Supplementary Information for "Development of a Genetic Priority Score to Predict Drug Side Effects Using Human Genetic Evidence"

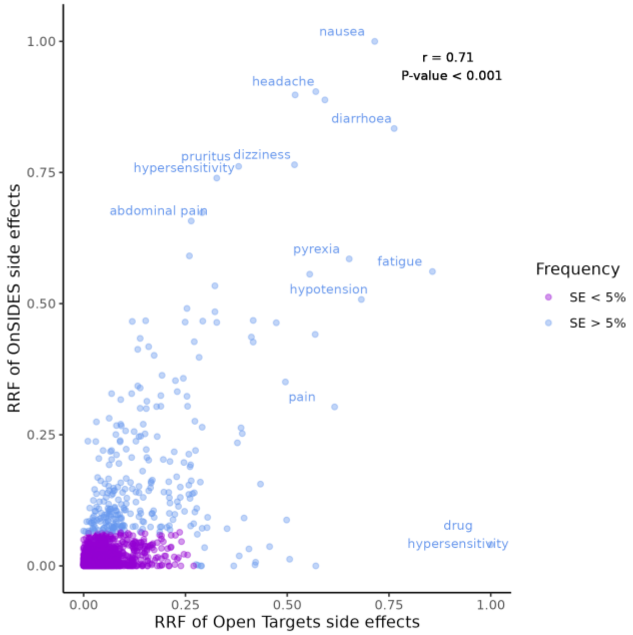
Áine Duffy[1,2,3,4], Robert Chen[1,2,3,5], David Stein[1,2,3,6], Joshua K. Park[1,2,3,5], Matthew Mort[7], Marie Verbanck[8,9], Avner Schlessinger[6,10], Yuval Itan[1,2,3,11], David N. Cooper[7], Daniel M. Jordan[1,2,3,4], Ghislain Rocheleau[1,2,3,4], Ron Do[1,2,3,4]

1. The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA
2. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3. Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA
4. Center for Genomic Data Analytics, Icahn School of Medicine at Mount Sinai, New York City, NY, USA
5. Medical Scientist Training Program, Icahn School of Medicine at Mount Sinai, New York, NY, USA
6. Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, New York City, NY, USA.
7. Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XN, UK.
8. Université Paris Cité, UR 7537 BioSTM, Paris, France.
9. Institut Curie, PSL Research University, Inserm U1331 Computational Oncology, Team Genetic Epidemiology of Cancers, Paris, France.
10. Small Molecule AI Drug Discovery Center, Icahn School of Medicine at Mount Sinai, New York, New York 10029 USA
11. Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York City, NY, USA

*Corresponding Author:

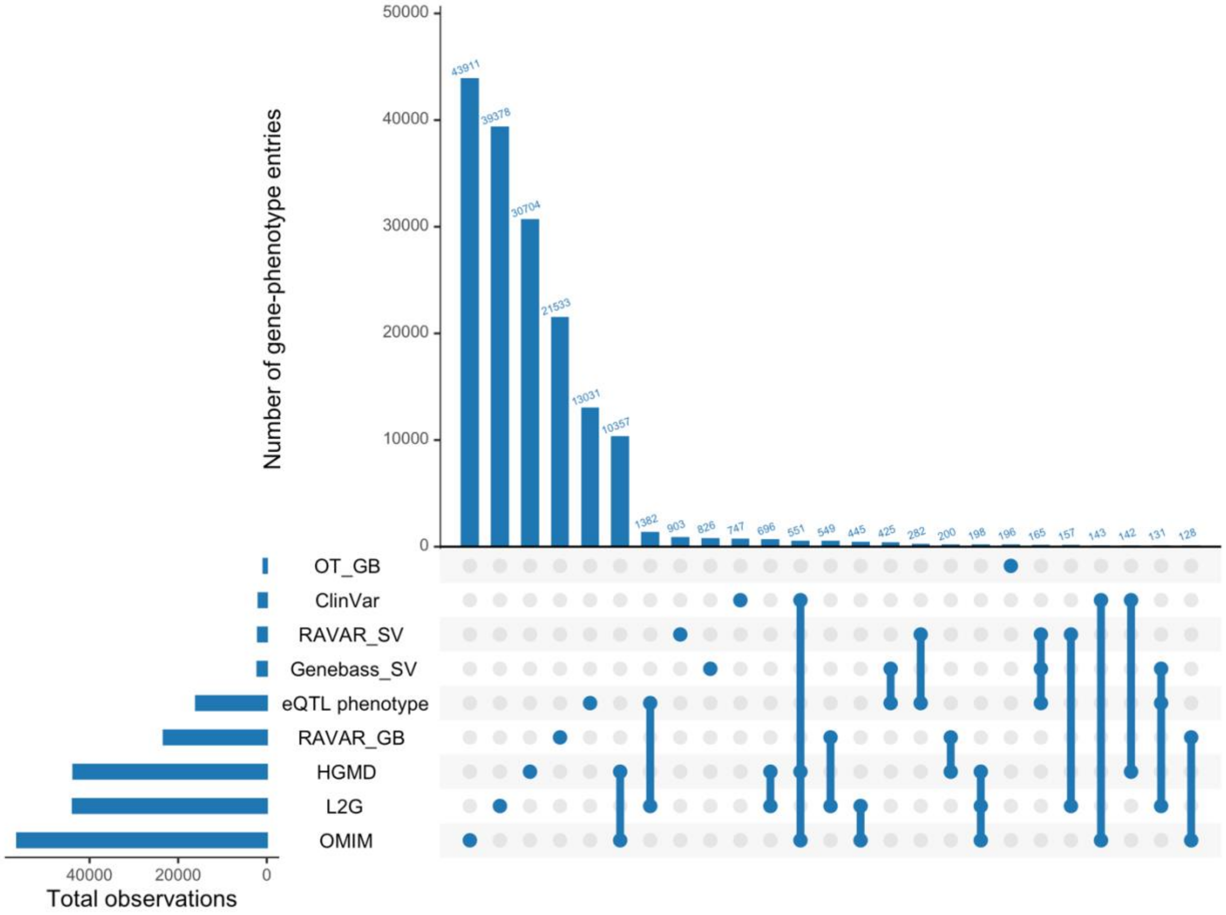
Ron Do, PhD
Professor, Department of Genetics and Genomic Sciences
The Charles Bronfman Institute for Personalized Medicine
Icahn School of Medicine at Mount Sinai
James Building, Floor 8 Room 803
3 East 101st St
New York, NY-10029
Phone Number: 212-241-6206 | Fax Number: 212-849-2643
ron.do@mssm.edu

Supplementary Fig. 1 Variation in side effect reports between the Open Target dataset and OnSIDES.



