Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects

Phuong A. Nguyen[1], Aimee M. Deaton[1], Paul Nioi[1], Lucas D. Ward[1*]

[1]Amgen, Inc., Cambridge, MA, USA
*To whom correspondence should be addressed. Email: luke.ward@amgen.com

ABSTRACT

Biomedical scientists face major challenges in developing novel drugs for unmet medical needs. Only a small fraction of early drug programs progress to the market, due to safety and efficacy failures, despite extensive efforts to predict drug and target safety as early as possible using a variety of assays in vitro and in preclinical species. In principle, characterizing the effect of natural variation in the genes encoding drug targets should present a powerful alternate approach to predict not only whether a protein will be an effective drug target, but also whether a protein will be an inherently safe drug target, while avoiding the challenges of translating biology from experiments in non-human species. We have embarked on a retrospective analysis, demonstrating for the first time a statistical link between the organ systems involved in genetic syndromes of drug target genes and the organ systems in which side effects are observed clinically. Across 1,819 drugs and 21 organ system phenotype categories analyzed, drug side effects are more likely to occur in organ systems where there is genetic evidence of a link between the drug target and a phenotype involving that organ system, compared to when there is no such genetic evidence (30.0% vs 19.2%; OR = 1.80). Conversely, we find that having genetic evidence of a Mendelian syndrome involving a drug target in which a certain organ system is unaffected decreases the likelihood that side effects will manifest in that organ system, relative to having no informative syndrome (18.5% vs 20.2%; OR = 0.89). We find a relationship between genetics and side effects even when controlling for known confounders such as drug delivery route and indication. We highlight examples where genetics of drug targets could have anticipated side effects observed during clinical trials. This result suggests that human genetic data should be routinely used to predict potential safety issues associated with novel drug targets. This may lead to selection of better targets, appropriate monitoring of putative side effects early in development, reduction of the use of preclinical animal experiments, and ultimately increased success of molecules. Furthermore, deeply phenotyping human knockouts will be critically important to understand the full spectrum of effects that a new drug may elicit.

INTRODUCTION

Approximately 90% of drug candidates fail to progress through clinical trials because of issues with safety or efficacy[1-4] and this problem is magnified for novel targets[5-7]. Safety is an especially big hurdle for drug development because molecules' safety must be tested in preclinical species before reaching humans, and an unacceptable side-effect profile may not become apparent until clinical testing, because of the poor translatability from animal studies[8,9]. As a result, it has been estimated that although the median cost of developing a new drug from Phase I to approval is around $250 million, the total R&D cost per new drug is over $2.5 billion[10]. Therefore it is critical to develop better preclinical assessments for choosing only the safest drug targets and ideal molecules for clinical testing, and even modest improvements in preclinical predictions of toxicity can be massively valuable to drug development success[9].

There are many factors that contribute towards the safety and efficacy profile of a drug, and many of these have been studied through retrospective analyses to improve drug design. These factors include chemical properties of the drug that affect its pharmacokinetics and pharmacodynamics, biological properties of the target such as its expression pattern, the activity profile of the drug against off-target proteins, and genetic differences between patients that modulate inter-individual differences in response [11-22]. While all of these factors can contribute to the safety profile of a drug, the physiological role of the intended target is the most obvious and unavoidable consideration. While some drug targets have very narrow and specific roles in certain diseases, others can have broader biological roles – meaning that when they are drugged, there can be numerous unintended effects. One way of understanding biological roles of the target has been though human genetics, with the idea that phenotypes arising from natural variation in the gene encoding a protein will predict the phenotypes that

would result from drugging that protein. Family studies of rare Mendelian syndromes or genome-wide association studies (GWAS) of common diseases can both serve as "natural experiments" that point to new targets to pursue (with PCSK9 being a celebrated example)[4,23-25]. Retrospective analysis has shown that this genetic validation method results in finding efficacious targets and indications that are more likely to succeed through development and reach patients[26]. However, no such retrospective analysis has been published linking the genetic phenotypes of drug targets – which may include additional, pleiotropic phenotypes besides the therapeutic indication – to corresponding unintended side effects.

To explore this relationships between drug targets' genetic associations and drugs' side effects, we compiled a set of drugs for which data are available for both (a) the drugs' intended human target proteins and (b) the side effects observed during clinical trials of that drug. We then compared those side effects to the phenotypes associated with mutations in the genes encoding those target proteins. Thus, we investigated the ability of human genetics to predict the adverse events and side effects that will arise when drugging given target proteins.

RESULTS

Enrichment analysis

We merged data from several databases to compare genetic phenotypes with side effects (**Figure 1**). To examine the ability of genetics to predict drug treatment related side effects we first needed a large set of standardized side effect information. We obtained drug, indication, and side effect information from the Cortellis Clinical API and Cortellis Drug Design API (Clarivate Analytics, Inc.) for 31,194 clinical trials, all of which had at least one clinical side effect recorded in a controlled vocabulary of phenotype terms. We excluded oncology trials because of cancer drugs' inherent cytotoxicity and difference in acceptable side effect profiles[27], we excluded combination therapies to reduce complexity in our analysis, and we limited our analysis to small-molecule and biologic drugs. We also excluded very commonly-observed side effects such as headache (see Methods) to increase the sensitivity of our analysis. We then identified the accepted human protein targets of these drugs using the union of three databases: Drugbank[28-32], Pharmaprojects (Informa PLC), and a recently published curated map of drugs' molecular "efficacy targets." [32] A total of 1,819 of the drugs had at least one target annotated by one of these databases, and there were 1,046 unique targets pursued. For the genes encoding these drug targets, we obtained associated phenotypes from two sources: information from genome-wide association studies (GWAS) collated by the STOPGAP database[33], which uses coding and noncoding variant annotations to assign target genes to GWAS SNPs; and Mendelian information collated by the Human Phenotype Ontology (HPO)[34], specifically the subset of HPO phenotypes from Online Mendelian Inheritance in Man (OMIM)[35] which is mapped to controlled phenotype terms. A total of 1,394 of the 1,819 drugs had at least one target with genetic phenotype information.

In order to examine whether the genetic phenotypes of these drug targets matched the drug's side effects for each drug, we mapped its indication(s), side effect(s), and targets' genetic associations to 21 phenotype categories using the MedDRA ontology at the System Organ Class (SOC) level. We classified each of these drug-organ system pairs as (a) whether the drug has ever been pursued for a therapeutic indication involving that organ system, (b) whether that drug has ever elicited a clinical side effect involving that organ system, and (c) whether that drugs' targets have genetic phenotype associations involving that organ system (**Figure 1**; see **Table S1** for full dataset).

