PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Systematic interrogation of diverse Omic data reveals interpretable, robust, and generalizable transcriptomic features of clinically successful therapeutic targets

Andrew D. Rouillard, Mark R. Hurle, Pankaj Agarwal*

Computational Biology, GSK, Collegeville, PA, United States of America

* pankaj.agarwal@gsk.com

Check for updates

Abstract

Target selection is the first and pivotal step in drug discovery. An incorrect choice may not manifest itself for many years after hundreds of millions of research dollars have been spent. We collected a set of 332 targets that succeeded or failed in phase III clinical trials, and explored whether Omic features describing the target genes could predict clinical success. We obtained features from the recently published comprehensive resource: Harmonizome. Nineteen features appeared to be significantly correlated with phase III clinical trial outcomes, but only 4 passed validation schemes that used bootstrapping or modified permutation tests to assess feature robustness and generalizability while accounting for target class selection bias. We also used classifiers to perform multivariate feature selection and found that classifiers with a single feature performed as well in cross-validation as classifiers with more features (AUROC = 0.57 and AUPR = 0.81). The two predominantly selected features were mean mRNA expression across tissues and standard deviation of expression across tissues, where successful targets tended to have lower mean expression and higher expression variance than failed targets. This finding supports the conventional wisdom that it is favorable for a target to be present in the tissue(s) affected by a disease and absent from other tissues. Overall, our results suggest that it is feasible to construct a model integrating interpretable target features to inform target selection. We anticipate deeper insights and better models in the future, as researchers can reuse the data we have provided to improve methods for handling sample biases and learn more informative features. Code, documentation, and data for this study have been deposited on GitHub at https://github.com/arouillard/omic-features-successful-targets.

Author summary

Drug discovery often begins with a hypothesis that changing the abundance or activity of a target—a biological molecule, usually a protein—will cure a disease or ameliorate its symptoms. Whether a target hypothesis translates into a successful therapy depends in part on the

OPEN ACCESS

Citation: Rouillard AD, Hurle MR, Agarwal P (2018) Systematic interrogation of diverse Omic data reveals interpretable, robust, and generalizable transcriptomic features of clinically successful therapeutic targets. PLoS Comput Biol 14(5): e1006142. https://doi.org/10.1371/journal.pcbi.1006142

Editor: Edwin Wang, University of Calgary Cumming School of Medicine, CANADA

Received: November 21, 2017

Accepted: April 13, 2018

Published: May 21, 2018

Data Availability Statement: All relevant data are within the paper, in supporting information files, or on GitHub at https://github.com/arouillard/omic-features-successful-targets.

Funding: The author(s) received no specific funding for this work.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: The authors are employed by GlaxoSmithKline. This does not alter our adherence to all PLOS Computational Biology policies on sharing data and materials.

characteristics of the target, but it is not completely understood which target characteristics are important for success. We sought to answer this question with a supervised machine learning approach. We obtained outcomes of target hypotheses tested in clinical trials, scoring targets as successful or failed, and then obtained thousands of features (i.e. properties or characteristics) of targets from dozens of biological datasets. We statistically tested which features differed between successful and failed targets, and built a computational model that used these features to predict success or failure of targets in clinical trials. We found that successful targets tended to have more variable mRNA abundance from tissue to tissue and lower average abundance across tissues than failed targets. Thus, it is probably favorable for a target to be present in the tissue(s) affected by a disease and absent from other tissues. Our work demonstrates the feasibility of predicting clinical trial outcomes from target features.

Introduction

More than half of drug candidates that advance beyond phase I clinical trials fail due to lack of efficacy [1, 2]. One possible explanation for these failures is sub-optimal target selection [3]. Many factors must be considered when selecting a target for drug discovery [4, 5]. Intrinsic factors include the likelihood of the target to be tractable (can the target's activity be altered by a compound, antibody, or other drug modality?), safe (will altering the target's activity cause serious adverse events?), and efficacious (will altering the target's activity provide significant benefit to patients?). Extrinsic factors include the availability of investigational reagents and disease models for preclinical target validation, whether biomarkers are known for measuring target engagement or therapeutic effect, the duration and complexity of clinical trials required to prove safety and efficacy, and the unmet need of patients with diseases that might be treated by modulating the target.

Over the past decade, technologies have matured enabling high-throughput genome-, transcriptome-, and proteome-wide profiling of cells and tissues in normal, disease, and experimentally perturbed states. In parallel, researchers have made substantial progress curating or text-mining biomedical literature to extract and organize information about genes and proteins, such as molecular functions and signaling pathways, into structured datasets. Taken together, both efforts have given rise to a vast amount of primary, curated, and text-mined data about genes and proteins, which are stored in online repositories and amenable to computational analysis [6, 7].

To improve the success rate of drug discovery projects, researchers have investigated whether any features of genes or proteins are useful for target selection. These computational studies can be categorized according to whether the researchers were trying to predict tractability [8, 9], safety [10–13], efficacy (no publications to our knowledge), or overall success (alternatively termed "drug target likeness") [8, 13–26]. Closely related efforts include disease gene prediction, where the goal is to predict genes mechanistically involved in a given disease [27–32], and disease target prediction, where the goal is to predict genes that would make successful drug targets for a given disease [33–35].

To our knowledge, we report the first screen for features of genes or proteins that distinguish targets of approved drugs from targets of drug candidates that failed in clinical trials. In contrast, related prior studies have searched for features that distinguish targets of approved drugs from the rest of the genome (or a representative subset) [13, 15–25]. Using the remainder of the genome for comparison has been useful for finding features enriched among successful targets, but it is uncertain whether these features are specific to successful targets or are enriched among targets of failed drug candidates as well. Our study aims to fill this knowledge gap by directly testing for features that separate targets by clinical outcome, expanding the scope of prior studies that have investigated how genetic disease associations [36] and publication trends [37] of targets correlate with clinical outcome.

Our work has five additional innovative characteristics. First, we included only targets of drugs that are presumed to be selective (no documented polypharmacology) to reduce ambiguity in assigning clinical trial outcomes to targets. Second, we included only phase III failures to enrich for target efficacy failures, as opposed to safety and target engagement failures, which are more common in phase I and phase II [2]. Third, we excluded targets of assets only indicated for cancer, as studies have observed that features of successful targets for cancer differ from features of successful targets for other indications [22, 23], moreover, cancer trials fail more frequently than trials for other indications [2]. Fourth, we interrogated a diverse and comprehensive set of features, over 150,000 features from 67 datasets covering 16 feature types, whereas prior studies have examined only features derived from protein sequence [16–18, 24, 25], protein-protein interactions [13, 15, 18–23], Gene Ontology terms [13, 15, 16], and gene expression profiles [15, 19, 21, 25]. Fifth, because targets of drugs and drug candidates do not constitute a random sample of the genome, we implemented a suite of tests to assess the robustness and generalizability of features identified as significantly separating successes from failures in the biased sample.

A handful of the initial 150,000+ features passed our tests for robustness and generalizability to new targets or target classes. Interestingly, these features were predominantly derived from gene expression datasets. *Notably, two significant features were discovered repeatedly in multiple datasets: successful targets tended to have lower mean mRNA expression across tissues and higher expression variance than failed targets*. We also trained a classifier to predict phase III success probabilities for untested targets (no phase III clinical trial outcomes reported for drug candidates that selectively modulate these targets). We identified 943 targets with sufficiently unfavorable expression characteristics to be predicted twice as likely to fail in phase III clinical trials as past phase III targets. Furthermore, we identified 2,700,856 target pairs predicted with 99% consistency to have a 2-fold difference in success probability. Such pairwise comparisons may be useful for prioritizing short lists of targets under consideration for a therapeutic program. We conclude this paper with a discussion of the biases and limitations faced when attempting to analyze, model, or interpret data on clinical trial outcomes.

Results

Examples of successful and failed targets obtained from phase III clinical trial reports

We extracted phase III clinical trial outcomes reported in Pharmaprojects [38] for drug candidates reported to be selective (single documented target) and tested as treatments for non-cancer diseases. We grouped the outcomes by target, scored targets with at least one approved drug as successful ($N_S = 259$), and scored targets with no approved drugs and at least one documented phase III failure as failed ($N_F = 72$) (S1 Table). The target success rate (77%) appears to be inflated relative to typically reported phase III success rates (58%) [2] because we scored targets by their best outcome across multiple trials.

Comprehensive and diverse collection of target features obtained from the Harmonizome

We obtained target features from the Harmonizome [39], a recently published collection of features of genes and proteins extracted from over 100 Omics datasets. We limited our analysis to 67 datasets that are in the public domain or GSK had independently licensed (Table 1).

Table 1. Datasets tested for features significantly separating successful targets from failed targets.

Dataset	Feature Type	Total Genes	Covered Samples	Total Features	Covered Features	Reduced Features
Roadmap Epigenomics Cell and Tissue DNA Methylation Profiles	cell or tissue DNA methylation	13835	227	26	26	4
Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles	cell or tissue expression	17979	287	416	416	2
Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles	cell or tissue expression	14248	287	2234	2234	2
BioGPS Human Cell Type and Tissue Gene Expression Profiles	cell or tissue expression	16383	320	86	86	2
BioGPS Mouse Cell Type and Tissue Gene Expression Profiles	cell or tissue expression	15443	313	76	76	2
GTEx Tissue Gene Expression Profiles	cell or tissue expression	26005	328	31	31	2
GTEx Tissue Sample Gene Expression Profiles	cell or tissue expression	19250	301	2920	2920	2
HPA Cell Line Gene Expression Profiles	cell or tissue expression	15868	259	45	45	1
HPA Tissue Gene Expression Profiles	cell or tissue expression	17496	314	33	33	2
HPA Tissue Protein Expression Profiles	cell or tissue expression	15788	266	46	46	11
HPA Tissue Sample Gene Expression Profiles	cell or tissue expression	16742	300	123	123	2
HPM Cell Type and Tissue Protein Expression Profiles	cell or tissue expression	7274	94	6	6	2
ProteomicsDB Cell Type and Tissue Protein Expression Profiles	cell or tissue expression	2776	28	55	55	5
Roadmap Epigenomics Cell and Tissue Gene Expression Profiles	cell or tissue expression	12824	164	59	59	6
TISSUES Curated Tissue Protein Expression Evidence Scores	cell or tissue expression	16216	317	645	245	106
TISSUES Experimental Tissue Protein Expression Evidence Scores	cell or tissue expression	17922	316	245	244	44
TISSUES Text-mining Tissue Protein Expression Evidence Scores	cell or tissue expression	16184	330	4189	2974	2118
ENCODE Histone Modification Site Profiles	cell or tissue histone modification sites	22382	330	437	432	91
Roadmap Epigenomics Histone Modification Site Profiles	cell or tissue histone modification sites	21032	313	385	295	282
ENCODE Transcription Factor Binding Site Profiles	cell or tissue transcription factor binding sites	22845	330	1681	1591	723
JASPAR Predicted Transcription Factor Targets	cell or tissue transcription factor binding sites	21547	330	113	80	77
COMPARTMENTS Curated Protein Localization Evidence Scores	cellular compartment associations	16738	330	1465	228	105
COMPARTMENTS Experimental Protein Localization Evidence Scores	cellular compartment associations	6495	73	61	37	10
COMPARTMENTS Text-mining Protein Localization Evidence Scores	cellular compartment associations	14375	330	2083	877	545
GO Cellular Component Annotations	cellular compartment associations	16757	328	1549	208	124
LOCATE Curated Protein Localization Annotations	cellular compartment associations	9639	269	80	50	20
LOCATE Predicted Protein Localization Annotations	cellular compartment associations	19747	325	26	23	10
CTD Gene-Chemical Interactions	chemical interactions	11125	321	9518	2222	2042
Guide to Pharmacology Chemical Ligands of Receptors	chemical interactions	899	209	4896	189	52
Kinativ Kinase Inhibitor Bioactivity Profiles	chemical interactions	232	9	28	28	25

(*Continued*)

Table 1. (Continued)