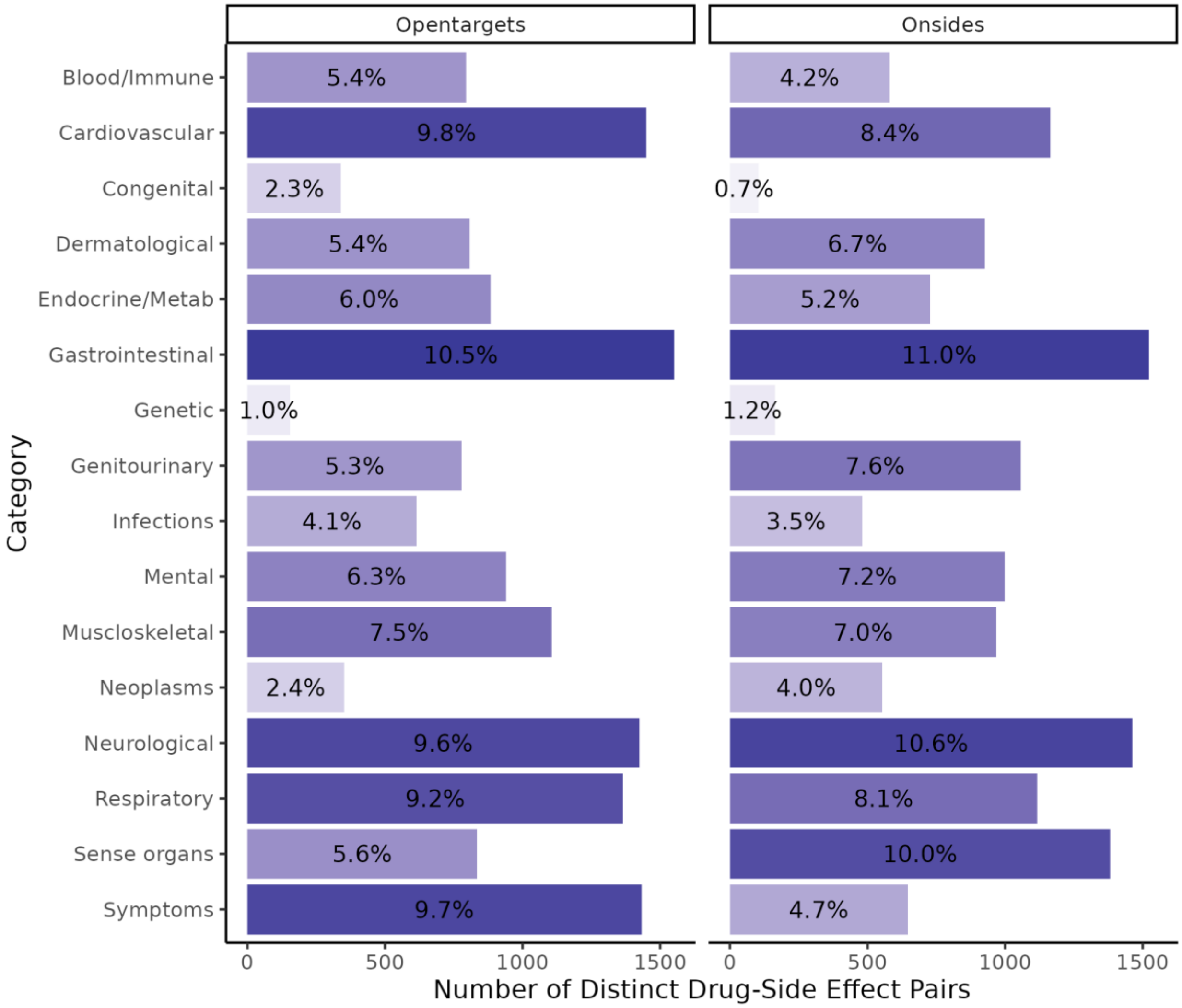
The ratio of reporting frequency (RRF) is calculated as the normalized count of drugs associated with a side effect. Each point represents a side effect with a RRF value in OnSIDES compared against its RRF value in Open Targets. The RRF of all side effects pre filtering (n = 10, 771 side effects). Side effects removed from either Open Targets or OnSIDES with a frequency greater than 5% are colored blue (n = 293) and the remaining side effects are colored purple (n = 10,478).

Supplementary Fig. 2 Gene-phenotype counts per genetic feature.



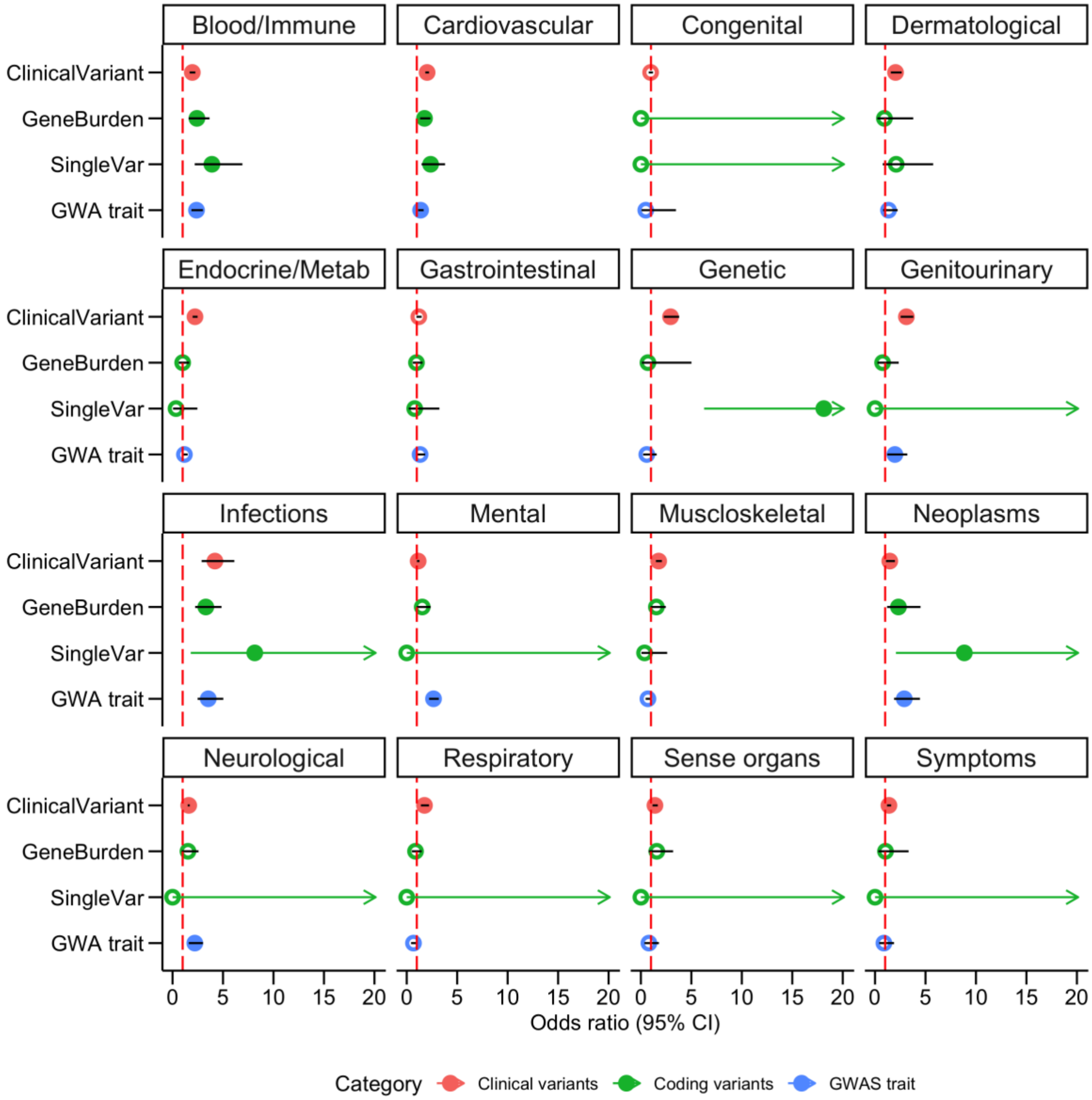
Total number of observations for each genetic feature across 19,422 genes and 502 phenotypes. OT_GB, Open Targets gene burden; Genebass_SV, Genebass single variant; RAVAR_SV, RAVAR single variant; RAVAR_GB, RAVAR gene burden