We found a striking relationship between drug target genetics and the organ systems in which side effects were observed (**Figure 2**). Across all 38,199 drug-side effect combinations, the existence of a genetic association between a drug's target(s) and phenotypes in a given organ system increases the probability that a side effect will be observed in that organ system during clinical trials from 19.2% to 30.0% (OR = 1.80; 95% CI = 1.71-1.90, Fisher's exact $P = 1.7 \times 10^{-94}$). This association could be largely driven by the correspondence between genetics and drug indications; side effects may tend to occur or be monitored more closely in the organ system of drug action, or side effects may be "exaggerated pharmacology" effects directly related to the therapeutic effect. To exclude this confounding effect, we performed the same enrichment test while removing from our analysis 4,710 drug-organ system pairs where the organ is the intended site of therapeutic efficacy for one of the drug's pursued indications, and found that genetic associations increase the probability of a side effect in the corresponding "off-indication" organ system from 15.6% to 22.4% (OR = 1.55; 95% CI = 1.45-1.66; Fisher's exact $P = 1.70 \times 10^{-36}$). This decreased off-indication effect confirms that some, but not all, of the genetics-side effect relationship is driven either by aspects of the intended therapeutic biology of the drug, or by other effects in the same organ system. Looking at individual organ systems (**Table 1**), we see that all statistically-significant effects are in the direction of enrichment, and that the strength of the enrichment detected is highly variable; for example, while gastrointestinal and

cardiac side effects are assigned to a similar number of drugs and genes across our data set, we find a strong signal of enriched off-indication cardiac side effects with target genetics (OR = 2.18, $P = 1.7 \times 10^{-94}$) and no significant enrichment for gastrointestinal side effects (OR = 1.12, $P = 0.37$). These differences may be due to the relative contribution of on- and off-target effects to these organ systems, or to the importance of modality to these organ systems; gastrointestinal side effects, for example, may be more often mediated by the absorption of a drug than its effect on its target.

We then explored whether there was any predictive value in the *lack* of a genetic association. When the gene encoding a drug target does not correlate with a specific phenotype, this could be for two reasons: either perturbing the gene causes exclusively other phenotypes, or there is no information relating to perturbation of the gene (which could be due to a lack of variation, or due to a lack of detectable effects when the gene is mutated.) We were interested in the difference between these two scenarios: if a drug target gene has been correlated with some phenotypes but not others, should drug developers accept this as evidence that those non-associated phenotypes are indeed less likely to manifest when the gene product is drugged? To test this difference we compared drug-side effect combinations where there was no genetic support, and split into two sets: those for which there was some genetic information for all of the targets of the drug, and those for which there were targets with no genetic associations and potentially missing information. As expected, a drug's targets having genetic support for other phenotypes – but not the phenotype being interrogated – is associated with a lower rate of side effects being observed in the interrogated phenotype, compared to a situation where there is a lack of genetic information for some or all of a drug's targets (OR = 0.91, $P = 0.0016$; **Figure 2**). Interestingly, this effect is stronger when considering only Mendelian syndromes as the source of genetic information (OR = 0.89, $P = 0.00017$); since Mendelian syndromes are generally more severe perturbations of the target, a lack of a phenotype finding in people carrying such a severe genetic effect in the target may represent a more confident "de-risking" of a target than GWAS or PheWAS allows.

Regression modeling

While a causal link between target genetics and side effects is biologically plausible – genetic and pharmacological perturbation of a protein should lead to similar phenotypes, and this logic is the basis of genetic validation of drug targets[24,25] – there are many potential confounders that could instead be explaining the enrichment we observe. For example, certain kinds of side effects may be correlated with certain drug indications simply because of the patient population treated, and indications will correlate with target genetics due to rational drug discovery and known biology. Indeed, we see that this is a strong correlation within our dataset (**Figure S1**); e.g., drugs with blood-related indications are nearly ten times more likely to have targets with Mendelian genetic syndromes involving blood. Other confounders include variables that have been previously linked to side effects: properties of the protein such as its cross-tissue expression pattern[15], properties of the gene such as evolutionary constraint[36] and corresponding intolerance of population genetic variation[26], or properties of the drug such as modality or delivery route. To explore the relationship of genetics and side effects with these potential confounders, we gathered additional drug and gene information and built a data set for regression modeling, to predict side effects in each organ system as dependent variables. For independent variables, in addition to the target genetic phenotypes, indications, and modality used for the enrichment analysis, we obtained drug delivery route information, expression breadth information, and population genetic constraint. We encoded all independent and dependent variables as binary values (**Table S1**) and built a logistic regression model for each of 18 side effect classes. Models used all of the independent variables described above including the source of genetic support (GWAS or Mendelian) for the side effect, to predict the presence or absence of the side effect class for each drug. We took two approaches for each side effect model: a multivariate logistic regression model that included all variables, and a machine learning logistic regression model with feature selection. We also ran these models two ways: using the complete dataset, and using a dataset that excluded "on-indication" side effects, as in the previous enrichment analysis.

Strikingly we found consistent, positive coefficients for genetics in models where genetics is significant in predicting drug side effects (**Figure 3**). This was true even when modeling other predictive variables and excluding potential exaggerated pharmacology (i.e. on-indication side effects). Mendelian and/or GWAS genetics was a significant predictor in the overall analysis for six of the side effect regression models, with four of these (metabolism, heart, endocrine, and reproductive) surpassing a conservative threshold corrected for the number of side effect classes tested ($p < 0.05/18$) (**Figures S2-S3**).

Splitting genetic data into Mendelian and GWAS evidence reduced our power to detect a joint effect on predictive power, but allowed us to ask about their individual contribution. We find that the only side effect classes for which GWAS was independently predictive (heart, mental, and reproductive) were also predictable from Mendelian genetics. This observation is consistent with the strength of the overall enrichment being similar when only focusing on Mendelian genetics (**Figure 2**).

Cross validation

Although the machine learning models perform cross-validation to ensure that the features selected add significant predictive power, there are two potential drawbacks that could lead to inaccurate effect and significance estimates. First, the substantial collinearity between our predictors (e.g. the confounding of indication with genetic phenotypes) may have led to genetics being selected as a predictor even though the causal factor was in fact indication. Second, different drugs in the model often share target genes, resulting in shared genetic information between the training and test sets while the machine learning models are testing new features to add to the model. To address both of these concerns, we implemented a custom leave-one-target-set-out cross-validation procedure to explicitly test the predictive power of genetics (measured as the area under the receiver operating characteristic curve, AUC) in side effect models, when the model has been trained on all other available variables and when the tested genetic data has never been seen in the training set. We use this cross-validation procedure to test the contribution of genetics to side effect prediction (measured as cross-validation AUC) relative to two baselines: the AUC of a model including no genetic information, and the AUC distribution of models with simulated data with permuted gene-phenotype relationships. We applied these tests to the six side effect models where either GWAS or Mendelian genetics was found to be significant (at $p < 0.05$). Genetics improved the cross-validation AUC (relative to simulated genetics and relative to no genetics) in five of the six models (**Figure 4**). The absolute increase in AUC is slight but statistically significant, indicating that many of the confounding factors are predictive of side effects, either because they are causal or because they are correlated with genetics. For example, the cross-validation AUC of the full model predicting cardiac side effects is 0.70; this drops to 0.68 when omitting genetics and never surpasses 0.70 in 1,000 simulations using permuted genetics, evidence that genetics is indeed adding predictive information.

DISCUSSION

In this study we show that human genetic studies of drug target proteins are able to predict not only therapeutic efficacy of drugs, but also can provide evidence about the likelihood of side effects. Although this approach has been proposed and used to make predictions about individual drug targets[37], it had not been systematically validated. This finding has a number of applications for drug discovery and could ultimately be used to help make safer therapeutics.

