and Alzheimer's disease had the best precision, 3.9% and 2.2% respectively. For most of the diseases, SNPs also predicted better than random sampling. Thirty-five diseases had no DE genes that also harbored disease-associated SNPs (NA in the last column, Table 1), thus the precision by DE genes and SNPs were calculated for 21 diseases. The average precision by both DE genes and disease-associated SNPs is 2.4% and 17 of them (80.9%) were 0%, where no DE genes that harbored SNPs were targets. The four diseases that had DE genes that harbored disease-associated SNPs and were targets were multiple sclerosis, rheumatoid arthritis, malignant melanoma, and breast cancer, with a precision of 25%, 15.4%, 6.2% and 3.6% respectively. The combined features of all four diseases predicted better than random sampling ($q < 0.05$).

To compare the performance between features, we plotted the precision and recall of each feature for all diseases (Figure 3). Although it was not the common precision and recall curve for

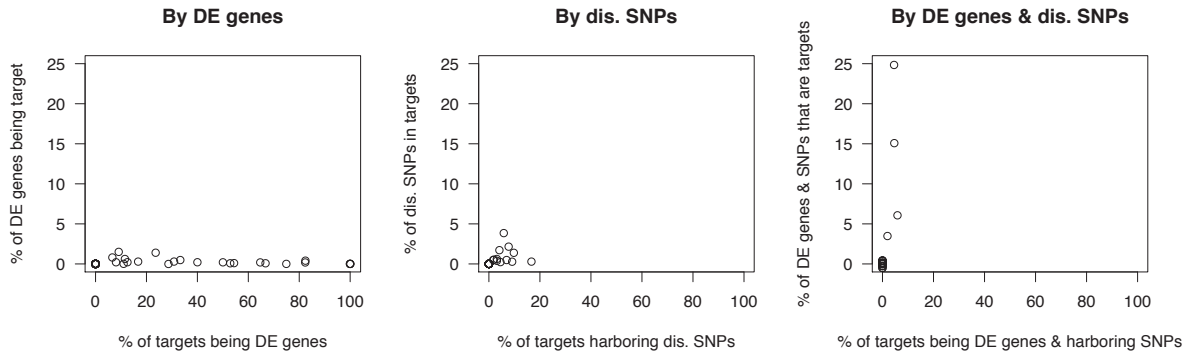

Figure 3. Performance of recovering known targets of each disease by DE genes, disease SNPs or both. Each point indicates the values of one disease.

evaluating the performance of a classifier, it showed how the performance of each feature varied in different diseases and allowed the comparison of performance between features. The performance of each feature varied greatly between diseases. When using DE genes, the recall values ranged from 0-100%, while the precision values were all below 3%. In other words, DE

genes could recover all known targets of a disease, but most of the DE genes were not disease targets. When using disease-associated SNPs, the recall values were between 0-20%, while the precision values were below 5%. In other words, disease-associated SNPs could recover a small portion of the known targets, but most of them were not in known targets. When using both DE genes and disease-associated SNPs, the recall values were below 6%, yet the precision values ranged from 0-25%. In other words, although the combination of both features could hardly recover any known targets, if a gene was differentially expressed and contained disease-associated SNPs, it would have higher chance to be a target for that disease.

When comparing between features, we found that DE genes gave better recall than disease-associated SNPs and disease-associated SNPs gave better recall than the combination of both. On the other hand, disease-associated SNPs gave better precision than DE genes, and the combination of both gave the best precision. We also identified the genes that were differentially expressed, harbored disease-associated SNPs and were targets. They were MOG (Myelin-oligodendrocyte glycoprotein) of multiple sclerosis, C5 (Complement C5) and TNF (Tumor necrosis factor) of rheumatoid arthritis, MC1R (Melanocyte-stimulating hormone receptor) of malignant melanoma, and ERBB4 (Receptor tyrosine-protein kinase erbB-4) of breast cancer (Table 2).