Supplementary Fig. 3 Distribution of side effect pairs by phecodeX category



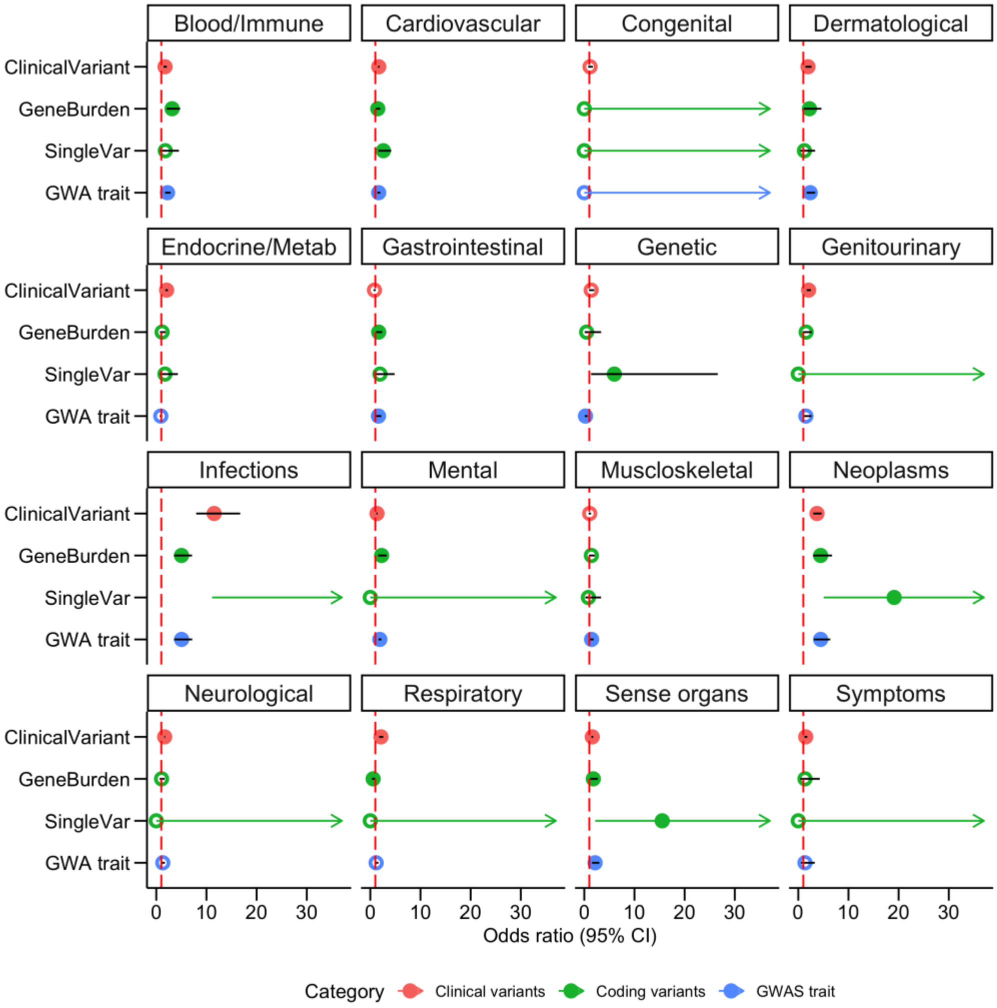
Bar plots showing the number of distinct drug side effect pairs grouped by phecodeX category in Open Targets (n = 14,827) and (n = 13,854) in OnSIDES.

Supplementary Fig. 4 Univariate association of each genetic feature with drug side effect outcome, stratified by Phecode category in the Open Target dataset.



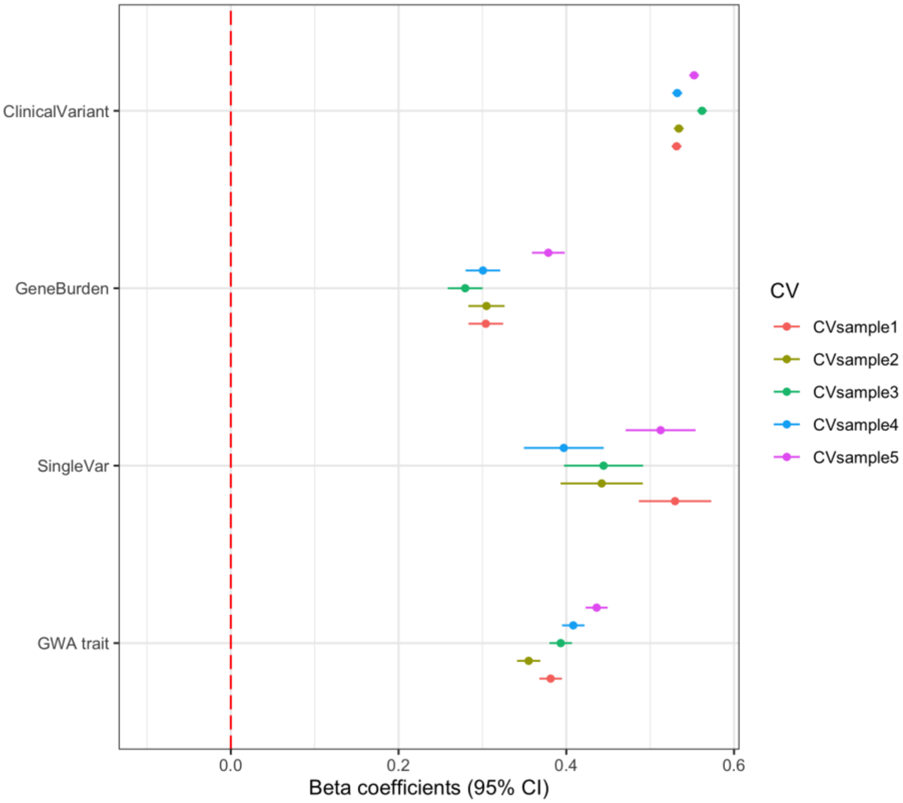
Forest plot of ORs with 95% CI from univariate logistic regression models, where each genetic feature was used as the predictor variable and drug side effect as the outcome. Analyses were stratified by phecode category within the Open Target dataset (*n* = 1,254,900 independent drug–gene–phenotype combinations), and the sample size for each stratified group is provided in the Source Data. The genetic features are grouped by color according to their genetic evidence category. The statistical test was two-sided and ORs with 95% CIs are defined in the forest plot as circles and error bars. The red dashed line represents the null odds ratio (OR=1). CI, confidence interval; OR, odds ratio.

Supplementary Fig. 5 Univariate association of each genetic feature with drug side effect outcome, stratified by Phecode category in the OnSIDES dataset.



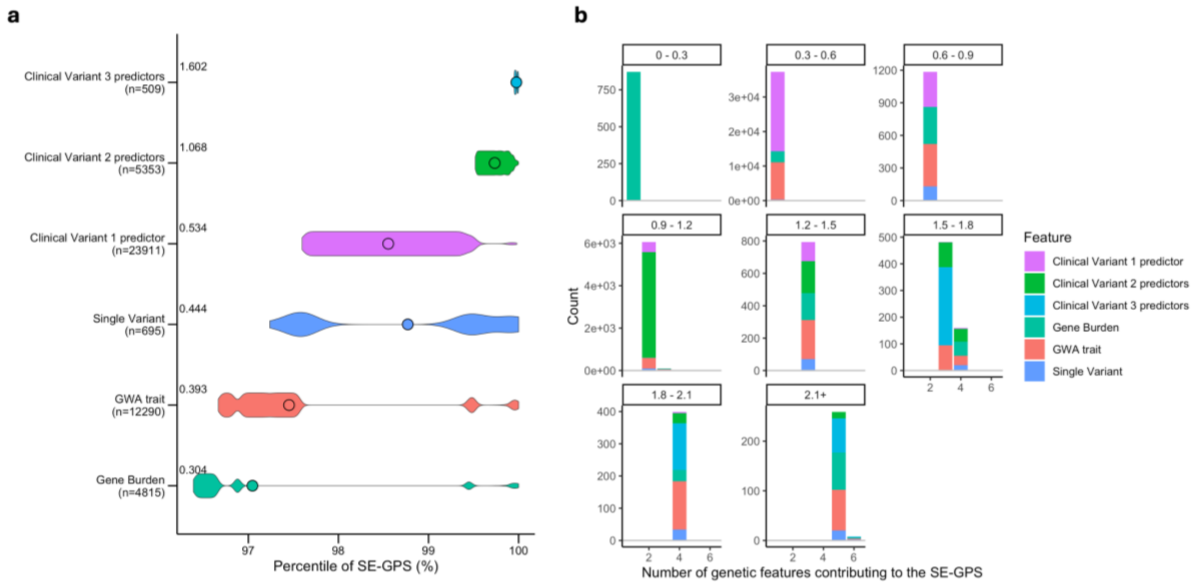
Forest plot of ORs with 95% CI from univariate logistic regression models, where each genetic feature was used as the predictor variable and drug side effect as the outcome. Analyses were stratified by phecode category within the OnSIDES dataset (*n* = 1,158,368 independent drug–gene–phenotype combinations) and the sample size for each stratified group is provided in the Source Data. The genetic features are grouped by color according to their genetic evidence category. The statistical test was two-sided and ORs with 95% CIs are defined in the forest plot as circles and error bars. The red dashed line represents the null odds ratio (OR=1). CI, confidence interval; OR, odds ratio.

Supplementary Fig. 6 Association of genetic features with drug side effects using a mixed effect regression model in the Open Target dataset.