First, examining the human genetic associations of potential drug targets early in the drug discovery process could help us more fully understand target biology and select safer targets. Choosing the safest targets at the beginning of drug development should help to make the process more efficient and reduce safety failures that could have been predicted. Previous retrospective analyses of drug development have focused on the relationship between a drugs' chemical properties and safety and efficacy[19,20]. Another class of methods attempts to predict the set of off-target proteins that interact with a drug and the resulting side effects[11]. However, such a focus on drug properties and interactions may not help to improve safety if drug developers are not pursuing safe targets. Altogether, our results suggests a practical framework for evaluating a new drug target and for understanding the effects that are seen during clinical development. When making the decision to pursue a novel target, Mendelian syndromes, GWAS associations, and PheWAS databases should be consulted to interrogate the gene(s) encoding the target. If a the novel target has been reported as a Mendelian disease gene, all aspects of that disease should be investigated, and researchers should seek to understand whether the genetic disease has aspects unrelated to the intended therapeutic action of the drug. If so, these should be investigated as potential liabilities for the drug; if not, the lack of concerning phenotypes can be used as evidence in favor of target safety. Searching GWAS and PheWAS associations is complicated by the problem of interpreting variants; however, if a variant in a novel drug target gene has an association with the intended therapeutic effect, then this variant can be presumed as functional (in light of other supporting functional genomics evidence) and examined for pleiotropic, non-therapeutic associations. These can then be considered potential "target liabilities."

Secondly, the genetics of drug targets can continue to inform drug development even after the decision has been made to pursue a target. As drug experiments in preclinical animals and clinical trials in humans proceed, human genetic information can guide researchers to monitor and design assays to test for side effects of most. When side effects do manifest in preclinical species or in clinical trial participants, human genetics of the target can be used to build hypotheses of whether the side effects are target-mediated phenotypes intrinsic to the therapeutic mechanism, or off-target effects that could be engineered out of the drug.

Our analysis revealed a number of examples where side effects could have been predicted though examination of target genetics. These examples include:

Basilizimab, *IL2RA*, and diabetes. New-onset diabetes after transplantation (NODAT) has been noted as a side effect in trials of basiliximab, an IL-2 receptor antibody indicated to prevent transplant rejection; this effect has been confirmed in an observational trial[38] and mouse studies[39]. The *IL2RA* gene is implicated in Type 1 diabetes by GWAS[40], and heterozygous truncating mutations in the *IL2RA* gene cause a Mendelian immunodeficiency syndrome sometimes associated with diabetes[41]. *IL2RA*, encoding the IL-2Rα (CD25) subunit of the trimeric IL-2 receptor, is responsible for high affinity of the receptor for IL-2, and is expressed on activated T lymphocytes including the chronically activated regulatory T (T reg) subset. Imbalance between T reg cells and T effector cells in the pancreatic islet has been implicated in breakdown of self-tolerance and development of type 1 diabetes in a mouse model; IL-2 administration is protective in this model by promoting T reg survival[42].

AChE inhibitors, *ACHE*, and bradycardia. Heart rate dysregulation manifesting as bradycardia has been noted in several trials of acetylcholinesterase (AchE) inhibitors for Alzheimer's disease, and this observation has been discussed in pharmacovigilance reports[43,44]; the gene encoding AchE, *ACHE*, has been linked to tachycardia by GWAS[45,46]. Acetylcholine is an important neurotransmitter in the parasympathetic nervous system, which has a cardioinhibitory affect on heartbeat via cardiac M2 muscarinic receptors. AchE inhibition results in the neurotransmitter persisting at the synapse, prolonging the cardioinhibitory effect[47].

Esreboxetine, SLC6A2, and tachycardia. In trials of esreboxetine, a norepinephrine reuptake inhibitor being pursued for fibromyalgia, an increased heart rate has been noted in the treatment arm[48-50]. Esreboxetine acts on the norepinephrine transporter, encoded by the *SLC6A2* gene. Loss-of-function mutation in *SLC6A2* causes a Mendelian syndrome characterized by orthostatic intolerance with tachycardia, resulting from decreased reuptake of norepinephrine and higher plasma concentrations[51,52]. Norepinephrine has a wide range of effects in the sympathetic nervous system through signaling to adrenergic receptors throughout the body, such as increasing heart rate and blood pressure, but reuptake inhibition has been proposed to particularly influence its effect in the heart, because clearance in the narrower synaptic clefts in the heart is far more dependent on reuptake than enzymatic breakdown. This difference is reflected in the fact that *SLC6A2* loss of function affects heart rate disproportionately relative to blood pressure[51].

There are some limitations to our study. The major one is that we are limited by small and biased sample size, both in terms of genetic knowledge and drug trials. Clinical trial side effects are frequently not reported, and our drug trial information is limited to the subset (~12% of trials in the Cortellis Clinical database) that have both outcome data and controlled-vocabulary side effect annotations. We attempt to control for this uncertainty by discarding frequently-noted side effects and by modeling the indication (which will correlate with phenotypes particular to a patient segment). However, individual drug-side effect observations are low confidence and may not be treatment-related, and conclusions should only be drawn from global patterns. By necessity, our phenotype matching is coarse-grained because of the small sample size and will introduce false positive matches. We anticipate that cleaner, larger sets of clinical trial data would improve our analysis. In addition, there is a paucity of genetic information for some genes either because they are highly constrained or lack common SNPs interrogated in GWAS. Finally, there are confounding factors that make it impossible to accurately determine the effect of genetics independently of other factors such as indication and delivery route. The raw enrichments are likely over-estimates, while the results of the multivariate regression may be over- or under-estimates. The results from the cross-validation procedure are likely to be under-estimates, since this procedure controls for all other potential covariates before adding genetics to the model – therefore it is these cross-validation results that give us confidence that genetics indeed adds predictive value.

We have demonstrated that the "natural experiments" afforded by human genetics can be helpful for anticipating the full spectrum of phenotypes elicited by modulating a target pharmacologically - not only the therapeutic effect, but also side effects and adverse events mediated by pleiotropic biological roles of the target. An extension of our work is to understand how the genetics of a drug's "off-target" proteins also contribute to its side effect profile. Human genetics data are becoming increasingly rich, especially with comprehensive efforts to deep-phenotype complete human knockouts [53-55] and to perform phenome-wide association studies (PheWAS) across electronic medical records to detect pleiotropic effects of genes[37,56]. This analysis underscores the importance of comprehensively characterizing human knockouts and people affected by Mendelian syndromes, because in some cases their biology can help anticipate drug safety issues before they occur, while in other cases their lack of concerning phenotypes can help build conviction that certain proteins are intrinsically safe to drug. Our analysis suggests that this growing body of knowledge will aid in selecting not only more effective targets but also developing safer drugs at lower cost.

URLs

DrugBank, http://www.drugbank.ca
STITCH, http://stitch.embl.de
Cortellis Clinical API, https://www.cortellislabs.com/page/?api=api-CLI and Drug Design API, https://www.cortellislabs.com/page/?api=api-DD *(Note: Information reported in this article is derived from Cortellis for Clinical Trials Intelligence database and Drug Design APIs, which are produced and owned by Clarivate Analytics. Clarivate Analytics will not be liable for any inaccuracy in the information provided in this article or the way in which it is used by any reader in this article.)*
HPO, http://human-phenotype-ontology.github.io
UMLS and related software, https://www.nlm.nih.gov/research/umls/ *(Note: MedDRA® is the international medical terminology developed under the auspices of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). MedDRA® trademark is owned by IFPMA on behalf of ICH.)*
Citeline Pharmaprojects, https://pharmaintelligence.informa.com/
STOPGAP, http://stopgapwebapp.com:3838/SWApp/

ONLINE METHODS

Drug databases

The following pharmacology databases were used: DrugBank[28-31] (v5.0.6, r2017-04-01), STITCH[57] (v5.0, r2016-06-27), Citeline Pharmaprojects (Informa PLC., d2016-11-22), Cortellis Clinical API (Clarivate Analytics, Inc., d2017-03-31), and Cortellis Drug Design API (Clarivate Analytics, Inc., d2017-04-03; filtered for drugs that have been tested in humans, i.e. with highest phase of Phase 0 or higher).