Dataset	Feature Type	Total Genes	Covered Samples	Total Features	Covered Features	Reduced Features
KinomeScan Kinase Inhibitor Targets	chemical interactions	287	10	75	75	72
CMAP Signatures of Differentially Expressed Genes for Small Molecules	chemical perturbation differentially expressed genes	12148	300	6102	5066	5065
ClinVar SNP-Phenotype Associations	disease or phenotype associations	2458	143	3293	3	2
CTD Gene-Disease Associations	disease or phenotype associations	21582	331	6327	2926	2116
dbGAP Gene-Trait Associations	disease or phenotype associations	5668	147	512	51	49
DISEASES Curated Gene-Disease Assocation Evidence Scores	disease or phenotype associations	2252	115	772	94	49
DISEASES Experimental Gene-Disease Assocation Evidence Scores	disease or phenotype associations	4055	131	352	106	43
DISEASES Text-mining Gene-Disease Assocation Evidence Scores	disease or phenotype associations	15309	330	4630	2559	1850
GAD Gene-Disease Associations	disease or phenotype associations	10705	318	12780	1189	980
GAD High Level Gene-Disease Associations	disease or phenotype associations	8016	314	20	19	16
GWAS Catalog Gene-Disease Associations	disease or phenotype associations	4356	127	1009	30	28
GWASdb SNP-Disease Associations	disease or phenotype associations	11805	253	587	252	126
GWASdb SNP-Phenotype Associations	disease or phenotype associations	12488	261	824	397	150
HPO Gene-Disease Associations	disease or phenotype associations	3158	171	6844	1187	667
HuGE Navigator Gene-Phenotype Associations	disease or phenotype associations	12055	322	2755	1241	1153
MPO Gene-Phenotype Associations	disease or phenotype associations	7798	299	8581	2434	1444
OMIM Gene-Disease Associations	disease or phenotype associations	4553	209	6177	5	4
GeneSigDB Published Gene Signatures	gene signatures or modules	19723	331	3517	1363	1313
MSigDB Cancer Gene Co-expression Modules	gene signatures or modules	4869	135	358	135	95
MiRTarBase microRNA Targets	microRNA targets	12086	218	598	93	91
TargetScan Predicted Conserved microRNA Targets	microRNA targets	14923	283	1539	1020	791
TargetScan Predicted Nonconserved microRNA Targets	microRNA targets	18210	324	1541	1534	1236
GO Biological Process Annotations	pathway, function, or process associations	15717	328	13214	2436	1215
GO Molecular Function Annotations	pathway, function, or process associations	15777	327	4164	367	204
HumanCyc Pathways	pathway, function, or process associations	932	41	288	11	8
KEGG Pathways	pathway, function, or process associations	7016	298	303	185	179
PANTHER Pathways	pathway, function, or process associations	1962	138	147	40	39
Reactome Pathways	pathway, function, or process associations	9005	309	1814	289	159
Wikipathways Pathways	pathway, function, or process associations	4958	263	301	140	137
DEPOD Substrates of Phosphatases	phosphatase interactions	293	19	114	13	9
NURSA Protein Complexes	protein complex associations	9785	141	1798	1182	1181
InterPro Predicted Protein Domain Annotations	protein domain associations	18002	329	11017	119	63
BioGRID Protein-Protein Interactions	protein interactions	15270	306	15272	1191	1163
DIP Protein-Protein Interactions	protein interactions	2709	140	2711	32	24
Guide to Pharmacology Protein Ligands of Receptors	protein interactions	187	46	213	5	4
IntAct Biomolecular Interactions	protein interactions	12303	269	12305	422	417

(Continued)

Table 1. (Continued)

Dataset	Feature Type	Total Genes	Covered Samples	Total Features	Covered Features	Reduced Features
GTEx eQTL	SNP eQTL targets	7898	107	7817	2	1
TOTALS	NA	NA	NA	174228	44092	28562

https://doi.org/10.1371/journal.pcbi.1006142.t001

Each dataset in the Harmonizome is organized into a matrix with genes labeling the rows and features such as diseases, phenotypes, tissues, and pathways labeling the columns. We included the mean and standard deviation calculated along the rows of each dataset as additional target features. These summary statistics provide potentially useful and interpretable information about targets, such as how many pathway associations a target has or how variable a target's expression is across tissues.

The datasets contained a total of 174,228 features covering 16 feature types (Table 1). We restricted our analysis to 44,092 features that had at least three non-zero values for targets assigned a phase III outcome. Many datasets had strong correlations among their features. To reduce feature redundancy and avoid excessive multiple hypothesis testing while maintaining interpretability of features, we replaced each group of highly correlated features with the group mean feature and assigned it a representative label (Fig 1, S2 Table). The number of features shrunk to 28,562 after reducing redundancy.

Target features tested for correlation with phase III outcome

We performed permutation tests [40, 41] on the remaining 28,562 target features to find features with a significant difference between the successful and failed targets, and we corrected p-values for multiple hypothesis testing using the Benjamini-Yekutieli method [42] (Fig 1, S2 Table). We used permutation testing to apply the same significance testing method to all features, since they had heterogeneous data distributions. We detected 19 features correlated with clinical outcome at a within-dataset false discovery rate of 0.05 (Table 2). The significant features were derived from 7 datasets, of which 6 datasets were gene expression atlases: Allen Brain Atlas adult human brain tissues [43, 44], Allen Brain Atlas adult mouse brain tissues [43, 45], BioGPS human cell types and tissues [46–48], BioGPS mouse cell types and tissues [46–48], Genotype-Tissue Expression Project (GTEx) human tissues [49, 50], and Human Protein Atlas (HPA) human tissues [51]. The remaining dataset, TISSUES [52], was an integration of experimental gene and protein tissue expression evidence from multiple sources. Two correlations were significant in multiple datasets: successful targets tended to have lower mean expression across tissues and higher expression variance than failed targets.

Significant features tested for robustness to sample variation and generalization across target classes

Because targets of drugs and drug candidates do not constitute a random sample of the genome, features that separate successful targets from failed targets in our sample may perform poorly as genome-wide predictors of success versus failure. We performed three analyses to address this issue (Fig 1).

Robustness to sample variation. We used bootstrapping [53, 54] (sampling with replacement from the original set of examples to construct sets of examples equal in size to the original set) to investigate how robust our significance findings were to variation in the success and failure examples. For each dataset that yielded significant features in our primary analysis, we repeated the analysis on 1000 bootstrap samples and quantified the replication probability [55]

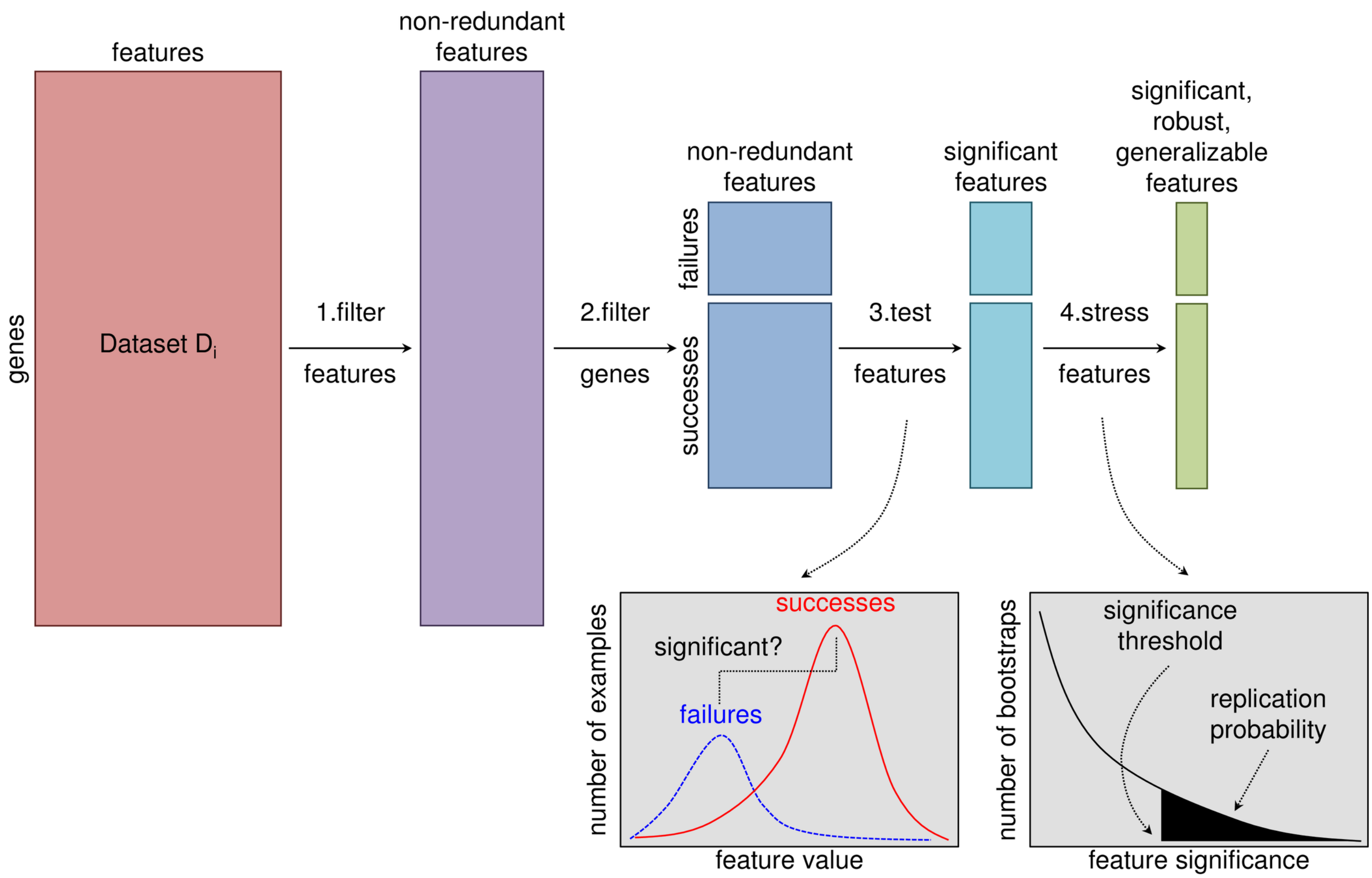

Fig 1. Feature selection pipeline. Each dataset took the form of a matrix with genes labeling the rows and features labeling the columns. We appended the mean and standard deviation computed across all features as two additional features. **Step 1:** We filtered the columns to eliminate redundant features, replacing each group of correlated features with the group average feature, where a group was defined as features with squared pair-wise correlation coefficient $r^2 \geq 0.5$. If the dataset mean feature was included in a group of correlated features, we replaced the group with the dataset mean. **Step 2:** We filtered the rows for targets with clinical trial outcomes of interest: targets of selective drugs approved for non-cancer indications (successes) and targets of selective drug candidates that failed in phase III clinical trials for non-cancer indications (failures). **Step 3:** We tested the significance of each feature as an indicator of success or failure using permutation tests to quantify the significance of the difference between the means of the successful and failed targets. We corrected for multiple hypothesis testing using the Benjamini-Yekutieli method to control the false discovery rate at 0.05 within each dataset. **Step 4:** We "stressed" the significant features with additional tests to assess their robustness and generalizability. For example, we used bootstrapping to estimate probabilities that the significance findings will replicate on similar sets of targets.

https://doi.org/10.1371/journal.pcbi.1006142.g001

of each feature as the fraction of bootstraps yielding a significant correlation with phase III outcome at a within-dataset false discovery rate of 0.05. Twelve features had less than 80% probability (considered a strong replication probability in [55]) that their correlation with clinical outcome will generalize to new examples (Table 2).

Robustness to target class variation. We tested if any of the significance findings depended upon the presence of targets from a single target class in our sample. We obtained target class labels (i.e. gene family labels) from the HUGO Gene Nomenclature Committee [56], tested if any target classes were significantly correlated with phase III outcome, and then tested if these classes were correlated with any features. The GPCR and integrin classes were correlated with phase III outcome as well as several features (Table 2). This raised the possibility that instead of these features being genome-wide indicators of clinical outcome, they were simply reflecting the fact that many GPCRs have succeeded (62/70, $p<0.05$) or that integrins have failed (3/3, $p<0.01$). To test this possibility, we repeated the bootstrapping procedure

Table 2. Features significantly correlated with phase III outcome.