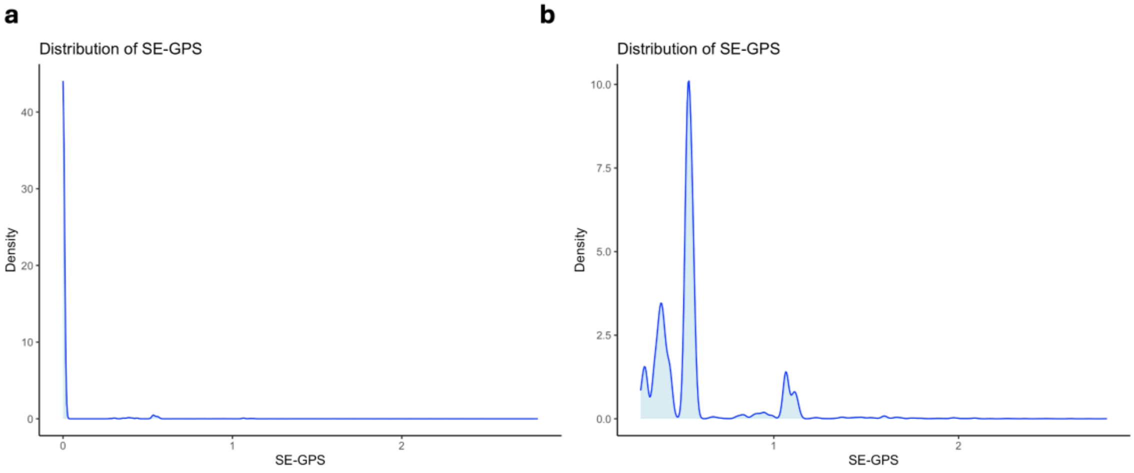
The Open Target dataset (n=1,003 drugs, 752 genes and 445 phenotypes) was split into 80% training and 20% test sets of non-overlapping groups of unique gene-phenotype pairs in five-fold cross-validation. A mixed effect regression was run for each cross-validation training set with drug side effect as the outcome, the four genetic features and 16 phecode categories as the predictor variables and the drug as the random effect variable. The side effect outcome was weighted by severity using a crowdsourced severity score[1]. Shown is a forest plot of beta coefficients with 95% CIs from the four genetic features included in each cross-validated model. The statistical test was two-sided. Each cross-validated sample is color labeled and filled circles indicate a beta coefficient with a significant *P*-value < 0.05 and the 95% CIs are defined as error bars. The red dashed line represents the null beta coefficient (β = 0). CI, confidence interval.

Supplementary Fig. 7 Contribution of each genetic feature to the SE-GPS in Open Targets.



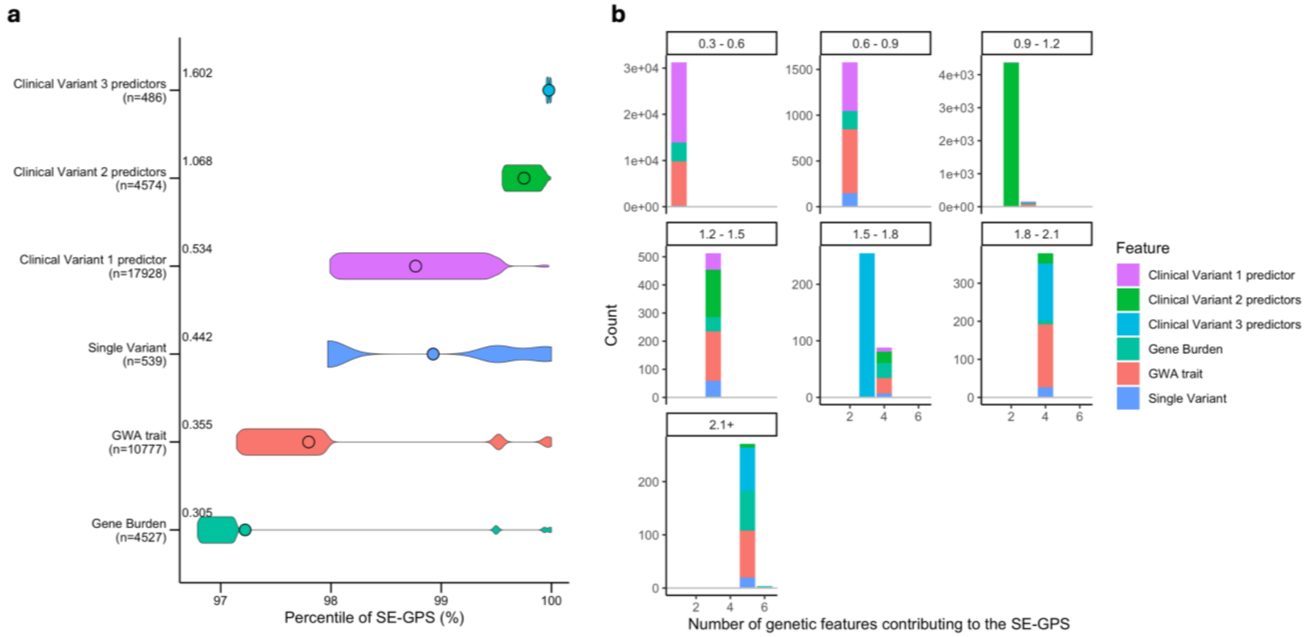
In panel a) violin plots show the distribution of each genetic feature that collectively sums to form the SE-GPS in the Open Target dataset for 1,254,900 SE-GPS across n = 334,640 gene-phecode combinations. The x-axis represents the percentile of the SE-GPS, starting at 97% to show non-zero scores only, while the y-axis separates the scores across each contributing genetic feature. The width of each violin plot represents the density of the genetic feature at each percentile, with the mean percentile marked as a circle. The total sample size of gene-phecodeX integer observations for each feature (n) and the mean weight from the five cross-validated samples is recorded under each feature on the y-axis, ordered by increasing value of these weights across the six features. The clinical variant feature was split according to the number of data sources (1, 2 or 3) for each gene phenotype observation. In panels b) bar plots show the contribution of the genetic features to the SE-GPS at 0.3 increment bins in the Open Target dataset. On the x-axis of each bar plot is the number of genetic features contributing to each score, colored by each feature present. The y-axis shows the count for each feature. In both plots, we demonstrate that as the SE-GPS increases, the number of features contributing to the score increases. SE-GPS, side-effect genetic priority score

Supplementary Fig. 8 Density distribution of the SE-GPS across the Open Target dataset



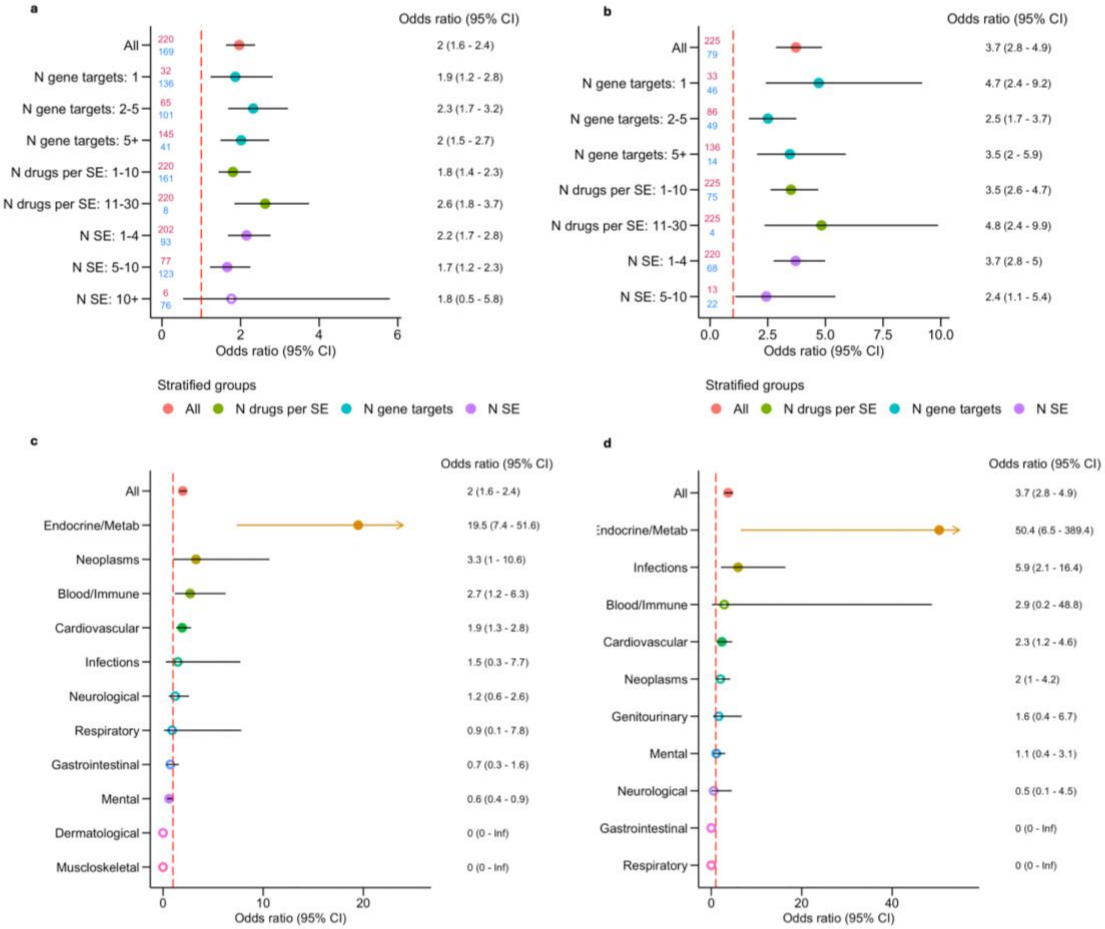
The left panel shows the full distribution of SE-GPS (n = 334,640 gene-phecode pairs), while the right panel displays the distribution restricted to gene–phecode pairs with SE-GPS greater than zero (n = 11,620 gene-phecode pairs).