To consolidate information about each drug to enter into our model, drugs from DrugBank, Cortellis Drug Design and Citeline Pharmaprojects were merged using an approach combining drug names and aliases/synonyms and CAS numbers (either original active ingredient or its various salt forms). Stereoisomers were also merged (e.g. baclofen and arbaclogen, tretinoin and isotretinoin). This process resulted in unique drug units that aimed to distinguish only active ingredients (but not different formulations or products from different companies; or different cellular expression system for protein-based therapeutics). Drug mixtures, i.e. those with more than one active ingredient, were removed from the analysis for simplicity. To each unique drug unit, intended targets were obtained from the union of DrugBank (targets annotated as having "known" pharmacological action), Citeline Pharmaprojects annotations, and a recent curated database of therapeutic efficacy targets of a subset of marketed drugs[32].

Adverse events reported from clinical trials were obtained from Cortellis Clinical API. Of a total of 264,735 trials on record, 92,397 trials had outcomes available. A subset of those trials (31,194) had adverse event terms extracted in computer-readable lists. We limited our analysis to this subset of trials and the corresponding drugs. Trials of combination therapies, oncology indications, and therapeutics that were not small molecules or proteins were also removed. Note that since we could not distinguish between lack of adverse events listed due to complete absence of adverse events or missing data, our dataset only included trials that had at least one adverse event listed as being higher than placebo. We found the frequency distribution of adverse events to be skewed towards a small number of side effects that were very common across different drugs (e.g., headache for 55% of drugs). Because these very common side effects may be less likely to reflect target-mediated effects, and in order to prevent them from overwhelming and biasing the downstream analysis, we excluded 27 of the most common side effects (defined as observed with at least 10% of drugs; **Table S2**), after which 1,179 side effect types remained.

Drug modality information was obtained from the 'TrialCategories' field of the Cortellis Clinical trial records (either "Biological" or "Small molecule"). Drug administration routes were obtained from the 'DevelopmentStatusSummary' field of the Cortellis Drug Design drug records, and were grouped into four categories: enteral, parenteral, topical, and other (**Table S4**). Drug disease indications were obtained from the Cortellis Clinical trial records and mapped to MedDRA SOC terms as described below.

Genetics databases

Genes involved in Mendelian traits were derived from the Human Phenotype Ontology (HPO)[58] (r2017-04-13). 3,404 genes were associated with 4,390 OMIM syndromes, each of which was described with a list of HPO phenotype terms.

Following the method of Nelson et al.[26], genome-wide association studies (GWAS) associated with each gene were obtained from STOPGAP [33], a database that systematically links published GWAS SNPs results to putative causal genes using a combination of linkage disequilibrium, functional genomics annotations, and variant effect predictions. Only the top-ranked gene (or genes, if tied for top rank) for each GWAS SNP was retained as a gene-phenotype association. Only associations with a genome-wide significant $p < 5 \times 10^{-8}$ were retained, and only associations derived from a publication in the EBI-NHGRI GWAS Catalog were retained. The final GWAS data set consisted of 4,265 genes associated with 431 Medical Subject Headings (MeSH) terms.

Phenotype mapping

To allow comparison between all phenotype databases (pharmacology and genetics), we used the Unified Medical Language System (UMLS) Metathesaurus (2015AB release), the MetaMap natural language processing (NLP) tool (MetaMap2016, r2016-01), and the UMLS-Interface software[59] to map phenotypes to the terms found in the Medical Dictionary for Regulatory Activities terminology (MedDRA). All phenotype terms were mapped to the most specific MedDRA term within the UMLS, which was in turn mapped to 26 MedDRA System Organ Class (SOC) terms (**Table S3**). When a single phenotype term was mapped to multiple SOC terms (e.g. 'kidney cancer' is mapped to both 'Neoplasm NOS' and 'Diseases, Urologic'), all SOC terms were retained. All mappings to the SOC terms 'Social circumstances' and 'Surgical and medical procedures' were excluded from analysis. All mappings to 'General disorders and administration site conditions', 'Injury, poisoning and procedural complications', and 'Investigation NOS' were further mapped to other, more organ system related, SOCs using the MedDRA terms one level below the SOCs ('High Level Group Terms').

Tissue expression and mutational constraint of targets.

Motivated by the observation of Gayvert *et al.*[15] and others that the general target tissue expression profile is associated with specific toxicities, we used expression breadth annotations from a cross-tissue analysis of the Human Protein Atlas [60]. Following this annotation, genes were put into four categories: "all" (corresponding to "expressed in all high" or "expressed in all low"), "mixed" (corresponding to "mixed high" and "mixed low"), "enriched" (corresponding to "tissue specific", "tissue enriched", or "group enriched), and "none detected". Motivated by the previous observation [26] that mutational constraint of targets is associated with drug success, we incorporated the annotation of Lek *et al.* [61] of Exome Aggregation Consortium variation data to define genes as constrained or unconstrained with a pLI score threshold of 0.9.

Multivariate regression and feature selection

To assess the correlation between the genetics of intended targets and the adverse event/side effect profile of a drug, we performed multivariate logistic regressions, both without feature selection (using the 'glm' R package) and with feature selection (using the lasso penalized maximum likelihood technique using the 'glmnet' R package[47,48]) (v2.0-10). Out of the 21 MedDRA SOC adverse event phenotype groups, we focused only on modeling phenotypes in the 18 groups with an sample size of at least 100 (~5%) drugs eliciting an adverse event in that phenotype group (AE). For each of these 18 phenotype groups, we built a logistic regression model with glm and glmnet using the following predictors: disease indications (20 MedDRA SOCs, excluding 'neoplasm'), drug modality ('biological' or 'small molecule', with the latter as the baseline), delivery routes from Cortellis Drug Design API, and genetic phenotypes of the drug's targets, encoded as follows: Mendelian genetics for each drug in each side effect model has three possible values: having genetic phenotypes matching the side effect, having genetic phenotypes mismatched to the side effect (there is a syndrome with phenotypes unrelated to the side effect), and having no Mendelian genetic information; and GWAS genetics for each drug in each side effect model has three possible values: having GWAS associations matching the AE, having GWAS associations mismatched to the AE, and having no GWAS hits linked to the target gene(s).

To choose the lambda parameter for the glmnet feature selection procedure, for each model, 10-fold cross-validation was run 100 times, and the final optimal lambda was selected as the average of mean lambda.min (i.e. lambda that resulted in the highest AUC) and the mean lamda.1se (i.e. largest lambda that resulted in an AUC one standard error away from the maximum AUC). The final model coefficients were derived from running 'glmnet' on the whole data set (1,819 drugs) with the optimal lambda.

Cross-validation procedure

To assess the predictive power of including genetics in a given side effect model, we performed a custom cross-validation procedure as follows: (1) Each unique *target set* (combination of target proteins for a given drug) among the list of drugs in the model was enumerated. (2) For each *target set*, the drugs sharing this target set were held out as tests. (3) The training set was all other drugs, excluding any drug that shared in its target set *any* of the targets contained in the test set, to avoid any sharing of genetic information between training and test set. (4) Logistic regression (without feature selection) was performed on the training set. (5) Side effect probabilities (the output of the model) were calculated for the drugs in the test set using the coefficients learned on the training set.

A cross-validation receiver-operating characteristic (ROC) curve was generated and its area under the curve (AUC) was calculated using the complete set of test predictions. For the side effect models tested, the cross-validation AUC was calculated: (1) With all predictors included in the model; (2) with genetics excluded from the model; and (3) with genetics randomized, such that within the set of drug targets, the assignment of genes to phenotype sets was permuted, maintaining the same fraction of genes with no genetic information and preserving the correlation between genetic phenotypes. This simulation was performed 1,000 times to generate a distribution of AUCs.