Dataset	Feature	Corr Pval	Correl-ation Sign	Correlated Target Classes (and sign)	Repl Prob (Bootstrap)	Repl Prob (Class Holdout Bootstrap)	Repl Prob (Within Class Permutation Bootstrap)
BioGPS Human Cell Type and Tissue Gene Expression Profiles	[mean]	0.001	-1	GPCRs (-1)	0.89	0.98	0.83
BioGPS Human Cell Type and Tissue Gene Expression Profiles	stdv	0.010	-1	GPCRs (-1), Integrins (+1)	*0.69*	*0.56*	*0.32*
BioGPS Mouse Cell Type and Tissue Gene Expression Profiles	[mean]	0.042	-1	GPCRs (-1)	*0.55*	*0.71*	*0.56*
Allen Brain Atlas Adult Human Brain Tissue Gene Expression Profiles	[mean]	0.006	-1	GPCRs (-1)	*0.78*	0.80	*0.78*
Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles	r3 roof plate	0.002	-1	None	0.88	1.00	0.89
Allen Brain Atlas Adult Mouse Brain Tissue Gene Expression Profiles	[mean]	0.007	-1	None	*0.76*	1.00	*0.79*
GTEx Tissue Gene Expression Profiles	[mean]	0.014	-1	GPCRs (-1)	*0.65*	*0.60*	*0.76*
GTEx Tissue Gene Expression Profiles	stdv	0.014	+1	GPCRs (+1)	*0.69*	0.94	*0.76*
HPA Tissue Gene Expression Profiles	[mean]	0.004	-1	GPCRs (-1)	0.80	0.90	0.85
HPA Tissue Gene Expression Profiles	stdv	0.004	+1	None	0.81	1.00	0.81
TISSUES Experimental Tissue Protein Expression Evidence Scores	bone marrow	0.001	-1	GPCRs (-1)	0.92	0.96	*0.66*
TISSUES Experimental Tissue Protein Expression Evidence Scores	[hematopoietic cells]	0.001	-1	GPCRs (-1), Integrins (+1)	0.93	1.00	*0.72*
TISSUES Experimental Tissue Protein Expression Evidence Scores	[mean]	0.001	-1	GPCRs (-1)	0.85	0.99	*0.76*
TISSUES Experimental Tissue Protein Expression Evidence Scores	[epithalamus and pineal gland]	0.012	-1	None	*0.73*	0.97	*0.49*
TISSUES Experimental Tissue Protein Expression Evidence Scores	erythroid cell	0.015	-1	None	*0.68*	0.94	*0.45*
TISSUES Experimental Tissue Protein Expression Evidence Scores	[t-lymphocyte]	0.017	-1	None	*0.65*	0.95	*0.65*
TISSUES Experimental Tissue Protein Expression Evidence Scores	[miscellaneous tissues]	0.017	-1	GPCRs (-1)	*0.64*	*0.64*	*0.63*
TISSUES Experimental Tissue Protein Expression Evidence Scores	[thymus and thorax]	0.017	-1	Integrins (+1)	*0.60*	*0.37*	*0.44*
TISSUES Experimental Tissue Protein Expression Evidence Scores	adrenal cortex	0.043	-1	None	*0.44*	*0.62*	*0.45*

Footnotes

Abbreviations: Corr Pval = p-value corrected for multiple hypothesis testing, Repl Prob = replication probability.

[Square brackets] denote groups of features.

[miscellaneous tissues] is a heterogeneous group of digestive, respiratory, urogenital, reproductive, nervous, cardiovascular, and hematopoietic system tissues.

White background indicates features that passed all tests for robustness and generalizability.

Gray background indicates features that failed at least one test for robustness or generalizability. Strikethrough italics indicates the failed test(s).

https://doi.org/10.1371/journal.pcbi.1006142.t002

described above to obtain replication probabilities, except excluded GPCRs and integrins from being drawn in the bootstrap samples. Six features had less than 80% probability that their correlation with clinical outcome will generalize to new target classes (Table 2).

Generalization across target classes. In the preceding analysis, we checked one target class at a time for its impact on our significance findings. To broadly test whether features generalize across target classes, we repeated the permutation testing described in our initial analysis, but only shuffled the success/failure labels within target classes, inspired by the work of Epstein et al. [57] on correcting for confounders in permutation testing. By generating a null distribution with preserved ratio of successes to failures within each target class, features must correlate with clinical outcome within multiple classes to be significant, while features that discriminate between classes will not be significant. We repeated the modified permutation tests on 1000 bootstrap samples to obtain replication probabilities. We rejected fifteen features that had less than 80% probability that their correlation with clinical outcome generalizes across target classes (Table 2). This set of fifteen features included all features with less than 80% replication probability in either of the previous two tests. The remaining robust and generalizable features were: 1) mean mRNA expression across tissues (HPA and BioGPS human tissue expression datasets), 2) standard deviation of expression across tissues (HPA human tissue expression dataset), and 3) expression in r3 roof plate (Allen Brain Atlas adult mouse brain tissue expression dataset). The r3 roof plate expression profile was correlated with mean expression across tissues in the Allen Brain Atlas dataset ($r^2 = 0.47$), falling just below the $r^2 = 0.5$ cut-off that would have grouped r3 roof plate with the mean expression profile during dimensionality reduction.

Classifier-based assessment of feature usefulness and interpretability

Statistical significance did not guarantee the remaining features would be useful in practice for discriminating between successes and failures. To test their utility, we trained a classifier to predict target success or failure, using cross-validation to select a model type (Random Forest or logistic regression) and a subset of features useful for prediction. Because we used all targets with phase III outcomes for the feature selection procedure described above, simply using the final set of features to train a classifier on the same data would yield overly optimistic performance, even with cross-validation. Therefore, we implemented a nested cross-validation routine to perform both feature selection and model selection [58].

Cross-validation routine. The outer loop of the cross-validation routine had five steps (Fig 2): 1) separation of targets with phase III outcomes into training and testing sets, 2) univariate feature selection using the training set, 3) aggregation of features from different datasets into a single feature matrix, 4) classifier-based feature selection and model selection using the training set, and 5) evaluation of the classifier on the test set. Step 4 used an inner loop with 5-fold cross-validation repeated 20 times to estimate the performance of different classifier types (Random Forest or logistic regression) and feature subsets (created by incremental feature elimination). The simplest classifier (least number of features, with logistic regression considered simpler than Random Forest) with cross-validation values for area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR) within 95% of maximum was selected. The outer loop used 5-fold cross-validation repeated 200 times, which provided 1000 train-test cycles for estimating the generalization performance of the classifier and characterizing the consistency of the selected features and model type.

Classifier consistency. Simple models were consistently selected for the classifier (Table 3, S3 Table). In 1000 train-test cycles, a logistic regression model with one feature was selected most the time (66%), followed in frequency by a logistic regression model with two features (8%), a Random Forest model with two features (8%), and a logistic regression model with three features (6%). Other combinations of model type (logistic regression or Random Forest) and number of features (ranging from 1 to 8) appeared 11% of the time (each 4% or less). For one of the train-test cycles (0.1%), no significant features were found in the univariate feature selection step, resulting in a null model. Note that the logistic regression models were selected primarily because we imposed a preference for simple and interpretable models, not because they performed better than Random Forest models. The Random Forest model tended to perform as well as the logistic regression model on the inner cross-validation loop, with AUROC = 0.62 ± 0.06 for Random Forest and 0.63 ± 0.05 for logistic regression (S4 Table).

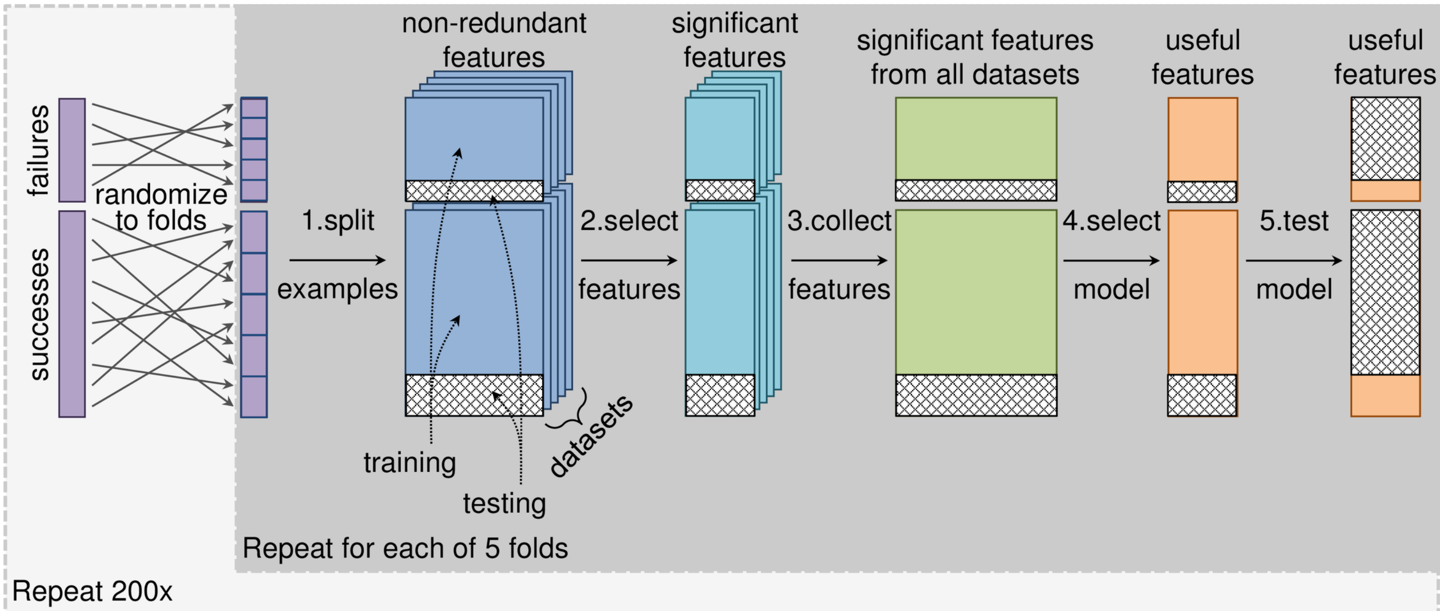

Fig 2. Modeling pipeline. We trained a classifier to predict phase III clinical trial outcomes, using 5-fold cross-validation repeated 200 times to assess the stability of the classifier and estimate its generalization performance. For each fold of cross-validation, modeling began with the non-redundant features for each dataset. **Step 1:** We split the targets with phase III outcomes into training and testing sets. **Step 2:** We performed univariate feature selection using permutation tests to quantify the significance of the difference between the means of the successful and failed targets in the training examples. We controlled for target class as a confounding factor by only shuffling outcomes within target classes. We accepted features with adjusted p-values less than 0.05 after correcting for multiple hypothesis testing using the Benjamini-Yekutieli method. **Step 3:** We aggregated significant features from all datasets into a single feature matrix. **Step 4:** We performed incremental feature elimination with an inner 5-fold cross-validation loop repeated 20 times to select the type of classifier (Random Forest or logistic regression) and smallest subset of features that had cross-validation area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR) values within 95% of maximum. **Step 5:** We refit the selected model using all the training examples and evaluated its performance on the test examples.

https://doi.org/10.1371/journal.pcbi.1006142.g002

Gene expression features were consistently selected for the classifier (Table 4, S3 Table). Mean mRNA expression across tissues and standard deviation of expression across tissues had frequencies of 69% and 59%, respectively. More precisely, 36% of the models used mean mRNA expression across tissues as the only feature, 31% used standard deviation of expression as the only feature, and 12% used mean and standard deviation as the only two features. Other expression features appeared in 21% of the models. These expression features tended to be

Table 3. Distribution of train-test cycles by classifier type and number of selected features.