Supplementary Fig. 9 Contribution genetic feature of each to the SE-GPS in OnSIDES.



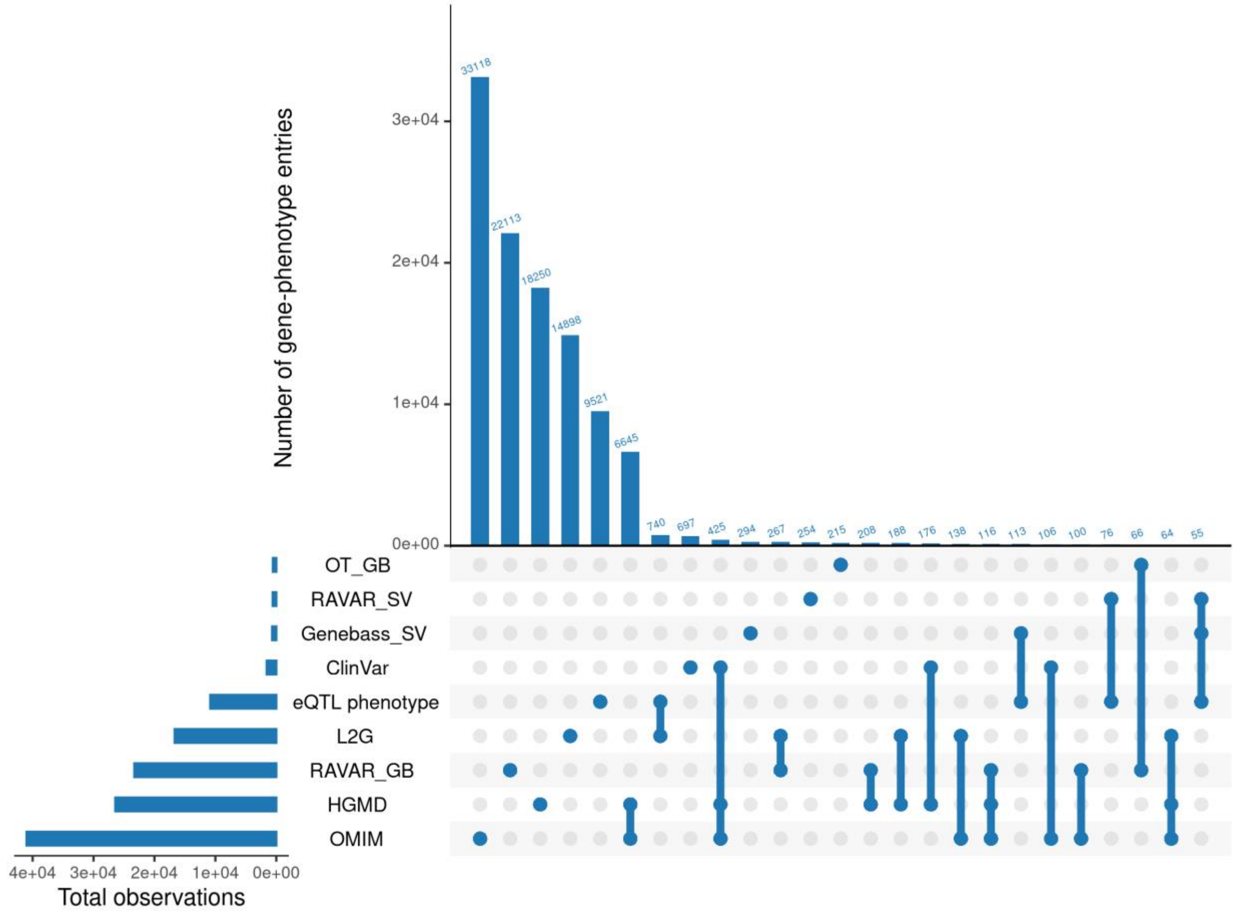
In panel a) violin plots show the distribution of each genetic feature that collectively sums to form the SE-GPS in the OnSIDES dataset for 1,158,368 SE-GPS across n = 291,712 gene-phecode combinations. The x-axis represents the percentile of the SE-GPS, starting at 97% to show non-zero scores only, while the y-axis separates the scores across each contributing genetic feature. The width of each violin plot represents the density of the genetic feature at each percentile, with the mean percentile marked as a circle. The total sample size of gene-phecodeX integer observations for each feature (n=) and the mean weight from the five cross-validated samples is recorded under each feature on the y-axis, ordered by increasing value of these weights across the six features. The clinical variant feature was split according to the number of data sources (1, 2 or 3) for each gene phenotype observation. In panels b) bar plots show the contribution of the genetic features to the SE-GPS at 0.3 increment bins in the OnSIDES dataset. On the x-axis of each bar plot is the number of genetic features contributing to each score, colored by each feature present. The y-axis shows the count for each feature. In both plots, we demonstrate that as the SE-GPS increases, the number of features contributing to the score increases. SE-GPS, side-effect genetic priority score

Supplementary Fig. 10 Association of the SE-GPS with drug side effects in the severe Open Target and OnSIDES datasets by drug side effect groupings.