REFERENCES

1 Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature reviews. Drug discovery* **3**, 711-715, doi:10.1038/nrd1470 (2004).

2 Waring, M. J. *et al.* An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature reviews. Drug discovery* **14**, 475-486, doi:10.1038/nrd4609 (2015).

3 Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature reviews. Drug discovery* **13**, 419-431, doi:10.1038/nrd4309 (2014).

4 Plenge, R. M. Disciplined approach to drug discovery and early development. *Sci Transl Med* **8**, 349ps315, doi:10.1126/scitranslmed.aaf2608 (2016).

5 Knowles, J. & Gromo, G. A guide to drug discovery: Target selection in drug discovery. *Nature reviews. Drug discovery* **2**, 63-69, doi:10.1038/nrd986 (2003).

6 Lindsay, M. A. Target discovery. *Nature reviews. Drug discovery* **2**, 831-838 (2003).

7 Ma, P. & Zemmel, R. Value of novelty? *Nature reviews. Drug discovery* **1**, 571-572, doi:10.1038/nrd884 (2002).

8 Bailey, J., Thew, M. & Balls, M. An analysis of the use of animal models in predicting human toxicology and drug safety. *Altern Lab Anim* **42**, 181-199 (2014).

9 Scannell, J. W. & Bosley, J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. *PloS one* **11**, e0147215, doi:10.1371/journal.pone.0147215 (2016).

10 DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of health economics* **47**, 20-33, doi:10.1016/j.jhealeco.2016.01.012 (2016).

11 Zhou, H., Gao, M. & Skolnick, J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. *Scientific reports* **5**, 11090, doi:10.1038/srep11090 (2015).

12 Duran-Frigola, M. & Aloy, P. Analysis of chemical and biological features yields mechanistic insights into drug side effects. *Chem Biol* **20**, 594-603, doi:10.1016/j.chembiol.2013.03.017 (2013).

13 Duran-Frigola, M., Rossell, D. & Aloy, P. A chemo-centric view of human health and disease. *Nature communications* **5**, 5676, doi:10.1038/ncomms6676 (2014).

14 Juan-Blanco, T., Duran-Frigola, M. & Aloy, P. IntSide: a web server for the chemical and biological examination of drug side effects. *Bioinformatics* **31**, 612-613, doi:10.1093/bioinformatics/btu688 (2015).

15 Gayvert, K. M., Madhukar, N. S. & Elemento, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chem Biol* **23**, 1294-1301, doi:10.1016/j.chembiol.2016.07.023 (2016).

16 Garcia-Serna, R. & Mestres, J. Anticipating drug side effects by comparative pharmacology. *Expert Opin Drug Metab Toxicol* **6**, 1253-1263, doi:10.1517/17425255.2010.509343 (2010).

17 LaBute, M. X. *et al.* Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PloS one* **9**, e106298, doi:10.1371/journal.pone.0106298 (2014).

18 Lounkine, E. *et al.* Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361-367, doi:10.1038/nature11159 (2012).

19 Scheiber, J. *et al.* Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J Chem Inf Model* **49**, 308-317, doi:10.1021/ci800344p (2009).

20 Leeson, P. Drug discovery: Chemical beauty contest. *Nature* **481**, 455-456, doi:10.1038/481455a (2012).

21 Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**, 3-26 (2001).

22 Alfirevic, A. & Pirmohamed, M. Genomics of Adverse Drug Reactions. *Trends in pharmacological sciences* **38**, 100-109, doi:10.1016/j.tips.2016.11.003 (2017).

23 Ferrero, E., Dunham, I. & Sanseau, P. In silico prediction of novel therapeutic targets using gene-disease association data. *J Transl Med* **15**, 182, doi:10.1186/s12967-017-1285-6 (2017).

24 Kamb, A., Harper, S. & Stefansson, K. Human genetics as a foundation for innovative drug development. *Nature biotechnology* **31**, 975-978, doi:10.1038/nbt.2732 (2013).

25 Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nature reviews. Drug discovery* **12**, 581-594, doi:10.1038/nrd4051 (2013).

26 Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature genetics* **47**, 856-860, doi:10.1038/ng.3314 (2015).

27 Dorato, M. A. & Engelhardt, J. A. The no-observed-adverse-effect-level in drug safety evaluations: use, issues, and definition(s). *Regulatory toxicology and pharmacology : RTP* **42**, 265-274, doi:10.1016/j.yrtph.2005.05.004 (2005).

28 Knox, C. *et al.* DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **39**, D1035-1041, doi:10.1093/nar/gkq1126 (2011).

29 Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **42**, D1091-1097, doi:10.1093/nar/gkt1068 (2014).

30 Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research* **34**, D668-672, doi:10.1093/nar/gkj067 (2006).

31 Wishart, D. S. *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research* **36**, D901-906, doi:10.1093/nar/gkm958 (2008).

32 Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nature reviews. Drug discovery* **16**, 19-34, doi:10.1038/nrd.2016.230 (2017).

33 Shen, J., Song, K., Slater, A., Ferrero, E. & Nelson, M. R. STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. *Bioinformatics*, doi:10.1093/bioinformatics/btx274 (2017).

34 Robinson, P. N. *et al.* The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics* **83**, 610-615, doi:10.1016/j.ajhg.2008.09.017 (2008).

35 Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic acids research* **43**, D789-798, doi:10.1093/nar/gku1205 (2015).

36 Begum, T., Ghosh, T. C. & Basak, S. Systematic Analyses and Prediction of Human Drug Side Effect Associated Proteins from the Perspective of Protein Evolution. *Genome biology and evolution* **9**, 337-350, doi:10.1093/gbe/evw301 (2017).

37 Diogo, D. *et al.* Phenome-wide association studies (PheWAS) across large "real-world data" population cohorts support drug target validation. *bioRxiv*, doi:10.1101/218875 (2017).

38 Prasad, N. *et al.* Is basiliximab induction, a novel risk factor for new onset diabetes after transplantation for living donor renal allograft recipients? *Nephrology* **19**, 244-250, doi:10.1111/nep.12209 (2014).

39 Setoguchi, R., Hori, S., Takahashi, T. & Sakaguchi, S. Homeostatic maintenance of natural Foxp3(+) CD25(+) CD4(+) regulatory T cells by interleukin (IL)-2 and induction of autoimmune disease by IL-2 neutralization. *The Journal of experimental medicine* **201**, 723-735, doi:10.1084/jem.20041982 (2005).

40 Plagnol, V. *et al.* Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet* **7**, e1002216, doi:10.1371/journal.pgen.1002216 (2011).

41 Caudy, A. A., Reddy, S. T., Chatila, T., Atkinson, J. P. & Verbsky, J. W. CD25 deficiency causes an immune dysregulation, polyendocrinopathy, enteropathy, X-linked-like syndrome, and defective IL-10 expression from CD4 lymphocytes. *The Journal of allergy and clinical immunology* **119**, 482-487, doi:10.1016/j.jaci.2006.10.007 (2007).

42 Tang, Q. *et al.* Central role of defective interleukin-2 production in the triggering of islet autoimmune destruction. *Immunity* **28**, 687-697, doi:10.1016/j.immuni.2008.03.016 (2008).

43 Madhani, P., Hughes, J. & Ballard, C. G. Cholinesterase inhibitors and the heart in old age. *Advances in Psychiatric Treatment* **10**, 399-399, doi:10.1192/apt.10.5.399 (2004).