		Selected Model Type		
		Logistic Regression	Random Forest	Total
Selected Features	1	662	5	667
	2	82	84	166
	3	57	41	98
	4	22	2	24
	5	24	1	25
	6	11	0	11
	7	6	0	6
	8	2	0	2
	Total	866	133	999*

Footnotes

* 1 train-test cycle yielded no significant features for modeling

https://doi.org/10.1371/journal.pcbi.1006142.t003

Table 4. Number of train-test cycles in which feature was selected for the classifier.

Feature Type	Feature	Count
cell or tissue expression	mean across tissues	685
cell or tissue expression	standard deviation across tissues	585
cell or tissue expression	other	214
disease or phenotype associations	mean across diseases	2
disease or phenotype associations	other	2
pathway, function, or process associations	any	1

https://doi.org/10.1371/journal.pcbi.1006142.t004

correlated with mean expression across tissues (median $r^2 = 0.49$). Disease association features appeared in 0.4% of the models.

Classifier performance. The classifier consistently had better than random performance in cross-validation (Fig 3, Table 5, S5 Table). The 2.5th, 50th, and 97.5th percentiles for AUROC were 0.51, 0.57, and 0.61. For comparison, a random ordering of targets would yield an AUROC of 0.50. The receiver operating characteristic curve showed that there was no single cut-off that would provide satisfactory discrimination between successes and failures (Fig 3A). For an alternative view, we used kernel density estimation [59] to fit distributions of the probability of success predicted by the classifier for the successful, failed, and unlabeled targets (Fig 3B, S1 Table). The distributions for successes and failures largely overlapped, except in the tails.

We attempted to identify subsets of targets with high positive predictive value (PPV) or high negative predictive value (NPV). The median PPV rose as high as 0.99, but uncertainty in the PPV was so large that we could not be confident in identifying any subset of targets with a predicted success rate better than the historical 0.77 (Fig 3C). The median NPV rose to 0.40, roughly twice the historical failure rate of 0.23. Furthermore, at 0.40 median NPV, 99% of the cross-validation repetitions had an NPV greater than the historical failure rate (Fig 3D). Using this cut-off, we identified 943 unlabeled targets expected to be twice as likely to fail in phase III clinical trials as past phase III targets.

We reasoned that a more practical use of the classifier would be to make pair-wise comparisons among a short list of targets already under consideration for a therapeutic program. To assess the utility of the classifier for this purpose, for every pair of targets T_A and T_B, we computed the fraction of cross-validation runs in which the classifier predicted greater probability of success for T_B than T_A. We identified 67,270,678 target pairs (39%) with at least a 0.1 difference in median success probability where the classifier was 95% consistent in predicting greater probability of success for T_B than T_A. The classifier was 99% consistent for 41528043 target pairs (24%). Requiring at least a 2-fold difference in median success probability between T_B and T_A reduced these counts to 2,730,437 target pairs (1.6%) at 95% consistency and 2,700,856 target pairs (1.6%) at 99% consistency. We visualized these results by plotting the 95% and 99% consistency fraction thresholds smoothly interpolated as a function of the median predicted probabilities of success of T_A and T_B (Fig 3E). For a median probability of success of T_A around 0.2, T_B must have a median probability of success of 0.5 or greater at the 99% threshold. For lower T_A success probabilities, the T_B success probability must be even higher because there is greater uncertainty about the low T_A probabilities. For higher T_A success probabilities, the T_B success probability at the 99% threshold increases steadily until a T_A success probability of about 0.6, where the T_B success probability reaches 1. For T_A success probabilities above 0.6, no targets are predicted to have greater probability of success with 99% consistency.

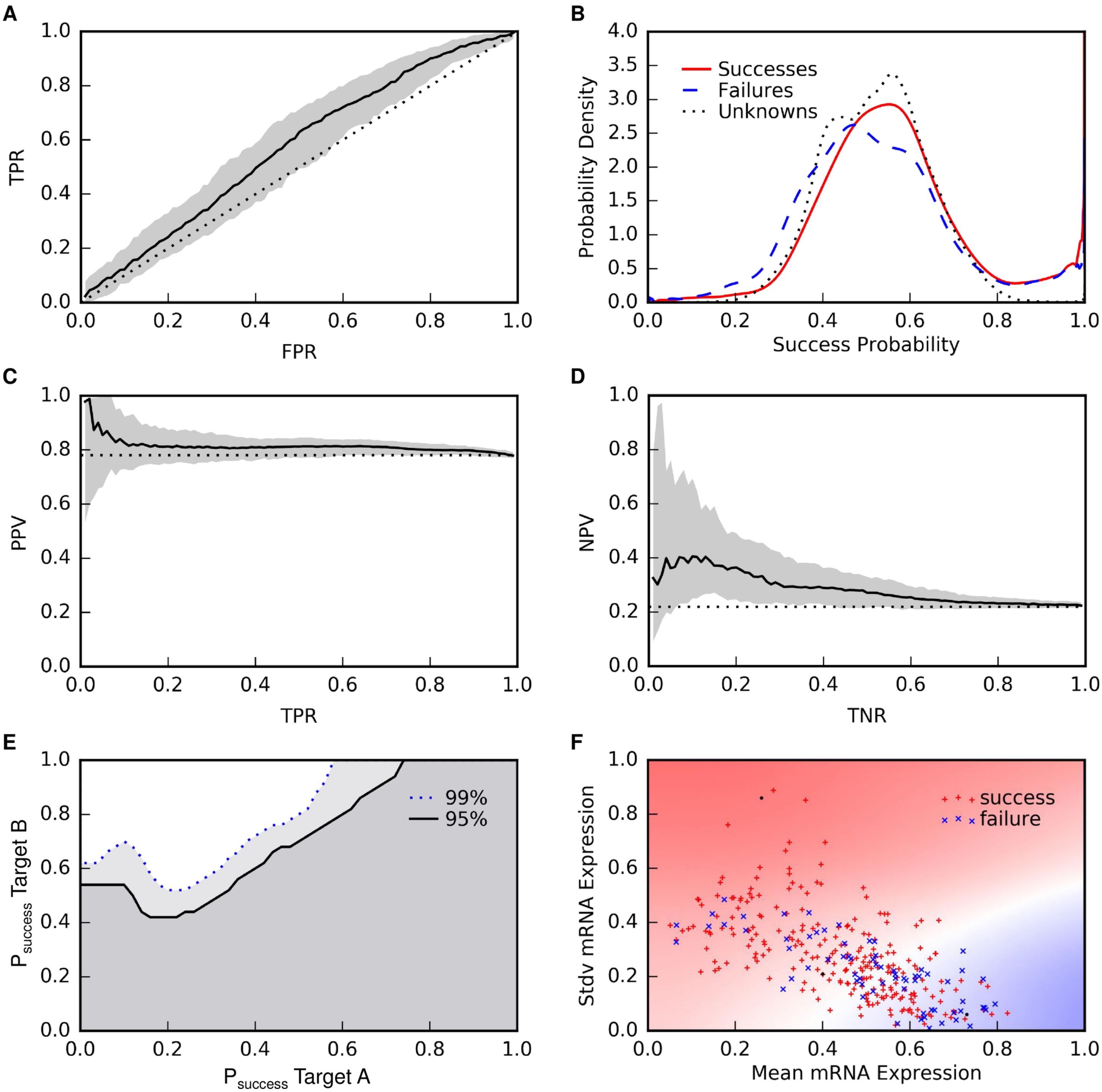

Fig 3. Classifier performance. **(A)** Receiver operating characteristic (ROC) curve. The solid black line indicates the median performance across 200 repetitions of 5-fold cross-validation and the gray area indicates the range of the 2.5 and 97.5 percentiles. The dotted black line indicates the performance of random rankings. **(B)** Distributions of the probability of success predicted by the classifier for the successful, failed, and unlabeled targets. **(C)** Precision-recall curve for success predictions. **(D)** Precision-recall curve for failure predictions. **(E)** Pairwise target comparisons. For each pair of targets, we computed the fraction of repetitions of cross-validation in which Target B had a higher predicted probability of success greater than Target A. The heatmap illustrates this fraction, thresholded at 0.95 or 0.99, plotted as a function of the median predicted probabilities of success of two targets. The upper left region is where the classifier is 95% (above solid black line) or 99% (above dotted blue line) consistent in predicting greater probability of success of Target B than Target A. **(F)** Relationship between features and phase III outcomes. Heat map showing the projection of the predicted success probabilities onto the two dominant features selected for the classifier: mean expression across tissues and standard deviation of expression across tissues. Red, white, and blue background colors correspond to 1, 0.5, and 0 success probabilities. Red plusses and blue crosses mark the locations of the success and failure examples. It appears the model has learned that failures tend to have high mean expression and low standard deviation of expression across tissues,

while successes tend to have low mean expression and high standard deviation of expression. The success and failure examples are not well separated, indicating that we did not discover enough features to fully explain why targets succeed or fail in phase III clinical trials.

https://doi.org/10.1371/journal.pcbi.1006142.g003

Feature interpretation. To interpret the relationship inferred by the classifier between target features and outcomes, we created a heatmap of the probability of success predicted by the classifier projected onto the two features predominantly selected for the model: mean expression and standard deviation of expression across tissues (Fig 3F). The probability of success was high in the subspace with low mean expression and high standard deviation of expression, and transitioned to low probability in the subspace with high mean expression and low standard deviation of expression. This trend appeared to be consistent with the distribution of the success and failure examples in the space.

Discussion

Gene expression predicts phase III outcome

We searched over 150,000 target features from 67 datasets covering 16 feature types for predictors of target success or failure in phase III clinical trials (Table 1, Fig 1). We found several features significantly correlated with phase III outcome, robust to re-sampling, and generalizable across target classes (Table 2). To assess the usefulness of such features, we implemented a nested cross-validation routine to select features, train a classifier to predict the probability a target will succeed in phase III clinical trials, and estimate the stability and generalization performance of the model (Figs 2 and 3, Tables 3, 4 and 5). Ultimately, we found two features useful for predicting success or failure of targets in phase III clinical trials. Successful targets tended to have low mean mRNA expression across tissues and high standard deviation of

Table 5. Classifier performance statistics.

Statistic	2.5 Percentile	Median	97.5 Percentile
True Positives (TP)	91	220	243
False Positives (FP)	16	52	65
True Negatives (TN)	5	16	52
False Negatives (FN)	1	24	154
True Positive Rate (TPR)	0.370	0.903	0.995
False Positive Rate (FPR)	0.232	0.762	0.928
False Negative Rate (FNR)	0.005	0.096	0.630
True Negative Rate (TNR)	0.072	0.237	0.768
Misclassification Rate (MCR)	0.206	0.241	0.542
Accuracy (ACC)	0.458	0.759	0.794
False Discovery Rate (FDR)	0.149	0.194	0.213
Positive Predictive Value (PPV)	0.787	0.806	0.851
False Omission Rate (FOMR)	0.233	0.583	0.741
Negative Predictive Value (NPV)	0.259	0.417	0.767
Area Under Receiver Operating Characteristic Curve (AUROC)	0.512	0.574	0.615
Area Under Precision-Recall Curve (AUPR)	0.777	0.811	0.836
Positive Likelihood Ratio (PLR)	1.058	1.184	1.619
Negative Likelihood Ratio (NLR)	0.086	0.402	0.819
Diagnostic Odds Ratio (DOR)	1.748	3.066	13.344
Risk Ratio (RR)	1.143	1.387	3.447
Matthews Correlation Coefficient (MCC)	0.100	0.178	0.251

https://doi.org/10.1371/journal.pcbi.1006142.t005

mRNA expression across tissues (Fig 3F). These features were significant in multiple gene expression datasets, which increased our confidence that their relationship to phase III outcome was real, at least for the targets in our sample, which included only targets of selective drugs indicated for non-cancer diseases.

One interpretation of why the gene expression features were predictive of phase III outcome is that they are informative of the specificity of a target's expression across tissues. A target with tissue specific expression would have a high standard deviation relative to its mean expression level. Tissue specific expression has been proposed by us and others as a favorable target characteristic in the past [4, 14, 60–62], but the hypothesis had not been evaluated empirically using examples of targets that have succeeded or failed in clinical trials. For a given disease, if a target is expressed primarily in the disease tissue, it is considered more likely that a drug will be able to exert a therapeutic effect on the disease tissue while avoiding adverse effects on other tissues. Additionally, specific expression of a target in the tissue affected by a disease could be an indicator that dysfunction of the target truly causes the disease.