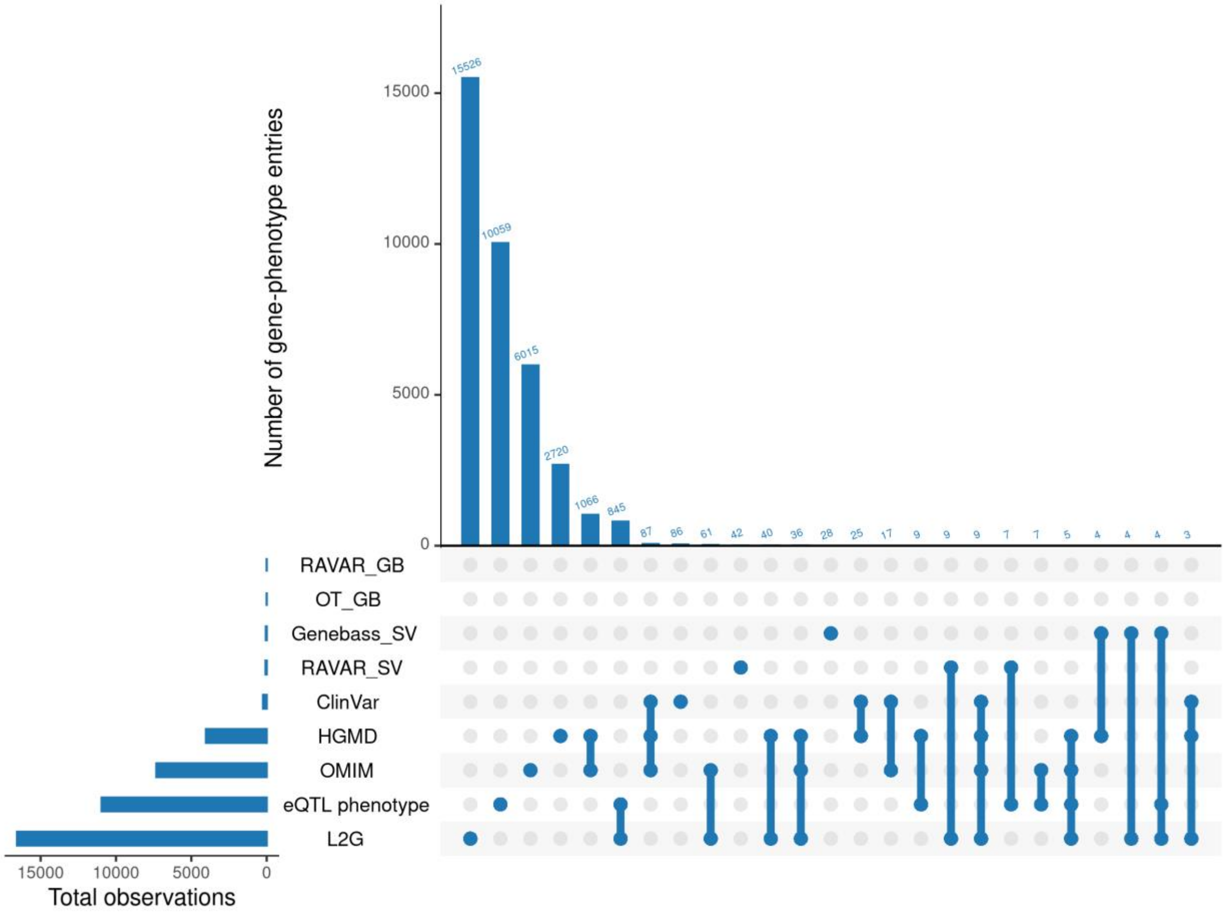
A) Forest plot showing ORs with 95% CI for the association between the presence of a SE-GPS > 0 (binarized as 1) and drug side effects, adjusted for 16 phecode categories using logistic regression. This was performed across the severe Open Target dataset (*n* = 69,290 independent drug– gene–phenotype combinations) with the OR colored in red and stratified by the number of gene targets per drug (1, 2-5, 5+; blue), the number of side effects per drug (1-4, 5-10, 10+; purple) and the number of drugs per side effect (1-10,11-30, 30+; green). For each feature, unique genes (red) and unique phenotypes (blue) are recorded on the y-axis. **B)** Replication analysis of A) using the OnSIDES severe dataset (*n* = 30, 652 independent drug– gene–phenotype combinations). **C)** Forest plot showing ORs with 95% CI for the association between the presence of a SE-GPS > 0 (binarized as 1) and drug side effects, stratified by phenotype category in the severe Open Target dataset. **D)** Replication analysis of C) using the severe OnSIDES dataset. The sample size of each stratified group is detailed in the Source Data. The statistical test was two-sided and ORs with 95% CIs are defined in the forest plot as circles and error bars. Filled circles indicate an OR with a significant *P*-value. The red dashed line represents the null odds ratio (OR=1). CI, confidence interval; N, number; OR, odds ratio, SE, side effect.

Supplementary Fig. 11 Gene-phenotype LOF observations per genetic feature.



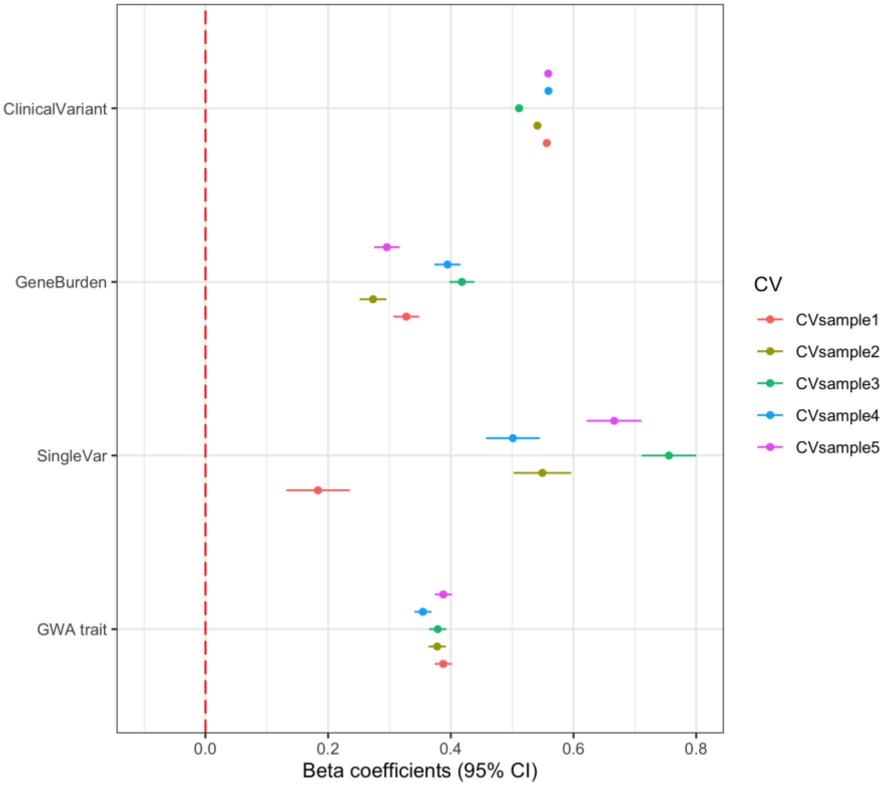
Total number of LOF observations for each genetic feature across 19,422 genes and 502 phenotypes. LOF, loss-of-function; OT_GB, Open Targets gene burden; Genebass_SV, Genebass single variant; RAVAR_SV, RAVAR single variant; RAVAR_GB, RAVAR gene burden

Supplementary Fig. 12 Gene-phenotype GOF observations per genetic feature.



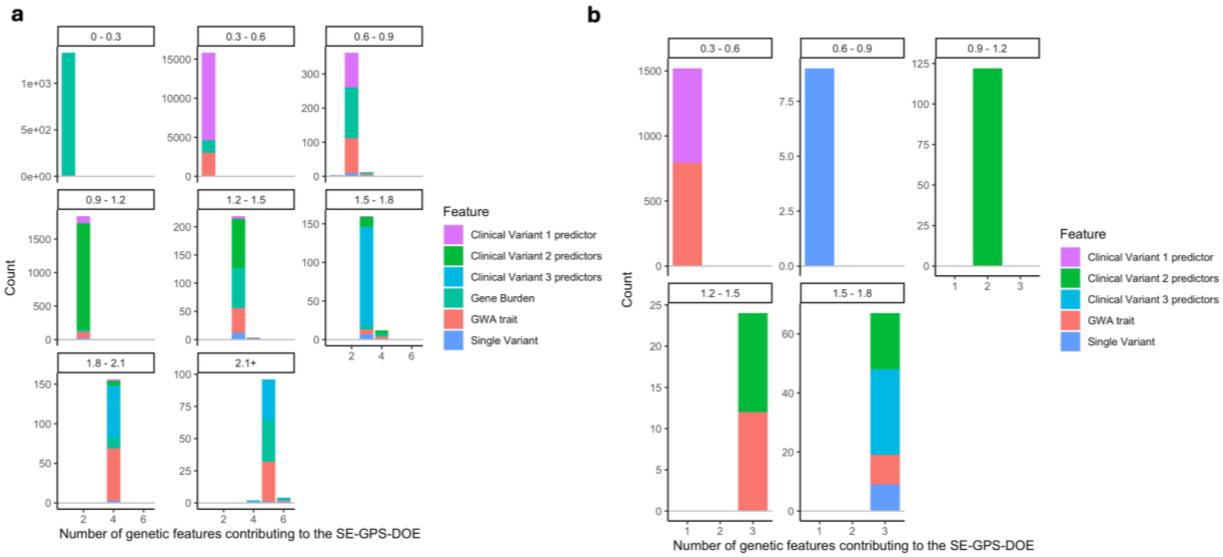
Total number of GOF observations for each genetic feature across 19,422 genes and 502 phenotypes. GOF, gain-of-function; OT_GB, Open Targets gene burden; Genebass_SV, Genebass single variant; RAVAR_SV, RAVAR single variant; RAVAR_GB, RAVAR gene burden

Supplementary Fig. 13 Association of genetic features with drug side effects using a mixed effect regression model in the Open Target dataset restricting to activator and inhibitor drugs.