44 Cholinesterase inhibitors and cardiac arrhythmias. *Australian Adverse Drug Reactions Bulletin* **23**.

45 den Hoed, M. *et al.* Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nature genetics* **45**, 621-631, doi:10.1038/ng.2610 (2013).

46 Eppinga, R. N. *et al.* Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. *Nature genetics* **48**, 1557-1563, doi:10.1038/ng.3708 (2016).

47 Masuda, Y. Cardiac effect of cholinesterase inhibitors used in Alzheimer's disease--from basic research to bedside. *Current Alzheimer research* **1**, 315-321 (2004).

48 Arnold, L. M., Hirsch, I., Sanders, P., Ellis, A. & Hughes, B. Safety and efficacy of esreboxetine in patients with fibromyalgia: a fourteen-week, randomized, double-blind, placebo-controlled, multicenter clinical trial. *Arthritis and rheumatism* **64**, 2387-2397, doi:10.1002/art.34390 (2012).

49 Arnold, L. M., Chatamra, K., Hirsch, I. & Stoker, M. Safety and efficacy of esreboxetine in patients with fibromyalgia: An 8-week, multicenter, randomized, double-blind, placebo-controlled study. *Clinical therapeutics* **32**, 1618-1632, doi:10.1016/j.clinthera.2010.08.003 (2010).

50 Denolle, T., Pellizzoni, C., Jannuzzo, M. G. & Poggesi, I. Hemodynamic effects of reboxetine in healthy male volunteers. *Clinical pharmacology and therapeutics* **66**, 282-287, doi:10.1016/S0009-9236(99)70036-6 (1999).

51 Shannon, J. R. *et al.* Orthostatic intolerance and tachycardia associated with norepinephrine-transporter deficiency. *The New England journal of medicine* **342**, 541-549, doi:10.1056/NEJM200002243420803 (2000).

52 Paczkowski, F. A., Bonisch, H. & Bryan-Lluka, L. J. Pharmacological properties of the naturally occurring Ala(457)Pro variant of the human norepinephrine transporter. *Pharmacogenetics* **12**, 165-173 (2002).

53 Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nature genetics* **47**, 448-452, doi:10.1038/ng.3243 (2015).

54 Mullard, A. Calls grow to tap the gold mine of human genetic knockouts. *Nature reviews. Drug discovery* **16**, 515-518, doi:10.1038/nrd.2017.139 (2017).

55 Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235-239, doi:10.1038/nature22034 (2017).

56 Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205-1210, doi:10.1093/bioinformatics/btq126 (2010).

57 Szklarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic acids research* **44**, D380-384, doi:10.1093/nar/gkv1277 (2016).

58 Kohler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic acids research* **45**, D865-D876, doi:10.1093/nar/gkw1039 (2017).

59 McInnes, B. T., Pedersen, T. & Pakhomov, S. V. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. *AMIA Annu Symp Proc* **2009**, 431-435 (2009).

60 Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Molecular & cellular proteomics : MCP* **13**, 397-406, doi:10.1074/mcp.M113.035600 (2014).

61 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

MAIN FIGURES AND TABLES

Figure 1. Data sources and analysis workflow.

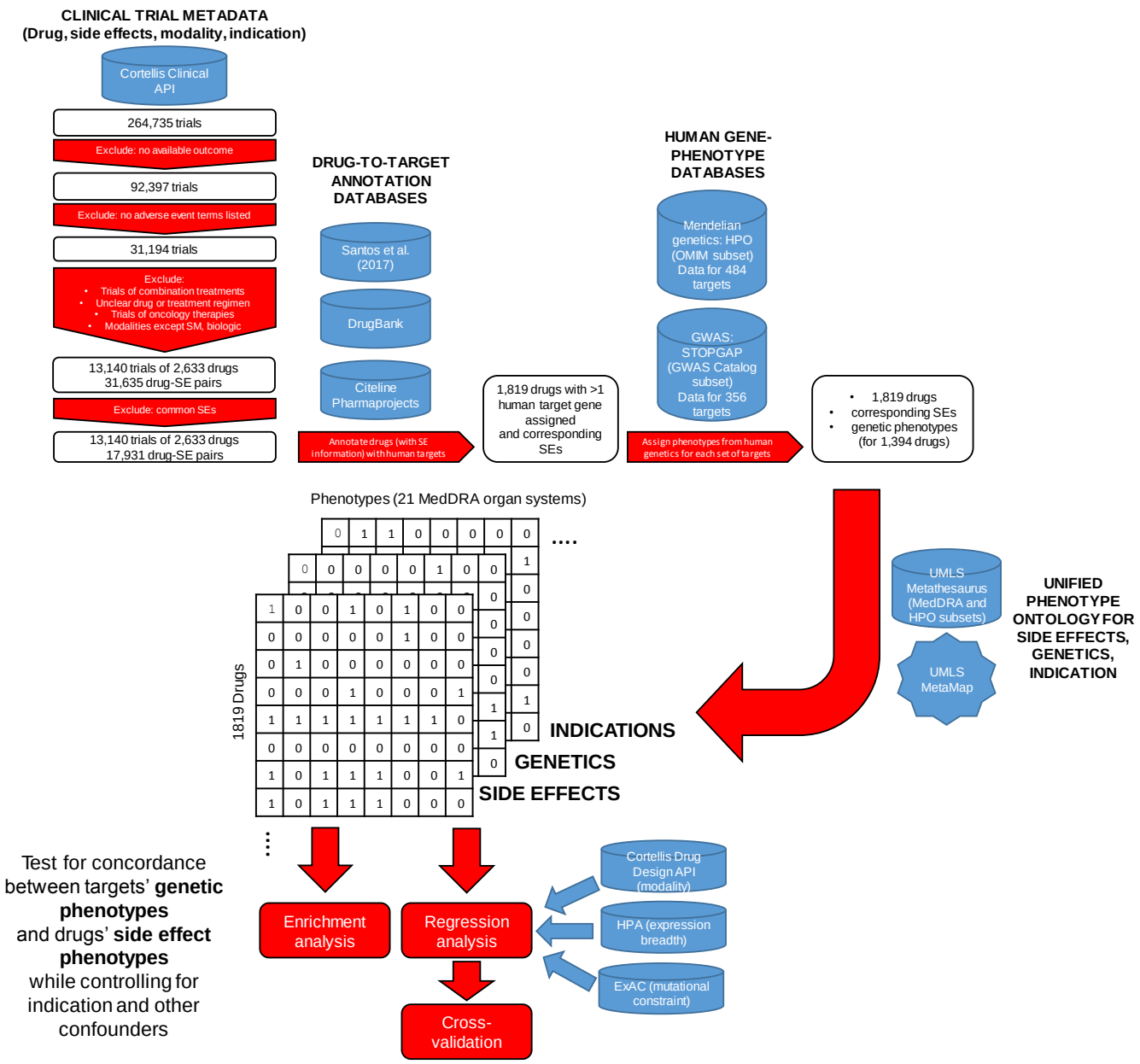

Figure 2. Comparison of the rate at which all possible drug-SE (side effect) combinations are manifested as side effects in the clinical trial data set, between drug-SE pairs with or without genetic support. OR, odds ratio; p, p-value from Fisher's exact test. The top two panels show analyses based on all 38,199 drug-SE combinations; the bottom two panels show analyses based on 33,489 "off-indication" drug-SE combinations. The left two panels show analyses based on all genetic data (GWAS and Mendelian); the right two panels show analyses based on only Mendelian data. In each panel, "no" genetic support is split into two categories: on the left, drug-SE combinations where there is incomplete genetic knowledge for the drugs' target(s), and on the right, drug-SE combinations where there is genetic knowledge for all of the drugs' target(s), and the drug-SE combination is not one of the associated phenotypes of the target(s).