The distribution of the success and failure examples in feature space (Fig 3F) partially supports the hypothesis that tissue specific expression is a favorable target feature. Successes were enriched among targets with low mean expression and high standard deviation of expression (tissue specific expression), and failures were enriched among targets with high mean expression and low standard deviation of expression (ubiquitous expression). However, it does not hold in general that, at any given mean expression level, targets with high standard deviation of expression tend to be more successful than targets with low standard deviation of expression. To further investigate the relationship between these features and phase III clinical trial outcomes, we re-ran the entire modeling pipeline (Fig 2) with gene expression entropy, a feature explicitly quantifying specificity of gene expression across tissues [21], appended to each tissue expression dataset (S1 Text). Model performance was unchanged (S1 Fig); gene expression entropy across tissues became the dominant selected feature, appearing in 610 models over 1000 train-test cycles; and mean gene expression across tissues remained an important feature, appearing in 381 models (S6 Table). To find concrete examples illustrating when tissue expression may be predictive of clinical trial outcomes, we pulled additional information from the Pharmaprojects database about targets at the two extremes of tissue expression (tissue specific or ubiquitous). We found examples of: 1) successful tissue specific targets where the target is specifically expressed in the tissue affected by the disease (Table 6), 2) failed tissue specific targets with plausible explanations for failure despite tissue specific expression (Table 7), 3) failed ubiquitously expressed targets (Table 8), and 4) successful ubiquitously expressed targets with plausible explanations for success despite ubiquitous expression (Table 9). Our results encourage further investigation of the relationship between tissue specific expression and clinical trial outcomes. Deeper insight may be gleaned from analysis of clinical trial outcomes of target-indication pairs using gene expression features explicitly designed to quantify specificity of a target's expression in the tissue(s) affected by the disease treated in each clinical trial.

Table 6. Examples of successful tissue specific targets.

Target	Indication	Expression	Outcome
PNLIP	Pancreatic insufficiency	Pancreas 5-fold higher than other tissues	Success
MMP8	Osteoarthritis	Bone marrow 4-fold higher than other tissues	Success
ATP4A	Ulcer, gastro-esophageal reflux	Stomach 3-fold higher than other tissues	Success
GABRA1	Neurological diseases (anxiety, depression, addiction, pain, insomnia, epilepsy)	Brain 3-fold higher than other tissues	Success

https://doi.org/10.1371/journal.pcbi.1006142.t006

Table 7. Examples of failed tissue specific targets with plausible exceptions.

Target	Indication	Expression	Outcome	Exception
BPI	Bacterial infections	Bone marrow 3-fold higher than other tissues	Failure	The drug is recombinant BPI, which is used for its anti-bacterial properties, thus modulation of endogenous BPI is not directly relevant to efficacy of the therapy
TSHR	Goiter	Thyroid 4-fold higher than other tissues	Failure	The trial was canceled before enrollment, thus perhaps TSHR should not be counted as a phase III failure

https://doi.org/10.1371/journal.pcbi.1006142.t007

Caveats and limitations

Latent factors (variables unaccounted for in this analysis) could confound relationships between target features and phase III outcomes. For example, diseases pursued vary from target to target, and a target's expression across tissues may be irrelevant for diseases where drugs can be delivered locally or for Mendelian loss-of-function diseases where treatment requires systemic replacement of a missing or defective protein. Also, clinical trial failure rates vary across disease classes [2]. Although we excluded targets of cancer therapeutics from our analysis, we otherwise did not control for disease class as a confounding explanatory factor. Modalities (e.g. small molecule, antibody, antisense oligonucleotide, gene therapy, or protein replacement) and directions (e.g. activation or inhibition) of target modulation also vary from target to target and could be confounding explanatory factors or alter the dependency between target features and outcomes.

The potential issues described above are symptoms of the fact that our analysis (and any analysis of clinical trial outcomes) attempts to draw conclusions from a small (331 targets with only 72 failures) and biased sample [63, 64]. The large uncertainty in the performance of the classifier across 200 repetitions of 5-fold cross-validation is evidence of the difficulty in finding robust signal in such a small dataset (Fig 3). For example, in the region where the model predicts highest probability of success (low mean expression and high standard deviation of expression), there are no failed phase III targets (Fig 3F), which is why the median PPV rises nearly to 1 (Fig 3C), but targets with phase III outcomes sparsely populate this region, so the PPV varies widely depending upon how targets happen to fall into training and testing sets during cross-validation. The small sample issue is compounded by latent factors, such as target classes, disease classes, modalities, and directions of target modulation, that are not uniformly represented in the sample. Correlations between target features and clinical trial outcomes likely depend on these factors, but attempts to stratify, match, or otherwise control for these factors are limited by the sample size. (The number of combinations of target class, disease class, modality, and direction of modulation exceeds the sample size.) We employed several tests to build confidence that our findings generalize across target classes, but did not address other latent factors. Consequently, we cannot be sure that conclusions drawn from this study apply equally to targets modulated in any direction, by any means, to treat any disease. For specific cases, expert knowledge and common sense should be relied upon to determine whether conclusions from this study (or similar studies) are relevant.

Another limitation is selection bias [63, 64]. Targets of drugs are not randomly selected from the genome and cannot be considered representative of the population of all possible

Table 8. Examples of failed ubiquitously expressed targets.

Target	Indication	Expression	Outcome
DPP8	Heart failure	Ubiquitous	Failure
CSNK2B	Human papilloma virus infection	Ubiquitous	Failure

https://doi.org/10.1371/journal.pcbi.1006142.t008

Table 9. Examples of successful ubiquitously expressed targets with plausible exceptions.

Target	Indication	Expression	Outcome	Exception
MTOR	Restenosis	Ubiquitous	Success	Tissue specificity is achieved via the delivery method (drug eluting stent)
IFNAR1	Eye infections	Ubiquitous	Success	Tissue specificity is achieved via the delivery method (eye drops)
GBA	Gaucher's disease	Ubiquitous	Success	Gaucher's disease is a loss-of-function genetic disorder affecting multiple organ systems, thus therapy requires system-wide replacement of the defective enzyme

https://doi.org/10.1371/journal.pcbi.1006142.t009

targets. Likewise, diseases treated by drugs are not randomly chosen; therefore, phase III clinical trial outcomes for each target cannot be considered representative of the population of all possible outcomes. Although we implemented tests to build confidence that our findings can generalize to new targets and new target classes, ultimately, no matter how we dissect the sample, a degree of uncertainty will always remain about the relevance of any findings for new targets that lack a representative counterpart in the sample.

Additionally, data processing and modeling decisions have introduced bias into the analysis. For example, we restricted the analysis to phase III clinical trial outcomes because failures in phase III are more likely to be due to lack of target efficacy than failures in earlier phases, but factors unrelated to target efficacy still could explain many of the phase III failures, such as poor target engagement, poorly defined clinical trial endpoints, and a poorly defined patient population. Also, we scored each target as successful or failed by its best outcome in all applicable (selective drug, non-cancer indication) phase III clinical trials. This approach ignores nuances. A target that succeeded in one trial and failed in all others is treated as equally successful as a target that succeeded in all trials. Also, the outcome of a target tested in a single trial is treated as equally certain as the outcome of a target tested in multiple trials. Representing target outcomes as success rates or probabilities may provide better signal for discovering features predictive of outcomes.

Another decision was to use datasets of features as we found them, rather than trying to reason about useful features that could be derived from the original data. Because of the breadth of data we interrogated, the effort and expertise necessary to hand engineer features equally well across all datasets exceeded our resources. Others have had success hand engineering features for similar applications in the past, particularly with respect to computing topological properties of targets in protein-protein interaction networks [18, 20, 21]. This analysis could benefit from such efforts, potentially changing a dataset or feature type from yielding no target features correlated with phase III outcomes to yielding one or several useful features [22]. On a related point, because we placed a priority on discovering interpretable features, we performed dimensionality reduction by averaging groups of highly correlated features and concatenating their (usually semantically related) labels. Dimensionality reduction by principal components analysis [65] or by training a deep auto-encoder [66] could yield more useful features, albeit at the expense of interpretability.

We also employed a stringent univariate feature selection step (Fig 2, Step 2) to bias our analysis toward yielding a simple and interpretable model. In doing so, we diminished the chance of the multivariate feature selection step (Fig 2, Step 4) finding highly predictive combinations of features that individually were insignificantly predictive. We addressed this concern by re-running the entire modeling pipeline (Fig 2) with the threshold for the univariate feature selection step made less stringent by eliminating the multiple hypothesis testing correction and accepting features with nominal p-values less than 0.05 (S2 Text). This allowed hundreds of features to pass through to the multivariate feature selection step (Random Forest with incremental feature elimination) and ultimately dozens of features (median of 73) were selected for each of the final models in the 1000 train-test cycles (S7 Table). Despite this increase in number of features, the mean expression and standard deviation of expression features were still robustly selected, appearing in 958 and 745 models, respectively, and the models had a median AUROC of 0.56 and AUPRC of 0.75, performing no better than the simple models (S2 Fig). This finding suggests that our sample size was not large enough to robustly select predictive combinations of features from a large pool of candidate features [67, 68].

We cannot stress enough the importance of taking care not to draw broad conclusions from our study, particularly with respect to the apparent dearth of features predictive of target success or failure. We examined only a specific slice of clinical trial outcomes (phase III trials of selective drugs indicated for non-cancer diseases) summarized in a particular way (net outcome per target, as opposed to outcome per target-indication pair). Failure of a feature to be significant in our analysis should not be taken to mean it has no bearing on target selection. For example, prior studies have quantitatively shown that genetic evidence of disease association(s) is a favorable target characteristic [3, 36], but we did not find a significant correlation between genetic evidence and target success in phase III clinical trials. Our finding is consistent with the work of Nelson et al. [36], who investigated the correlation between genetic evidence and drug development outcomes at all phases and found a significant correlation overall and at all phases of development except phase III. As a way of checking our work, we applied our methods to test for features that differ between targets of approved drugs and the remainder of the druggable genome (instead of targets of phase III failures), and we recovered the finding of Nelson et al. that targets of approved drugs have significantly more genetic evidence than the remainder of the druggable genome (S8 Table). This example serves as a reminder to be cognizant of the domain of applicability of research findings. Though we believe we have performed a rigorous and useful analysis, we have shed light on only a small piece of a large and complex puzzle.

Advances in machine learning enable and embolden us to create potentially powerful predictive models for target selection. However, as described in the limitations, scarce training data are available, the data are far from ideal, and we must be cautious about building models with biased data and interpreting their predictions. For example, many features that appeared to be significantly correlated with phase III clinical trial outcomes in our primary analysis did not hold up when we accounted for target class selection bias. This study highlights the need for both domain knowledge and modeling expertise to tackle such challenging problems.

Conclusion

Our analysis revealed several features that significantly separated targets of approved drugs from targets of drug candidates that failed in phase III clinical trials. This suggested that it is feasible to construct a model integrating multiple interpretable target features derived from Omics datasets to inform target selection. Only features derived from tissue expression datasets were promising predictors of success versus failure in phase III, specifically, mean mRNA expression and standard deviation of expression across tissues. Although these features were significant at a false discovery rate cut-off of 0.05, their effect sizes were too small to be useful for classification of the majority of untested targets, however, even a two-fold improvement in target quality can dramatically increase R&D productivity [69]. We identified 943 targets predicted to be twice as likely to fail in phase III clinical trials as past phase III targets, and, therefore, should be flagged as having unfavorable expression characteristics. We also identified 2,700,856 target pairs predicted with 99% consistency to have a 2-fold difference in success probability, which could be useful for prioritizing short lists of targets with attractive disease relevance.