Table 2. Diseases that have targets that are DE genes and harbor dis. SNPs

Disease Name	**All Known Targets** (DE genes in *italic*, harboring SNPs in **bold**, DE genes & harboring SNPs in ***bold italic***)
Multiple Sclerosis	ADRB2, CASP3, CNP, CRH, LPAR1, CXCR3, ICAM1, IFNAR2, IFNG, ITGA4, KCNA3, LEP, MMP9, ***MOG***, MPO, PPARG, KLK6, CFLAR, NR1I2, CCR2, *SPP1*
Rheumatoid Arthritis	***C5***, *CD4*, CD80, CCL2, CCR2, CD86, CFLAR, CTSK, F2RL1, FGF2, IKBKB, IKBKE, IL12A, IL13, IL15, IL17A, IL1R1, IL4, *IL6ST*, ITGA4, ITGB1, **LTA**, ITGB7, JAK3, JUN, LIF, LTB4R, MAPK11, MAPK12, MAPK14, MIF, MMP8, MMP9, *MYD88*, OSM, PTGES, PTGS2, SYK, TLR9, ***TNF***, TNFRSF1B, TRBV7-9, VEGFB
Malignant Melanoma	*ALOX12, BIRC5, BRAF, CDH2,* CTLA4, CTSL1, *DCT, EDNRB, FN1,* HDAC4, *HSP90AA1, IFNAR2, JUN, MAP3K4*, ***MC1R,*** *PLAU, TXNIP*
Breast Cancer	AKT1, ANGPT2, CDH2, CYP1B1, *BRCA2, CCND1, CDC25A, CLU, COPS5, CTSD, CXCL12, CXCR4, CYP19A1, DNMT3B, EGFR, EPHA2, ERBB2,* ***ERBB4,*** *ESR2,* ESRRA, *FOS, HSD17B1, JUN, LHCGR,* MAP2K1, *MAP3K4, MDM2, MFGE8, MMP2, MUC1, NCOA3, NRG1, PGR, PLAUR,* PRL, *PRLR, PTGS2,* PTK6, *PTN, SERPINB5, SNCG, SRC, ST14,* STC1, *TPBG, VDR, HSP90AA1, MAP2K5, SCGB2A2, STS, TYMP*

4. Discussion

Gene expression and genetic variants are the two most commonly measured and used features for selecting the best candidate genes for target validation. Their efficiency in target prioritization is often studied in specific disease conditions and their performance between diseases is largely unknown. Here we incorporated three diverse datasets from GEO microarray database, VARIMED disease-associated SNPs database and TTD target database, and systematically evaluated each feature and the combination of them in recovering and predicting known targets of 56 human diseases.

We found that the performance of each feature varied between diseases, which indicates that each feature could have different therapeutic utility for different diseases. However, overall, both DE genes and SNPs had lower precision than recall, which suggests that the DE or disease-associated SNP feature by itself is not good at predicting a target. The combination of being DE genes and harboring disease-associated SNPs had significantly improved precision ($q < 0.05$) compared to each feature alone (Figure 3). This implies that genes that are differentially expressed and harbor disease-associated SNPs are more likely to be targets. Indeed, for example, TNF (Table 2) is a successful target for RA validated by others (17) and carries risk variants via genome-wide association studies (18). Thus this combinatory feature could be used as a new criterion for prioritizing candidate genes for target validation.

In this study, DE genes, disease-associated SNPs or the combination of them was directly evaluated to allow objective assessment of their performance in target prioritization. Although the combination of DE and SNPs showed increased predictive power, it was still not great (< 25%). Optimizing the two features may improve their performance in prioritizing targets. A common alternative way to prioritize DE genes is their fold change (fc). Disease-associated SNPs can be ranked by how often they are associated with DE genes (%SNPs), since genetic variants associated with disease traits are likely to influence gene expression (1). Then the rank sum of fc and %SNPs can be used combinatorially. Many other prioritization methods can be incorporated with each feature, including the use of protein-protein interaction network, pathway involvement, literature and ontology. However, their effect on the performance may not necessarily improve the overall performance and need to be evaluated on a disease-by-disease basis.