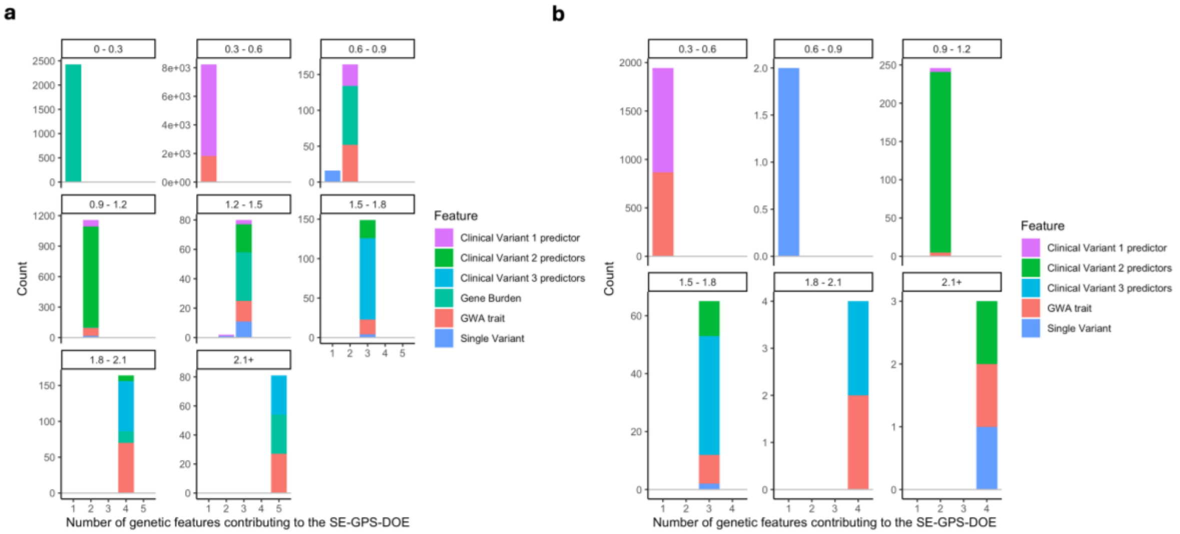
The Open Targets dataset was restricted to drugs classified as inhibitor or activator (n=913 drugs, 723 genes and 445 phenotypes). The dataset was split into 80% training and 20% test sets of non-overlapping groups of unique gene-phenotype pairs in five-fold cross-validation. A mixed effect regression was run for each cross-validation training set, with drug side effect as the outcome variable, the five genetic features, 16 phecode categories and the mechanism of action (categorized as either inhibitor or activator) as the predictor variables and the drug as the random effect variable. The side effect outcome was weighted by severity using a crowdsourced severity score[1].Shown is a forest plot of beta coefficients with 95% CIs from the four genetic features included in each cross-validated model. The statistical test was two-sided. Each cross-validated sample is color labeled and filled circles indicate a beta coefficient with a significant *P*-value < 0.05 and the 95% CIs are defined as error bars. The red dashed line represents the null beta coefficient (β = 0). CI, confidence interval.

Supplementary Fig. 14 Contribution of each genetic feature to the SE-GPS DOE in Open Targets



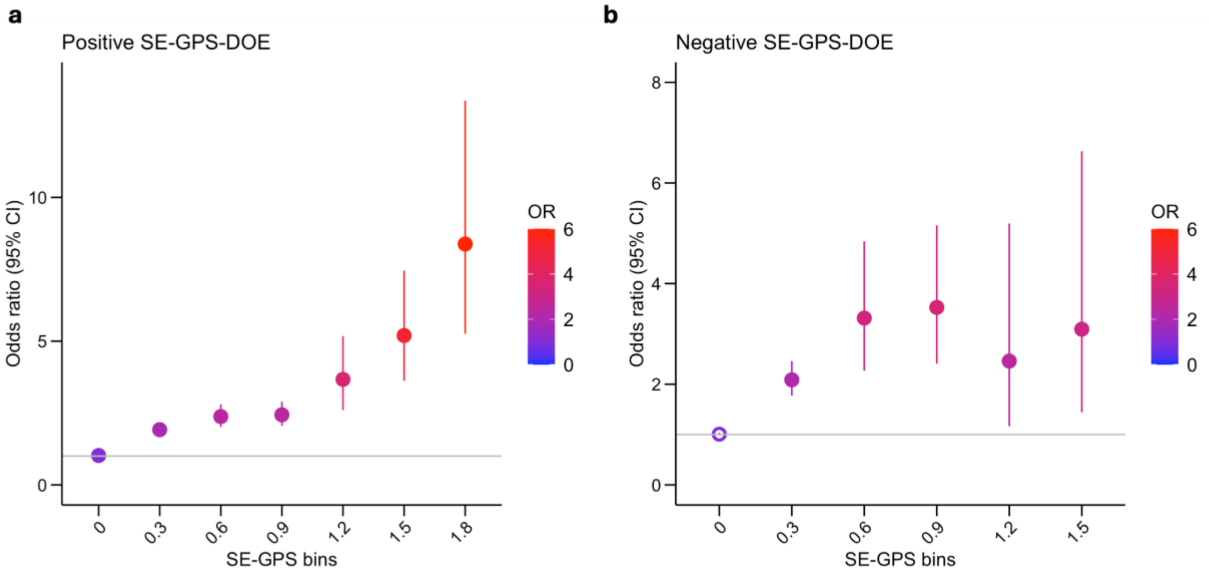
In panel **a**) bar plots show the contribution of the genetic features to the positive SE-GPS DOE at 0.3 increments bins in the Open Target dataset reflecting LOF mutations across n = 6,343 gene-phecode combinations with a non-zero SE-GPS DOE. In panel **b**) bar plots show the contribution of the genetic features to the negative SE-GPS DOE at 0.3 increments bins in the Open Target dataset reflecting GOF mutations across n = 569 gene-phecode combinations with a non-zero SE-GPS DOE. On the x-axis of each bar plot is the number of genetic features contributing to each score, colored by each feature present. The y-axis shows the count for each feature. The number of observations for the positive SE-GPS DOE is much greater than the negative SE-GPS DOE.

Supplementary Fig. 15 Contribution of each genetic feature to the SE-GPS DOE in OnSIDES



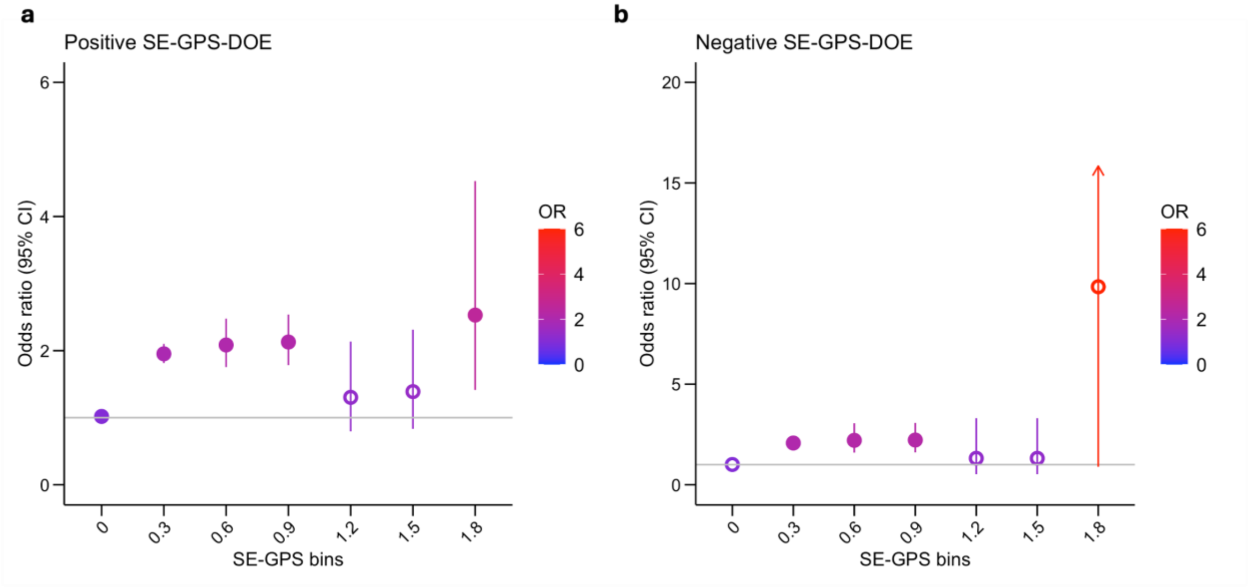
In panel **a**) bar plots show the contribution of the genetic features to the positive SE-GPS DOE at 0.3 increments bins in the OnSIDES dataset reflecting LOF mutations across n = 3,913 gene-phecode combinations with a non-zero SE-GPS DOE. In panel **b**) bar plots show the contribution of the genetic features to the negative SE-GPS DOE at 0.3 increments bins in the OnSIDES dataset reflecting GOF mutations across n = 853 gene-phecode combinations with a non-zero SE-GPS DOE. On the x-axis of each bar plot is the number of genetic features contributing to each score, colored by each feature present. The y-axis shows the count for each feature. The number of observations for the positive SE-GPS DOE is much greater than the negative SE-GPS DOE.