It should be noted that our analysis was not designed or powered to show that specific datasets or data types have no bearing on target selection. There are many reasons why a dataset may not have yielded any significant features in our analysis. In particular, data processing and filtering choices could determine whether or not a dataset or data type has predictive value. Also, latent factors, such as target classes, disease classes, modalities, and directions of target modulation, could confound or alter the dependency between target features and clinical trial outcomes. Finally, although we implemented tests to ensure robustness and generalizability of the target features significantly correlated with phase III outcomes, selection bias in the sample of targets available for analysis is a non-negligible limitation of this study and others of its kind. Nevertheless, we are encouraged by our results and anticipate deeper insights and better models in the future, as researchers improve methods for handling sample biases and learn more informative features.

Methods

Data

Clinical outcomes. We extracted data from Citeline's Pharmaprojects database [38] (downloaded May 27, 2016), reformatting available XML data into a single tab-delimited form having one row for each asset (i.e. drug or drug candidate)/company combination. For each asset, known targets, identified with EntrezGene [70] IDs and symbols, and indications are reported. We obtained 107,120 asset-indication pairs and 37,211 asset-target pairs, correcting a single outdated EntrezGene ID, for SCN2A, which we updated from 6325 to 6326.

An overall pipeline status of each asset (e.g. "Launched", "Discontinued", "No Development Reported") is reported in a single field ("Status"), and detailed information for each indication being pursued is dispersed throughout several other fields (e.g., "Key Event Detail", "Overview", etc.). While many assets have been tried against a single indication, and thus the status of the asset-indication pair is certain, the majority (N = 61,107) of asset-indication pairs are for assets with multiple indications. For those pairs, we used a combination of string searching of these fields and manual review of the results to determine the likely pipeline location and status of each indication. For example, we excluded efforts where a trial of an asset was reported as planned, but no further information was available. Asset-indication pairs were thus assigned a status of Successful ("Launched", "Registered", or "Pre-registration"), Failed ("Discontinued", "No Development Reported", "Withdrawn", or "Suspended"), or In Progress, consisting of 9,337, 72,269 and 25,159 pairs, respectively. We then used the pipeline location to assign each asset-indication pair to one of 10 outcomes: Succeeded, In Progress-Preclinical, In Progress-Phase I, In Progress-Phase II, In Progress-Phase III, Failed-Preclinical, Failed-Phase I, Failed-Phase II, Failed-Phase III, and Failed-Withdrawn. We discarded indications which were diagnostic in nature or unspecified, mapping the remainder to Medical Subject Headings (MeSH) [71]. We also observed that only 24% of the failures reported in Pharmaprojects are clinical failures, suggesting a clinical success rate of nearly 35%, much higher than typically cited [69].

We joined the list of asset-indication-outcome triples with the list of asset-target pairs to produce a list of asset-target-indication-outcome quadruples. We then filtered the list to remove: 1) assets with more than one target, 2) non-human targets, 3) cancer indications (indications mapped to MeSH tree C04), and 4) outcomes labeled as In Progress at any stage or Failed prior to Phase III. We scored the remaining targets (N = 331) as Succeeded (N = 259), if the target had at least one successful asset remaining in the list, or Failed (N = 72), otherwise.

Target features. We obtained target features from the Harmonizome [39], a recently published collection of features of genes and proteins extracted from over 100 Omics datasets. We downloaded (on June 30, 2016) a subset of Harmonizome datasets that were in the public domain or GSK had independently licensed (Table 1). Each dataset was structured as a matrix with genes labeling the rows and features such as diseases, phenotypes, tissues, and pathways labeling the columns. Genes were identified with EntrezGene IDs and symbols, enabling facile integration with the clinical outcome data from Pharmaprojects. Some datasets were available on the Harmonizome as a "cleaned" version and a "standardized" version. In all instances, we used the cleaned version, which preserved the original data values (e.g. gene expression values), as opposed to the standardized version, in which the original data values were transformed into scores indicating relative strengths of gene-feature associations intended to be comparable across datasets. The data matrices were quantitative and filled-in (e.g. gene expression measured by microarray), quantitative and sparse (e.g. protein expression measured by immunohistochemistry), or categorical (i.e. binary) and sparse (e.g. pathway associations curated by experts). We standardized quantitative, filled-in features by subtracting the mean and then dividing by the standard deviation. We scaled quantitative, sparse features by dividing by the mean. We included the mean and standard deviation calculated along the rows of each dataset as additional target features. We excluded features that had fewer than three non-zero values for the targets with phase III clinical trial outcomes. The remaining features, upon which our study was based, have been deposited at https://github.com/arouillard/omic-features-successful-targets.

Dimensionality reduction

Our goals in performing dimensionality reduction were to identify groups of highly correlated features, avoid excessive multiple hypothesis testing, and maintain interpretability of features. For each dataset, we computed pair-wise feature correlations (r) using the Spearman correlation coefficient [72–74] for quantitative, filled-in datasets, and the cosine coefficient [73, 74] for sparse or categorical datasets. We thresholded the correlation matrix at $r^2 = 0.5$ (for the Spearman correlation coefficient, this corresponds to one feature explaining 50% of the variance of another feature, and for the cosine coefficient, this corresponds to one feature being aligned within 45 degrees of another feature) and ordered the features by decreasing number of correlated features. We created a group for the first feature and its correlated features. If the dataset mean was included in the group, we replaced the group of features with the dataset mean. Otherwise, we replaced the group of features with the group mean and assigned it the label of the first feature (to indicate that the feature represents the average of features correlated with the first feature), while also retaining a list of the labels of all features included in the group. We continued through the list of features, repeating the grouping process as described for the first feature, except excluding features already assigned to a group from being assigned to a second group.

Feature selection

We performed permutation tests [40, 41] to find features with a significant difference between successful and failed targets. We used permutation testing in order to apply the same significance testing method to all features. The features in our collection had heterogeneous shapes of their distributions and varying degrees of sparsity, and therefore no single parametric test would be appropriate for all features. Furthermore, individual features frequently violated assumptions required for parametric tests, such as normality for the t-test (for continuous-valued features) or having at least five observations in each entry of the contingency table for the Chi-squared test (for categorical features). For each feature, we performed 10^5 success/failure label permutations to obtain a null distribution for the difference between the means of successful and failed targets, and then calculated an empirical two-tailed p-value as the fraction of permutations that yielded a difference between means at least as extreme as the actual observed difference. We used the Benjamini-Yekutieli method [42] to correct for multiple hypothesis testing within each dataset and accepted features with corrected p-values less than 0.05 as significantly correlated with phase III clinical trial outcomes, thus controlling the false discovery rate at 0.05 within each dataset.

Feature robustness and generalizability

Robustness to sample variation. We used bootstrapping [53, 54] to investigate how robust our significance findings were to variation in the success and failure examples. We created a bootstrap sample by sampling with replacement from the original set of examples to construct an equal sized set of examples. For each dataset that yielded significant features in our primary analysis, we repeated the analysis on the bootstrap sample and recorded whether the features were still significant at the aforementioned 0.05 false discovery rate cut-off. We performed this procedure on 1000 bootstrap samples and quantified the replication probability [55] of each feature as the fraction of bootstraps showing a significant correlation between the feature and phase III clinical trial outcomes. We accepted features with replication probabilities greater than 0.8 [55] as robust to sample variation.

Robustness to target class variation. We tested if any of the significance findings depended upon the presence of targets from a single target class in our sample. We obtained target class labels (i.e. gene family labels) from the HUGO Gene Nomenclature Committee [56] (downloaded April 19, 2016) and created binary features indicating target class membership. Using the same permutation testing and multiple hypothesis testing correction methods described above for feature selection, we tested if any target classes were significantly correlated with phase III clinical trial outcomes. Then, we tested if the significant target classes were correlated with any significant features. Such features might be correlated with clinical outcome only because they are surrogate indicators for particular target classes that have been historically very successful or unsuccessful, as opposed to the features being predictors of clinical outcome irrespective of target class. To test this possibility, we performed a bootstrapping procedure as described above, except did not allow examples from target classes correlated with clinical outcome to be drawn when re-sampling. Thus, the modified bootstrapping procedure provided replication probabilities conditioned upon missing information about target classes correlated with clinical outcome. We accepted features with replication probabilities greater than 0.8 as robust to target class variation.

Generalization across target classes. We implemented a modified permutation test, inspired by the approach of Epstein et al. [57] to correct for confounders in permutation testing, to select features correlated with phase III clinical trial outcomes while controlling for target class as a confounding explanatory factor. In the modified permutation test, success/failure labels were shuffled only within target classes, so the sets of null examples had the same ratios of successes to failures within target classes as in the set of observed examples. Consequently, features had to correlate with clinical outcome within multiple classes to be significant, while features that discriminated between classes would not be significant. We performed bootstrapping as described previously to obtain replication probabilities for the significant features, in this case conditioned upon including target class as an explanatory factor. We accepted features with replication probabilities greater than 0.8 as generalizable across target classes represented in the sample.

Clinical outcome classifier

We trained a classifier to predict target success or failure in phase III clinical trials, using a procedure like the above for initial feature selection, then using cross-validation to select a model type (Random Forest or logistic regression) and subset of features useful for prediction. We used an outer cross-validation loop with 5-folds repeated 200 times, yielding a total of 1000 train-test cycles, to estimate the generalization performance and stability of the feature selection and model selection procedure [58]. Each train-test cycle had five steps: 1) splitting examples into training and testing sets, 2) univariate feature selection on the training data, 3) aggregation of significant features from different datasets into a single feature matrix, 4) model selection and model-based (multivariate) feature selection on the training data, and 5) evaluation of the classifier on the test data.

Step 2: Univariate feature selection. Beginning with the non-redundant features obtained from dimensionality reduction, we performed modified permutation tests to find features with a significant difference between successful and failed targets in the training examples. As described above, for the modified permutation test, success/failure labels were shuffled only within target classes. This was done to control for target class as a confounding factor that might explain correlations between phase III outcomes and features. For each feature, we performed 10^4 success/failure label permutations and calculated an empirical two-tailed p-value. We corrected for multiple hypothesis testing within each dataset and accepted features with corrected p-values less than 0.05.

Step 3. Feature aggregation. Significant features from different datasets, each having different target coverage, had to be aggregated into a single feature matrix prior to training a classifier. When features from many datasets were aggregated, we found that the set of targets with no missing data across all features could become very small. To mitigate this, we excluded features from non-human datasets and small datasets (fewer than 2,000 genes). We also excluded features from the Allen Brain Atlas human brain expression atlas, unless there were no other significant features, because we noticed it had poor coverage of targets with phase III outcomes (287) compared to other expression atlases, such as BioGPS (320), GTEx (328), and HPA (314), which almost always yielded alternative significant expression-based features. After aggregating features into a single matrix, we min-max scaled the features so that features from different datasets would have the same range of values (from 0 to 1).

To reduce redundancy in the aggregated feature matrix, we grouped features as described for the primary analysis. We used the cosine coefficient to compute pair-wise feature correlations because some features were sparse. Instead of replacing groups of correlated features with the group mean, we selected the feature in each group that was best correlated with phase III outcomes, because we preferred not to create features derived from multiple datasets.

Step 4. Model selection and model-based feature selection. We hypothesized that a Random Forest classifier [75] would be a reasonable model choice because the Random Forest model does not make any assumptions about the distributions of the features and can seamlessly handle a mixture of quantitative, categorical, filled-in, or sparse features. Furthermore, we expected each train-test cycle to yield only a handful of significant features. Consequently, we would have 10- to 100-fold more training examples than features and could potentially afford to explore non-linear feature combinations. We also trained logistic regression classifiers and used an inner cross-validation loop (described below) to choose between Random Forest and logistic regression for each train-test cycle of the outer cross-validation loop. We used the implementations of the Random Forest and logistic regression classifiers available in the Scikit-learn machine learning package for Python. To correct for unequal class sizes during training, the loss functions of these models weighted the training examples inversely proportional to the size of each example's class.