There are limitations in this study that should be recognized. First, the microarrays used to derive the DE genes were from the study with the largest sample size, which could be the reason for the over 10,000 DE genes in some diseases. Meta-analysis of all microarray studies of each disease might result in more robust set of DE genes and a better disease signature (19). Likewise, meta-analysis of genome-wide studies for the same disease, as well as accounting for LD structure among the associated variants, may increase the reliability of disease-SNPs pairs. In this work, we used stringent thresholds (i.e., FDR < 0.05 and p value < 10^{-6}), changing which can alter the number of DE genes and disease-associated SNPs that will affect the precision and recall. Second, the known targets of each disease were extracted from the TTD database. Other databases may help derive more known targets, such as the DrugBank (20) and PharmGKB (21). However, DrugBank does not provide direct relations between targets and diseases, while PharmGKB has more pharmacogenomic information than drug-therapeutic targets relations. It is also important to recognize that all of these databases capture the current knowledge, which is not complete or perfect. As we discover more therapeutic targets and evaluate their efficacy, these resources will become more comprehensive and serve as a better gold standard.

Our study revealed a baseline performance of the two most common gene features, DE and SNPs, on prioritizing candidate targets, and identified an increased predictive power of the combination of the two features than that of each feature alone.

5. Acknowledgments

We thank Dr. Hyojung Paik for the constructive discussion. The research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM079719. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The full data from these analyses is available on request from the authors (fminogue@stanford.edu).

References

1. Plenge, R. M., Scolnick, E. M., and Altshuler, D. (2013) Validating therapeutic targets through human genetics. *Nature reviews. Drug discovery* **12**, 581-594
2. Bromberg, Y. (2013) Chapter 15: disease gene prioritization. *PLoS computational biology* **9**, e1002902
3. Moreau, Y., and Tranchevent, L. C. (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature reviews. Genetics* **13**, 523-536
4. Murray, D., Doran, P., MacMathuna, P., and Moss, A. C. (2007) In silico gene expression analysis--an overview. *Molecular cancer* **6**, 50
5. Hudson, N. J., Dalrymple, B. P., and Reverter, A. (2012) Beyond differential expression: the quest for causal mutations and effector molecules. *BMC genomics* **13**, 356
6. Thomas, G., Jacobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., Crenshaw, A., Cancel-Tassin, G., Staats, B. J., Wang, Z., Gonzalez-Bosquet, J., Fang, J., Deng, X., Berndt, S. I., Calle, E. E., Feigelson, H. S., Thun, M. J., Rodriguez, C., Albanes, D., Virtamo, J., Weinstein, S., Schumacher, F. R., Giovannucci, E., Willett, W. C., Cussenot, O., Valeri, A., Andriole, G. L., Crawford, E. D., Tucker, M., Gerhard, D. S., Fraumeni, J. F., Jr., Hoover, R., Hayes, R. B., Hunter, D. J., and Chanock, S. J. (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nature genetics* **40**, 310-315
7. Chen, R., Morgan, A. A., Dudley, J., Deshpande, T., Li, L., Kodama, K., Chiang, A. P., and Butte, A. J. (2008) FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome biology* **9**, R170
8. Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., Kochi, Y., Ohmura, K., Suzuki, A., Yoshida, S., Graham, R. R., Manoharan, A., Ortmann, W., Bhangale, T., Denny, J. C., Carroll, R. J., Eyler, A. E., Greenberg, J. D., Kremer, J. M., Pappas, D. A., Jiang, L., Yin, J., Ye, L., Su, D. F., Yang, J., Xie, G., Keystone, E., Westra, H. J., Esko, T., Metspalu, A., Zhou, X., Gupta, N., Mirel, D., Stahl, E. A., Diogo, D., Cui, J., Liao, K., Guo, M. H., Myouzen, K., Kawaguchi, T., Coenen, M. J., van Riel, P. L., van de Laar, M. A., Guchelaar, H. J., Huizinga, T. W., Dieude, P., Mariette, X., Bridges, S. L., Jr., Zhernakova, A., Toes, R. E., Tak, P. P., Miceli-Richard, C., Bang, S. Y., Lee, H. S., Martin, J., Gonzalez-Gay, M. A., Rodriguez-Rodriguez, L., Rantapaa-Dahlqvist, S., Arlestig, L., Choi, H. K., Kamatani, Y., Galan, P., Lathrop, M., consortium, R., consortium, G., Eyre, S., Bowes, J., Barton, A., de Vries, N., Moreland, L. W., Criswell, L. A., Karlson, E. W., Taniguchi, A., Yamada, R., Kubo, M., Liu, J. S., Bae, S. C., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Raychaudhuri, S., Stranger, B. E., De Jager, P. L., Franke, L., Visscher, P. M., Brown, M. A., Yamanaka, H., Mimori, T., Takahashi, A., Xu, H., Behrens, T. W., Siminovitch, K. A., Momohara, S., Matsuda, F., Yamamoto, K., and Plenge, R. M. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-381
9. Butte, A. J., and Chen, R. (2006) Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 106-110
10. Dudley, J., and Butte, A. J. (2008) Enabling integrative genomic analysis of high-impact human diseases through text mining. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 580-591
11. Zhu, F., Shi, Z., Qin, C., Tao, L., Liu, X., Xu, F., Zhang, L., Song, Y., Liu, X., Zhang, J., Han, B., Zhang, P., and Chen, Y. (2012) Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery. *Nucleic acids research* **40**, D1128-1136