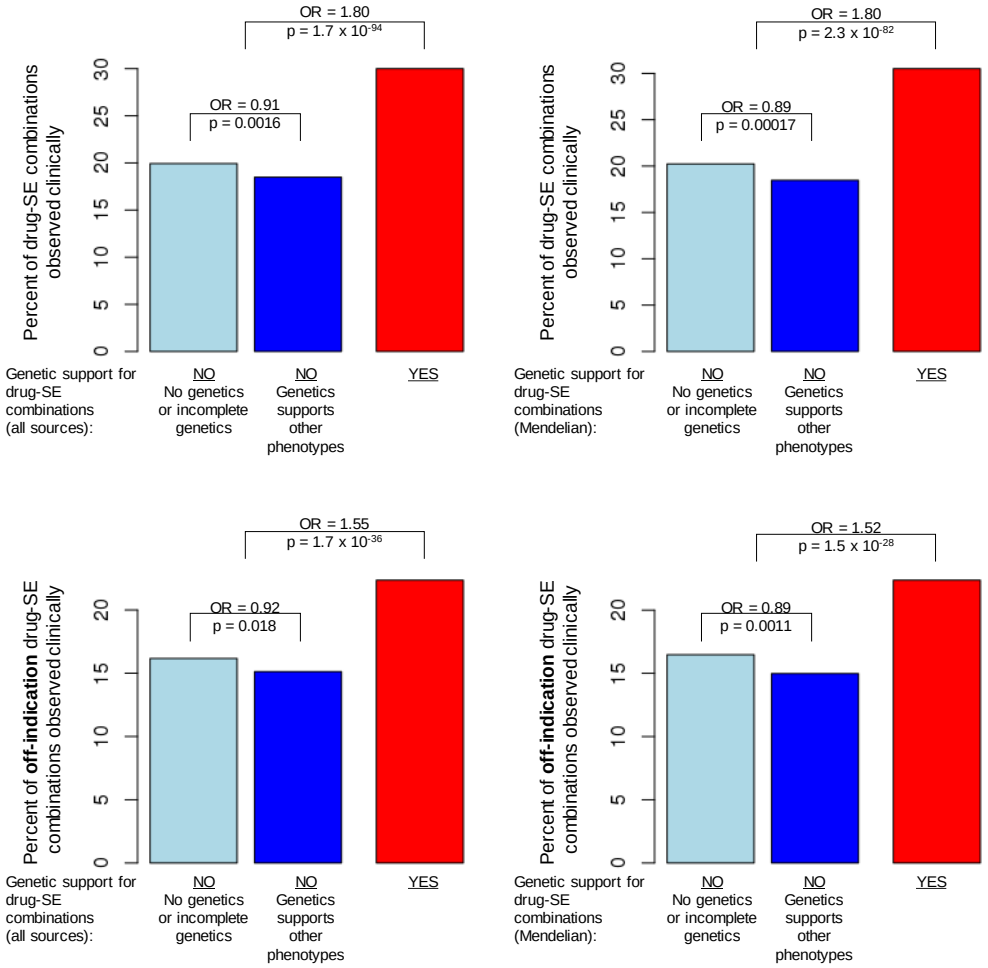

Table 1. Analysis of enrichment between genetic phenotypes and side effects in individual organ systems.

	Analysis of 38,199 (drug x organ system) combinations								Analysis of 33,489 off-indication (drug x organ system) combinations							
Phenotype category	**Drugs with genetics and side effect**	**Drugs with genetics and no side effect**	**Drugs with no genetics and side effect**	**Drugs with no genetics and no side effect**	**Rate of side effect when genetic association exists**	**Rate of side effect when genetic association does not exist**	**OR**	**p**	**Drugs with genetics and side effect**	**Drugs with genetics and no side effect**	**Drugs with no genetics and side effect**	**Drugs with no genetics and no side effect**	**Rate of side effect when genetic associati on exists**	**Rate of side effect when genetic associati on does not exist**	**OR**	**p**
blood	96	269	234	1220	26.3%	16.1%	1.86	1.34E-05	65	234	212	1193	21.7%	15.1%	1.56	0.00572
heart	145	176	451	1047	45.2%	30.1%	1.91	3.92E-07	112	162	318	992	40.9%	24.3%	2.16	5.90E-08
congenital	17	572	22	1208	2.9%	1.8%	1.63	0.165103	9	507	11	1135	1.7%	1.0%	1.83	0.2224
ear	5	112	74	1628	4.3%	4.3%	0.98	1	4	112	65	1622	3.4%	3.9%	0.89	1
endocrine	97	252	177	1293	27.8%	12.0%	2.81	3.71E-12	34	207	96	1204	14.1%	7.4%	2.06	0.00141
eye	56	256	193	1314	17.9%	12.8%	1.49	0.018631	34	250	148	1268	12.0%	10.5%	1.17	0.4617
gi	155	276	497	891	36.0%	35.8%	1.01	0.954208	126	244	379	823	34.1%	31.5%	1.12	0.37312
hepatobiliary	20	153	160	1486	11.6%	9.7%	1.21	0.42299	12	150	134	1443	7.4%	8.5%	0.86	0.76603
immune	113	335	251	1120	25.2%	18.3%	1.51	0.001757	41	232	116	914	15.0%	11.3%	1.39	0.09472
infection	64	104	536	1115	38.1%	32.5%	1.28	0.143709	49	89	455	1056	35.5%	30.1%	1.28	0.20933
metabolism	228	400	280	911	36.3%	23.5%	1.85	1.34E-08	107	292	185	820	26.8%	18.4%	1.62	0.0006
musculoskeletal	217	435	322	845	33.3%	27.6%	1.31	0.011847	138	373	221	749	27.0%	22.8%	1.25	0.07435
neoplasm	48	305	122	1344	13.6%	8.3%	1.73	0.003098	48	305	122	1344	13.6%	8.3%	1.73	0.0031
nervous	343	421	401	654	44.9%	38.0%	1.33	0.003732	190	312	266	558	37.8%	32.3%	1.28	0.04269
pregnancy	7	268	35	1509	2.5%	2.3%	1.13	0.826677	7	262	32	1486	2.6%	2.1%	1.24	0.64886
mental	208	350	255	1006	37.3%	20.2%	2.34	5.16E-14	92	278	177	912	24.9%	16.3%	1.71	0.00034
urologic	75	245	220	1279	23.4%	14.7%	1.78	0.00022	47	215	162	1220	17.9%	11.7%	1.65	0.00826
repro	75	199	186	1359	27.4%	12.0%	2.75	5.75E-10	37	185	148	1318	16.7%	10.1%	1.78	0.00542
respiratory	140	211	464	1004	39.9%	31.6%	1.44	0.003676	93	197	361	950	32.1%	27.5%	1.24	0.13058
skin	122	224	413	1060	35.3%	28.0%	1.40	0.008749	80	211	307	1004	27.5%	23.4%	1.24	0.15042
vascular	281	299	436	803	48.4%	35.2%	1.73	8.31E-08	133	245	304	739	35.2%	29.1%	1.32	0.03184
TOTAL	**2512**	**5862**	**5729**	**24096**	**30.0%**	**19.2%**	**1.80**	**1.73E-94**	**1458**	**5062**	**4219**	**22750**	**22.4%**	**15.6%**	**1.55**	**1.70E-36**

Figure 3. Summary of side effect modeling results. LEFT: results of raw enrichment analysis (from Table 1). CENTER: coefficients of the Mendelian genetics predictors in each of the 18 side effect models. RIGHT: coefficients of the GWAS genetics predictors in each of the 18 side effect models. Asterisks p-values of the the Fisher's exact test (LEFT) and coefficients from the generalized linear models (CENTER and RIGHT); (*, $p < 0.05$; **, $p < 0.005$; ***, $p < 0.0005$).