We performed incremental feature elimination with an inner cross-validation loop to 1) choose the type of classifier (Random Forest or logistic regression) and 2) choose the smallest subset of features needed to maximize the performance of the classifier. First, we trained Random Forest and logistic regression models using the significant features aggregated in Step 2, performing 5-fold cross-validation repeated 20 times to obtain averages for the area under the receiver operating characteristic curve (AUROC) and area under the precision recall curve (AUPR). We also obtained average feature importance scores from the Random Forest model. Next, we eliminated the feature with lowest importance score and trained the models using the reduced feature set, performing another round of 5-fold cross-validation repeated 20 times to obtain AUROC, AUPR, and feature importance scores. We continued eliminating features then obtaining cross-validation performance statistics and feature importance scores until no features remained. Then, we found all models with performance within 95% of the maximum AUROC and AUPR. If any logistic regression models satisfied this criterion, we selected the qualifying logistic regression model with fewest features. Otherwise, we selected the qualifying Random Forest model with fewest features.

Step 5. Classifier evaluation. For each train-test cycle, after selecting a set of features and type of model (Random Forest or logistic regression) in Step 4, we re-fit the selected model to the training data and predicted success probabilities for targets in the test set as well as unlabeled targets. For each round of 5-fold cross-validation, we computed the classifier's receiver operating characteristic curve, precision-recall curve, and performance summary statistics, including the true positive rate, false positive rate, positive predictive value, negative predictive value, and Matthews correlation coefficient.

We computed distributions of the log odds ratios predicted by the classifier (log of the ratio of the predicted probability of success over the probability of failure) for the successful, failed, and untested (unlabeled) targets, aggregating predicted probabilities from the 200 repetitions of 5-fold cross-validation. Histograms of the log odds ratios for the three groups of targets were roughly bell-shaped, so we fit the distributions using kernel density estimation [59] with a Gaussian kernel and applied Silverman's rule for the bandwidth. We transformed the fitted distributions from a function of log odds ratio to a function of probability of success using the rule $pdf(x) = pdf(y)^{*}|dy/dx|$.

We created a heatmap of the probability of success predicted by the classifier projected onto the two dominant features in the model: mean mRNA expression across human tissues and standard deviation of mRNA expression across human tissues. We examined the heatmap to interpret the classifier's decision function and assess its plausibility.

To more concretely assess the usefulness of the classifier, we found the probability cut-off corresponding to the maximum median positive predictive value and determined the number of unlabeled targets predicted to succeed at that cut-off. Likewise, we found the probability cut-off corresponding to the maximum median negative predictive value and determined the number of unlabeled targets predicted to fail at that cut-off. We also created a heatmap illustrating the separation needed between the median predicted success probabilities of two targets in order to be confident that one target is more likely to succeed than the other. This heatmap was created by calculating the fraction of times Target B had greater probability of success than Target A across the 200 repetitions of 5-fold cross-validation, for all pairs of targets.

Implementation

Computational analyses were written in Python 3.4.5 and have the following package dependencies: Fastcluster 1.1.20, Matplotlib 1.5.1, Numpy 1.11.3, Requests 2.13.0, Scikit-learn 0.18.1, Scipy 0.18.1, and Statsmodels 0.6.1. Code, documentation, and data have been deposited on GitHub at https://github.com/arouillard/omic-features-successful-targets.

Supporting information

S1 Table. List of targets with their phase III outcome labels and predicted success probabilities for 200 cross-validation repetitions.
(XLSX)

S2 Table. List of non-redundant features with their similar features and p-values from the basic permutation test.
(XLSX)

S3 Table. List of classifier attributes (selected features, selected model type, and test performance) for 1000 train-test cycles.
(XLSX)

S4 Table. Comparison of inner cross-validation loop AUROC and AUPR values between Random Forest and logistic regression models for 1000 train-test cycles.
(XLSX)

S5 Table. List of classifier test performance statistics for 200 cross-validation repetitions.
(XLSX)

S6 Table. For the modeling pipeline with gene expression entropy across tissues included as a candidate feature, list of classifier attributes (selected features, selected model type, and test performance) for 1000 train-test cycles.
(XLSX)

S7 Table. For the modeling pipeline with heavier reliance on multivariate feature selection (less stringent univariate feature selection), list of classifier attributes (selected features, selected model type, and test performance) for 1000 train-test cycles.
(XLSX)

S8 Table. Cases illustrating how the significance of genetic evidence (and likely other types of evidence) as a predictor of target success depends on which targets are compared.
(XLSX)

S1 Fig. Evaluation of entropy as an alternative feature quantifying tissue specificity of target expression. (A) Relationship between coefficient of variation (standard deviation/mean) of gene expression across tissues and entropy of gene expression across tissues. Entropy of a target was defined as sum[$P_i \log_2(1/P_i)$] where $P_i = E_i/\mathrm{sum}(E_i)$ and E_i is the target's expression in the i^{th} tissue). Coefficient of variation and entropy were computed using un-log-transformed expression values. The strong (nonlinear) correlation indicates that entropy captures similar information about the distribution of a target's expression across tissues as the pair of mean and standard deviation. **(B)** Distribution of area under the receiver operating characteristic curve (AUROC) values from 200 repetitions of 5-fold cross-validation. The light gray distribution corresponds to the original analysis that included the mean and standard deviation of gene expression across tissues as candidate target features. The dark blue distribution corresponds to the alternative analysis that replaced the mean and standard deviation features with entropy of gene expression across tissues as a candidate target feature. The models had nearly identical AUROC distributions. **(C)** Distribution of area under the precision-recall curve (AUPRC) values from 200 repetitions of 5-fold cross-validation. The models had nearly identical AUPRC distributions.
(TIF)

S2 Fig. Evaluation of multivariate feature selection. (A) Distribution of area under the receiver operating characteristic curve (AUROC) values from the alternative modeling

pipeline with weak univariate feature selection (nominal p-value less than 0.05), which allowed hundreds of features to pass through to the multivariate feature selection step (Random Forest feature ranking with incremental feature elimination). The light gray distribution corresponds to the inner cross-validation loop performance of the model. The dark blue distribution corresponds to the outer cross-validation loop performance of the model. There is a large discrepancy between the distributions, indicating failure of the inner cross-validation loop to appropriately tune the complexity of the model. **(B)** Distribution of area under the precision-recall curve (AUPRC) values from the alternative modeling pipeline with weak univariate feature selection. **(C and D)** Distribution of AUROC and AUPRC values from the original modeling pipeline with strong univariate feature selection (multiple hypothesis testing corrected p-value less than 0.05). There is little discrepancy between generalization performance estimated by the inner and outer cross-validation loops, indicating appropriate tuning of the complexity of the model. **(E)** Illustration of the mismatch between training and testing examples that arises from splitting our small and heterogeneous sample of targets. Each point in the scatterplot corresponds to a feature that passed through the weak univariate feature selection step of the alternative modeling pipeline. Plotted on the horizontal axis is the difference between the median of the Class I training examples and the median of the Class J training examples, where if Class I is success, then Class J is failure, and vice versa. Plotted on the vertical axis is the difference between the median of the Class I TESTING examples and the median of the Class J training examples. For features in the first quadrant, the Class I training and testing examples both have medians greater than the Class J training examples, so there is no mismatch between the training and testing distributions of Class I. In the third quadrant, the Class I training and testing examples both have medians less than the Class J training examples, so again there is no mismatch. In the second and fourth quadrants, however, the Class I training and testing examples have medians in opposite directions relative to the Class J training examples, so a decision boundary separating the Class I and Class J training examples will fail to generalize to the Class I testing examples. Without looking at the testing examples, there is no way to distinguish with certainty generalizable features from non-generalizable (mismatched) features. Note, however, that the probability of mismatch decreases as the magnitude of the difference between Class I and Class J in the training set increases. Consequently, stringent univariate feature selection is one way to guard against non-generalizable features being selected for the classifier. **(F)** Distribution of the fraction of features with mismatched distributions that pass through to the multivariate feature selection step when univariate feature selection is weak.
(TIF)

S1 Text. Description of analysis of gene expression entropy across tissues as a feature quantifying tissue specificity of target expression.
(DOCX)

S2 Text. Description of analysis of multivariate feature selection (modeling pipeline with less stringent univariate feature selection).
(DOCX)

Acknowledgments

Many thanks to Dr. Subhas Chakravorty for assisting with access to and processing of the Pharmaprojects data and to Dr. David Cooper for helpful direction regarding nested cross-validation.

Author Contributions

Conceptualization: Mark R. Hurle, Pankaj Agarwal.

Data curation: Andrew D. Rouillard, Mark R. Hurle.

Formal analysis: Andrew D. Rouillard.

Methodology: Andrew D. Rouillard, Mark R. Hurle, Pankaj Agarwal.

Software: Andrew D. Rouillard.

Supervision: Pankaj Agarwal.

Visualization: Andrew D. Rouillard.

Writing – original draft: Andrew D. Rouillard.

Writing – review & editing: Andrew D. Rouillard, Mark R. Hurle, Pankaj Agarwal.

References

1. Arrowsmith J, Miller P. Trial watch: phase II and phase III attrition rates 2011–2012. Nat Rev Drug Discov. 2013; 12(8):569. https://doi.org/10.1038/nrd4090 PMID: 23903212.
2. Harrison RK. Phase II and phase III failures: 2013–2015. Nat Rev Drug Discov. 2016; 15(12):817–8. https://doi.org/10.1038/nrd.2016.184 PMID: 27811931.
3. Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, et al. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov. 2014; 13 (6):419–31. https://doi.org/10.1038/nrd4309 PMID: 24833294.
4. Gashaw I, Ellinghaus P, Sommer A, Asadullah K. What makes a good drug target? Drug Discov Today. 2011; 16(23–24):1037–43. https://doi.org/10.1016/j.drudis.2011.09.007 PMID: 21945861.
5. Bunnage ME, Gilbert AM, Jones LH, Hett EC. Know your target, know your molecule. Nat Chem Biol. 2015; 11(6):368–72. https://doi.org/10.1038/nchembio.1813 PMID: 25978985.
6. Rouillard AD, Wang Z, Ma'ayan A. Abstraction for data integration: Fusing mammalian molecular, cellular and phenotype big datasets for better knowledge extraction. Comput Biol Chem. 2015; 59 Pt B:123–38. https://doi.org/10.1016/j.compbiolchem.2015.08.005 PMID: 26297300.
7. Rigden DJ, Fernandez-Suarez XM, Galperin MY. The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. Nucleic Acids Res. 2016; 44(D1):D1–6. https://doi.org/10.1093/nar/gkv1356 PMID: 26740669; PubMed Central PMCID: PMCPMC4702933.
8. Abi Hussein H, Geneix C, Petitjean M, Borrel A, Flatters D, Camproux AC. Global vision of druggability issues: applications and perspectives. Drug Discov Today. 2017; 22(2):404–15. https://doi.org/10.1016/j.drudis.2016.11.021 PMID: 27939283.
9. Fauman EB, Rai BK, Huang ES. Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. Curr Opin Chem Biol. 2011; 15(4):463–8. https://doi.org/10.1016/j.cbpa.2011.05.020 PMID: 21704549.
10. Perez-Lopez AR, Szalay KZ, Turei D, Modos D, Lenti K, Korcsmaros T, et al. Targets of drugs are generally, and targets of drugs having side effects are specifically good spreaders of human interactome perturbations. Sci Rep. 2015; 5:10182. https://doi.org/10.1038/srep10182 PMID: 25960144; PubMed Central PMCID: PMCPMC4426692.
11. Iwata H, Mizutani S, Tabei Y, Kotera M, Goto S, Yamanishi Y. Inferring protein domains associated with drug side effects based on drug-target interaction network. BMC Syst Biol. 2013; 7(Suppl 6):S18. https://doi.org/10.1186/1752-0509-7-S6-S18 PMID: 24565527
12. Wang X, Thijssen B, Yu H. Target essentiality and centrality characterize drug side effects. PLoS Comput Biol. 2013; 9(7):e1003119. https://doi.org/10.1371/journal.pcbi.1003119 PMID: 23874169
13. Kotlyar M, Fortney K, Jurisica I. Network-based characterization of drug-regulated genes, drug targets, and toxicity. Methods. 2012; 57(4):499–507. https://doi.org/10.1016/j.ymeth.2012.06.003 PMID: 22749929.
14. Kandoi G, Acencio ML, Lemke N. Prediction of Druggable Proteins Using Machine Learning and Systems Biology: A Mini-Review. Front Physiol. 2015; 6:366. https://doi.org/10.3389/fphys.2015.00366 PMID: 26696900; PubMed Central PMCID: PMCPMC4672042.