12. Phan, J. H., Young, A. N., and Wang, M. D. (2012) Robust microarray meta-analysis identifies differentially expressed genes for clinical prediction. *TheScientificWorldJournal* **2012**, 989637
13. Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116-5121
14. Dudley, J. T., Tibshirani, R., Deshpande, T., and Butte, A. J. (2009) Disease signatures are robust across tissues and experiments. *Molecular systems biology* **5**, 307
15. Chen, R., Davydov, E. V., Sirota, M., and Butte, A. J. (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PloS one* **5**, e13574
16. Ashley, E. A., Butte, A. J., Wheeler, M. T., Chen, R., Klein, T. E., Dewey, F. E., Dudley, J. T., Ormond, K. E., Pavlovic, A., Morgan, A. A., Pushkarev, D., Neff, N. F., Hudgins, L., Gong, L., Hodges, L. M., Berlin, D. S., Thorn, C. F., Sangkuhl, K., Hebert, J. M., Woon, M., Sagreiya, H., Whaley, R., Knowles, J. W., Chou, M. F., Thakuria, J. V., Rosenbaum, A. M., Zaranek, A. W., Church, G. M., Greely, H. T., Quake, S. R., and Altman, R. B. (2010) Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525-1535
17. Criscione, L. G., and St Clair, E. W. (2002) Tumor necrosis factor-alpha antagonists for the treatment of rheumatic diseases. *Current opinion in rheumatology* **14**, 204-211
18. International, M. H. C., Autoimmunity Genetics, N., Rioux, J. D., Goyette, P., Vyse, T. J., Hammarstrom, L., Fernando, M. M., Green, T., De Jager, P. L., Foisy, S., Wang, J., de Bakker, P. I., Leslie, S., McVean, G., Padyukov, L., Alfredsson, L., Annese, V., Hafler, D. A., Pan-Hammarstrom, Q., Matell, R., Sawcer, S. J., Compston, A. D., Cree, B. A., Mirel, D. B., Daly, M. J., Behrens, T. W., Klareskog, L., Gregersen, P. K., Oksenberg, J. R., and Hauser, S. L. (2009) Mapping of multiple susceptibility variants within the MHC region for 7 immune-mediated diseases. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 18680-18685
19. Chen, R., Khatri, P., Mazur, P. K., Polin, M., Zheng, Y., Vaka, D., Hoang, C. D., Shrager, J., Xu, Y., Vicent, S., Butte, A. J., and Sweet-Cordero, E. A. (2014) A meta-analysis of lung cancer gene expression identifies PTK7 as a survival gene in lung adenocarcinoma. *Cancer research* **74**, 2892-2902
20. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A. C., and Wishart, D. S. (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **39**, D1035-1041
21. Hernandez-Boussard, T., Whirl-Carrillo, M., Hebert, J. M., Gong, L., Owen, R., Gong, M., Gor, W., Liu, F., Truong, C., Whaley, R., Woon, M., Zhou, T., Altman, R. B., and Klein, T. E. (2008) The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic acids research* **36**, D913-918