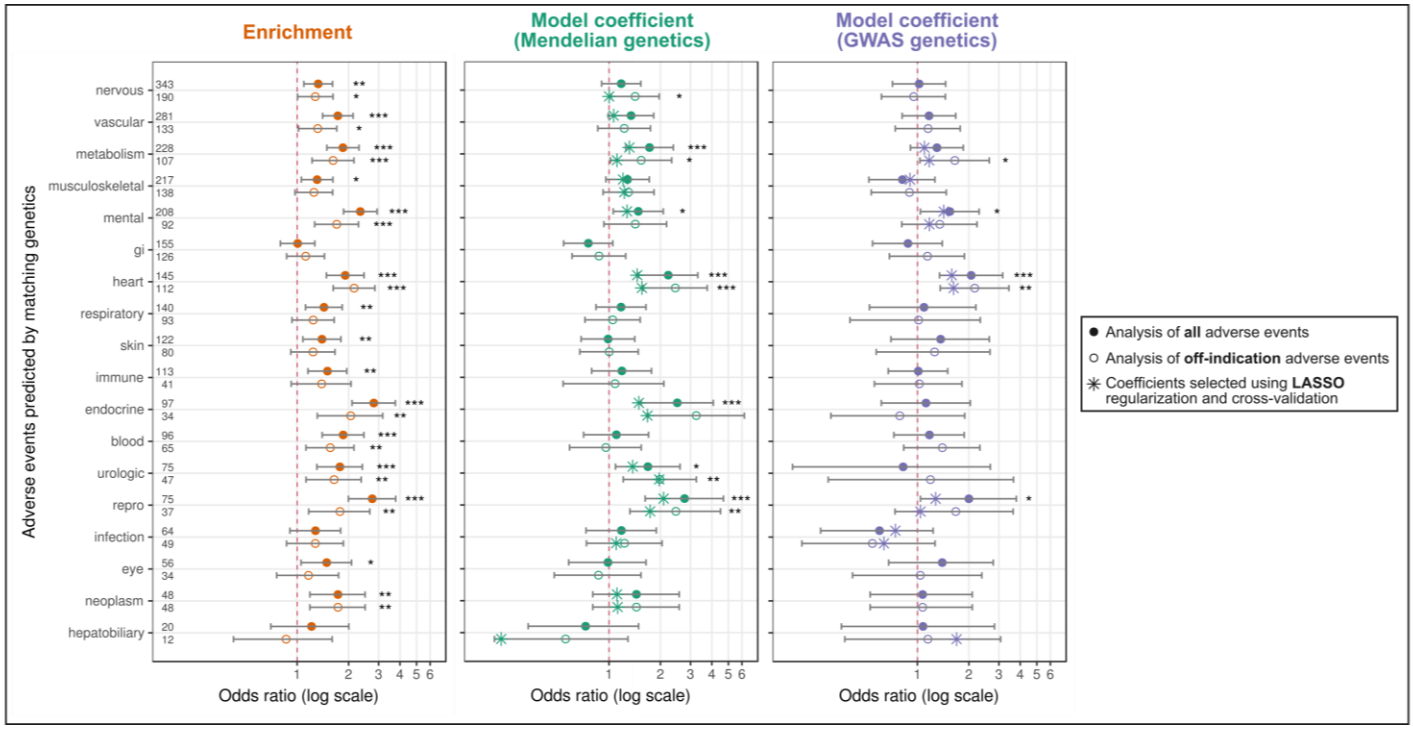
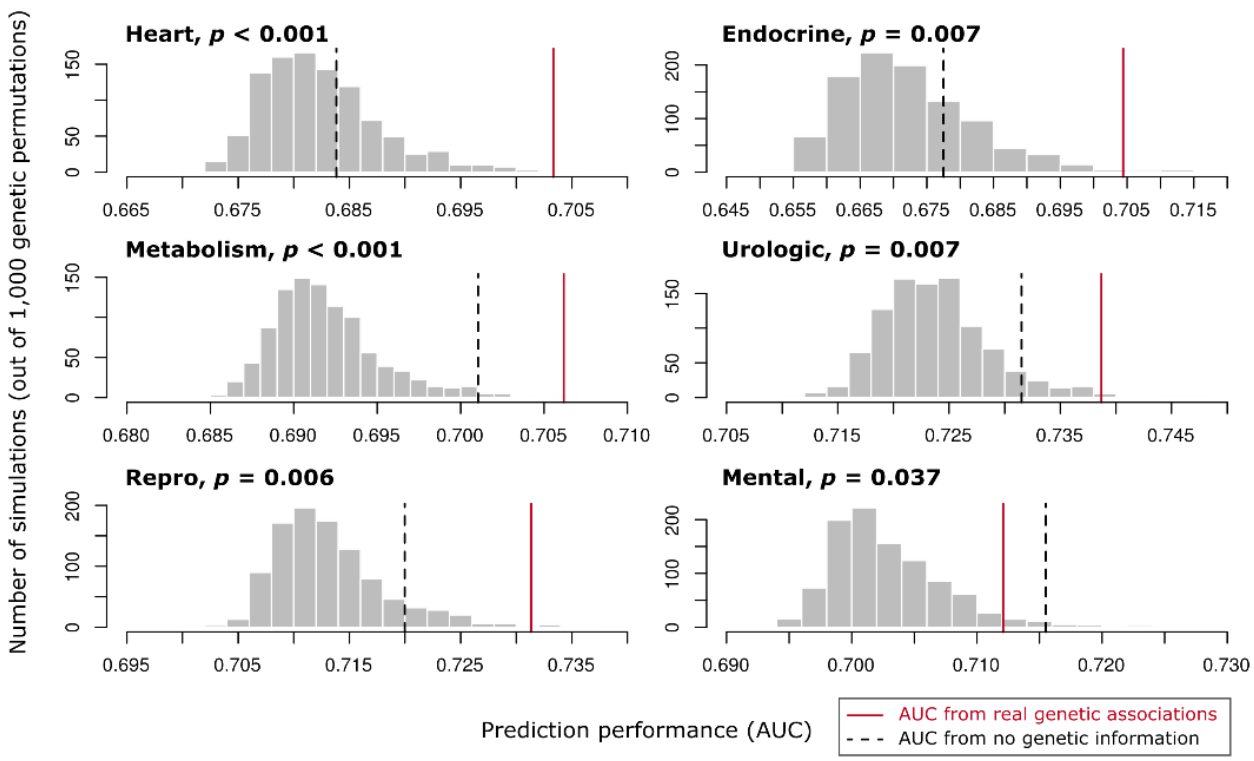
Figure 4. Results of cross validataion analysis of six side effect models where genetics was significant by regression modeling. Gray bars show distribution of leave-one-target-set-out cross-validation AUC values from 1,000 permutations of the genetic data; dashed line shows the cross-validation AUC value for a model omitting genetic information entirely; red line shows the cross-validation AUC of the regression model with real genetic data. The p values are derived from the frequency of permutation runs that exceed the AUC of the true model.