15. Costa PR, Acencio ML, Lemke N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. BMC Genomics. 2010; 11(Suppl 5):S9. https://doi.org/10.1186/1471-2164-11-S5-S9 PMID: 21210975

16. Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics. 2009; 25(4):451–7. https://doi.org/10.1093/bioinformatics/btp002 PMID: 19164304.

17. Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. BMC Bioinformatics. 2007; 8:353. https://doi.org/10.1186/1471-2105-8-353 PMID: 17883836; PubMed Central PMCID: PMCPMC2082046.

18. Li ZC, Zhong WQ, Liu ZQ, Huang MH, Xie Y, Dai Z, et al. Large-scale identification of potential drug targets based on the topological features of human protein-protein interaction network. Anal Chim Acta. 2015; 871:18–27. https://doi.org/10.1016/j.aca.2015.02.032 PMID: 25847157.

19. Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome Med. 2014; 6(7):57. https://doi.org/10.1186/s13073-014-0057-7 PMID: 25165489

20. Zhu M, Gao L, Li X, Liu Z, Xu C, Yan Y, et al. The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network. J Drug Target. 2009; 17(7):524–32. https://doi.org/10.1080/10611860903046610 PMID: 19530902.

21. Yao L, Rzhetsky A. Quantitative systems-level determinants of human genes targeted by successful drugs. Genome Res. 2008; 18(2):206–13. https://doi.org/10.1101/gr.6888208 PMID: 18083776; PubMed Central PMCID: PMCPMC2203618.

22. Mora A, Donaldson IM. Effects of protein interaction data integration, representation and reliability on the use of network properties for drug target prediction. BMC Bioinformatics. 2012; 12(13):294. https://doi.org/10.1186/1471-2105-13-294

23. Mitsopoulos C, Schierz AC, Workman P, Al-Lazikani B. Distinctive Behaviors of Druggable Proteins in Cellular Networks. PLoS Comput Biol. 2015; 11(12):e1004597. https://doi.org/10.1371/journal.pcbi.1004597 PMID: 26699810; PubMed Central PMCID: PMCPMC4689399.

24. Xu H, Xu H, Lin M, Wang W, Li Z, Huang J, et al. Learning the drug target-likeness of a protein. Proteomics. 2007; 7(23):4255–63. https://doi.org/10.1002/pmic.200700062 PMID: 17963289.

25. Bull SC, Doig AJ. Properties of protein drug target classes. PLoS One. 2015; 10(3):e0117955. https://doi.org/10.1371/journal.pone.0117955 PMID: 25822509; PubMed Central PMCID: PMCPMC4379170.

26. Li S, Yu X, Zou C, Gong J, Liu X, Li H. Are Topological Properties of Drug Targets Based on Protein-Protein Interaction Network Ready to Predict Potential Drug Targets? Comb Chem High Throughput Screen. 2016; 19(2):109–20. https://doi.org/10.2174/1386207319666151110122145 PMID: 26552443

27. Ghiassian SD, Menche J, Barabasi AL. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015; 11(4):e1004120. https://doi.org/10.1371/journal.pcbi.1004120 PMID: 25853560; PubMed Central PMCID: PMCPMC4390154.

28. Yang P, Li X, Chua HN, Kwoh CK, Ng SK. Ensemble positive unlabeled learning for disease gene identification. PLoS One. 2014; 9(5):e97079. https://doi.org/10.1371/journal.pone.0097079 PMID: 24816822

29. Carson MB, Lu H. Network-based prediction and knowledge mining of disease genes. BMC Med Genomics. 2015; 8(Suppl 2):S9. https://doi.org/10.1186/1755-8794-8-S2-S9 PMID: 26043920

30. Zhu C, Wu C, Aronow BJ, Jegga AG. Computational approaches for human disease gene prediction and ranking. Adv Exp Med Biol. 2014; 799:69–84. https://doi.org/10.1007/978-1-4614-8778-4_4 PMID: 24292962

31. Piro RM, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J. 2012; 279(5):678–96. https://doi.org/10.1111/j.1742-4658.2012.08471.x PMID: 22221742.

32. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet. 2012; 13(8):523–36. https://doi.org/10.1038/nrg3253 PMID: 22751426.

33. Emig D, Ivliev A, Pustovalova O, Lancashire L, Bureeva S, Nikolsky Y, et al. Drug target prediction and repositioning using an integrated network-based approach. PLoS One. 2013; 8(4):e60618. https://doi.org/10.1371/journal.pone.0060618 PMID: 23593264; PubMed Central PMCID: PMCPMC3617101.

34. Sun J, Zhu K, Zheng W, Xu H. A comparative study of disease genes and drug targets in the human protein interactome. BMC Bioinformatics. 2015; 16(Suppl 5):S1. https://doi.org/10.1186/1471-2105-16-S5-S1 PMID: 25861037

35. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS

Comput Biol. 2010; 6(2):e1000662. https://doi.org/10.1371/journal.pcbi.1000662 PMID: 20140234; PubMed Central PMCID: PMCPMC2816673.

36. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015; 47(8):856–60. https://doi.org/10.1038/ng.3314 PMID: 26121088.

37. Heinemann F, Huber T, Meisel C, Bundschus M, Leser U. Reflection of successful anticancer drug development processes in the literature. Drug Discov Today. 2016; 21(11):1740–4. https://doi.org/10.1016/j.drudis.2016.07.008 PMID: 27443674.

38. Pharmaprojects [Internet]. 2017. Available from: https://pharmaintelligence.informa.com/products-and-services/data-and-analysis/pharmaprojects.

39. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016; 2016. https://doi.org/10.1093/database/baw100 PMID: 27374120; PubMed Central PMCID: PMCPMC4930834.

40. Ernst MD. Permutation Methods: A Basis for Exact Inference. Statistical Science. 2004; 19(4):676–85. https://doi.org/10.1214/088342304000000396

41. Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol. 2010; 9:Article39. https://doi.org/10.2202/1544-6115.1585 PMID: 21044043.

42. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics. 2001; 29(4):1165–88.

43. Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. Nucleic acids research. 2013; 41(Database issue):D996–D1008. https://doi.org/10.1093/nar/gks1042 PMID: 23193282; PubMed Central PMCID: PMCPMC3531093.

44. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012; 489(7416):391–9. https://doi.org/10.1038/nature11405 PMID: 22996553; PubMed Central PMCID: PMCPMC4243026.

45. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature. 2007; 445(7124):168–76. https://doi.org/10.1038/nature05453 PMID: 17151600.

46. Wu C, MacLeod I, Su AI. BioGPS and MyGene. info: organizing online, gene-centric information. Nucleic acids research. 2012:gks1114.

47. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, et al. Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99(7):4465–70. https://doi.org/10.1073/pnas.012025199 PMID: 11904358; PubMed Central PMCID: PMCPMC123671.

48. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(16):6062–7. https://doi.org/10.1073/pnas.0400782101 PMID: 15075390; PubMed Central PMCID: PMCPMC395923.

49. Consortium G. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 2013; 45(6):580–5. https://doi.org/10.1038/ng.2653 PMID: 23715323.

50. Consortium G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348(6235):648–60. https://doi.org/10.1126/science.1262110 PMID: 25954001

51. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015; 347(6220):1260419. https://doi.org/10.1126/science.1260419 PMID: 25613900.

52. Santos A, Tsafou K, Stolte C, Pletscher-Frankild S, O'Donoghue SI, Jensen LJ. Comprehensive comparison of large-scale tissue expression datasets. PeerJ. 2015; 3:e1054. https://doi.org/10.7717/peerj.1054 PMID: 26157623; PubMed Central PMCID: PMCPMC4493645.

53. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman and Hall; 1991.

54. Calmettes G, Drummond GB, Vowler SL. Making do with what we have: use your bootstraps. J Physiol. 2012; 590(15):3403–6. https://doi.org/10.1113/jphysiol.2012.239376 PMID: 22855048; PubMed Central PMCID: PMCPMC3547254.

55. Jaffe AE, Storey JD, Ji H, Leek JT. Gene set bagging for estimating the probability a statistically significant result will replicate. BMC Bioinformatics. 2013; 14:360. https://doi.org/10.1186/1471-2105-14-360 PMID: 24330332

56. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015; 43(Database issue):D1079–85. https://doi.org/10.1093/nar/gku1071 PMID: 25361968; PubMed Central PMCID: PMCPMC4383909.

57. Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. Am J Hum Genet. 2012; 91 (2):215–23. https://doi.org/10.1016/j.ajhg.2012.06.004 PMID: 22818855; PubMed Central PMCID: PMCPMC3415546.

58. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006; 7:91. https://doi.org/10.1186/1471-2105-7-91 PMID: 16504092; PubMed Central PMCID: PMCPMC1397873.

59. Scott DW. Multivariate Density Estimation: Theory, Practice, and Visualization: John Wiley and Sons, Inc.; 1992.

60. Kumar V, Sanseau P, Simola DF, Hurle MR, Agarwal P. Systematic Analysis of Drug Targets Confirms Expression in Disease-Relevant Tissues. Sci Rep. 2016; 6:36205. https://doi.org/10.1038/srep36205 PMID: 27824084; PubMed Central PMCID: PMCPMC5099936.

61. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci U S A. 2008; 105(52):20870–5. https://doi.org/10.1073/pnas.0810772105 PMID: 19104045; PubMed Central PMCID: PMCPMC2606902.

62. Magger O, Waldman YY, Ruppin E, Sharan R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. PLoS Comput Biol. 2012; 8(9):e1002690. https://doi.org/10.1371/journal.pcbi.1002690 PMID: 23028288; PubMed Central PMCID: PMCPMC3459874.

63. Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet. 2002; 359 (9302):248–52. https://doi.org/10.1016/s0140-6736(02)07451-2

64. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Selection bias and information bias in clinical research. Nephron Clin Pract. 2010; 115(2):c94–9. https://doi.org/10.1159/000312871 PMID: 20407272.

65. Groth D, Hartmann S, Klie S, Selbig J. Principal components analysis. Methods Mol Biol. 2013; 930:527–47. https://doi.org/10.1007/978-1-62703-059-5_22 PMID: 23086856.

66. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006; 313(5786):504–7. https://doi.org/10.1126/science.1127647 PMID: 16873662.

67. Rubingh CM, Bijlsma S, Derks EP, Bobeldijk I, Verheij ER, Kochhar S, et al. Assessing the performance of statistical validation tools for megavariate metabolomics data. Metabolomics. 2006; 2(2):53–61. https://doi.org/10.1007/s11306-006-0022-6 PMID: 24489531; PubMed Central PMCID: PMCPMC3906710.

68. Cawley GC, Talbot NLC. On overfitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research. 2010; 11:2079–107.

69. Hurle MR, Nelson MR, Agarwal P, Cardon LR. Trial watch: Impact of genetically supported target selection on R&D productivity. Nature reviews Drug discovery. 2016; 15(9):596–7. https://doi.org/10.1038/nrd.2016.164 PMID: 27573226.

70. Brown GR, Hem V, Katz KS, Ovetsky M, Wallin C, Ermolaeva O, et al. Gene: a gene-centered information resource at NCBI. Nucleic Acids Res. 2015; 43(Database issue):D36–42. https://doi.org/10.1093/nar/gku1055 PMID: 25355515; PubMed Central PMCID: PMCPMC4383897.

71. Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. J Am Med Inform Assoc. 2001; 8(4):317–23. PMID: 11418538

72. Spearman C. The Proof and Measurement of Association between Two Things. American Journal of Psychology. 1904; 15(1):72–101. https://doi.org/10.2307/1412159

73. Frades I, Matthiesen R. Overview on techniques in cluster analysis. Methods Mol Biol. 2010; 593:81–107. https://doi.org/10.1007/978-1-60327-194-3_5 PMID: 19957146

74. Deshpande R, Vandersluis B, Myers CL. Comparison of profile similarity measures for genetic interaction networks. PLoS One. 2013; 8(7):e68664. https://doi.org/10.1371/journal.pone.0068664 PMID: 23874711; PubMed Central PMCID: PMCPMC3707826.

75. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32.