Supplementary Fig. 16 Association of the SE-GPS DOE at increments of 0.3 with drug side effects in the Open Target dataset.



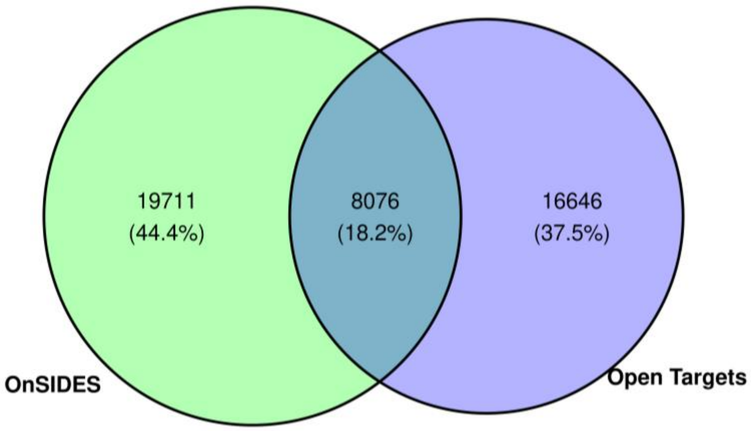
The association of drug side effects with positive SE-GPS DOE (reflecting LOF mutations) in **a**), and negative SE-GPS DOE (reflecting GOF mutations) in **b**) was investigated by binning the Open Target drug dataset (*n* = 1,150,770 independent drug– gene–phenotype combinations) into 0.3 increments of the SE-GPS DOE. Taking the absolute values of the scores, the SE-GPS DOE greater or equal to each increment was compared at each 0.3 increment with SE-GPS DOE equal to zero. A logistic regression model was performed for each increment bin with drug side effect as the outcome variable and the SE-GPS DOE bin as the predictor variable, adjusting for phecode categories as covariates. ORs with 95% CIs are defined in the forest plot as circles and error bars. The statistical test was two-sided and filled circles indicate an OR with a significant P-value < 0.05 after correcting for multiple testing. Points are colored along a blue-to-red gradient, with blue representing lower OR values and red representing higher OR values. The grey vertical line represents the null odds ratio (OR=1). CI, confidence interval; OR, odds ratio; SE-GPS-DOE, side-effect genetic priority score with direction of effect.

Supplementary Fig. 17 Association of the SE-GPS DOE at increments of 0.3 with drug side effects in the OnSIDES dataset.



The association of drug side effects with positive SE-GPS DOE (reflecting LOF mutations) in **a**), and negative SE-GPS DOE (reflecting GOF mutations) in **b**) was investigated by binning the OnSIDES drug dataset (*n* = 1,010,821independent drug– gene–phenotype combinations) into 0.3 increments of the SE-GPS DOE. Taking the absolute values of the scores, the SE-GPS DOE greater or equal to each increment was compared at each 0.3 increment with SE-GPS DOE equal to zero. A logistic regression model was performed for each increment bin with drug side effect as the outcome variable and the SE-GPS DOE bin as the predictor variable, adjusting for phecode categories as covariates. ORs with 95% CIs are defined in the forest plot as circles and error bars. The statistical test was two-sided and filled circles indicate an OR with a significant P-value < 0.05 after correcting for multiple testing. Points are colored along a blue-to-red gradient, with blue representing lower OR values and red representing higher OR values. The grey vertical line represents the null odds ratio (OR=1). CI, confidence interval; OR, odds ratio; SE-GPS-DOE, side-effect genetic priority score with direction of effect.

Supplementary Fig. 18 Overlap of gene-phecode side effect pairs in Open Targets and OnSIDES



Venn diagram showing the number of unique and shared gene-phecode side effect pairs between OnSIDES and Open Targets

Table S1: Beta estimates from mixed effect model of five genetic features on drug side effect in Open Targets five cross validated training sets

CV	Predictor	beta	95% CI	P-value
CVsample1	ClinicalVariant	0.53	0.53 - 0.54	2.225E-308
CVsample1	GWA trait	0.38	0.37 - 0.39	2.225E-308
CVsample1	GeneBurden	0.3	0.28 - 0.32	5.716E-184
CVsample1	SingleVar	0.53	0.49 - 0.57	5.477E-128
CVsample2	ClinicalVariant	0.53	0.53 - 0.54	2.225E-308
CVsample2	GWA trait	0.36	0.34 - 0.37	2.225E-308
CVsample2	GeneBurden	0.3	0.28 - 0.33	5.02E-169
CVsample2	SingleVar	0.44	0.39 - 0.49	5.936E-70
CVsample3	ClinicalVariant	0.56	0.56 - 0.57	2.225E-308
CVsample3	GWA trait	0.39	0.38 - 0.41	2.225E-308
CVsample3	GeneBurden	0.28	0.26 - 0.3	5.778E-153
CVsample3	SingleVar	0.44	0.4 - 0.49	3.3486E-76
CVsample4	ClinicalVariant	0.53	0.53 - 0.54	2.225E-308
CVsample4	GWA trait	0.41	0.39 - 0.42	2.225E-308
CVsample4	GeneBurden	0.3	0.28 - 0.32	5.151E-179
CVsample4	SingleVar	0.4	0.35 - 0.44	3.9405E-60
CVsample5	ClinicalVariant	0.55	0.55 - 0.56	2.225E-308
CVsample5	GWA trait	0.44	0.42 - 0.45	2.225E-308
CVsample5	GeneBurden	0.38	0.36 - 0.4	2.225E-308
CVsample5	SingleVar	0.51	0.47 - 0.55	4.708E-128

Abbreviations: CV, cross-validated; CI, confidence interval ; SE-GPS, side effect genetic priority score

Table S2: Association of the SE-GPS with drug side effects in the Open Target dataset by CV sample.

CV	OR	95% CI	P-value
CVsample2	2.63	2.31 - 2.99	2.4583E-48
CVsample4	2.51	2.21 - 2.85	1.3337E-45
CVsample1	2.46	2.17 - 2.79	3.5278E-44
CVsample3	2.2	1.93 - 2.5	1.7994E-33
CVsample5	2	1.75 - 2.28	3.3123E-25

Abbreviations: CV, cross-validated; OR, odds ratio; CI, confidence interval; SE-GPS, side effect genetic priority score

Table S3: Beta estimates from mixed effect model of five genetic features on drug side effect in directional Open Target five cross validated training sets

CV	Predictor	beta	95% CI	P-value
CVsample1	ClinicalVariant	0.56	0.55 - 0.56	2.225E-308
CVsample1	GWA trait	0.39	0.37 - 0.4	2.225E-308
CVsample1	GeneBurden	0.33	0.31 - 0.35	2.616E-205
CVsample1	SingleVar	0.18	0.13 - 0.24	3.8877E-12
CVsample2	ClinicalVariant	0.54	0.54 - 0.55	2.225E-308
CVsample2	GWA trait	0.38	0.36 - 0.39	2.225E-308
CVsample2	GeneBurden	0.27	0.25 - 0.29	8.504E-134
CVsample2	SingleVar	0.55	0.5 - 0.6	2.248E-116
CVsample3	ClinicalVariant	0.51	0.51 - 0.52	2.225E-308
CVsample3	GWA trait	0.38	0.36 - 0.39	2.225E-308
CVsample3	GeneBurden	0.42	0.4 - 0.44	2.225E-308
CVsample3	SingleVar	0.76	0.71 - 0.8	3.485E-243
CVsample4	ClinicalVariant	0.56	0.55 - 0.57	2.225E-308
CVsample4	GWA trait	0.35	0.34 - 0.37	2.225E-308
CVsample4	GeneBurden	0.39	0.37 - 0.42	2.588E-288
CVsample4	SingleVar	0.5	0.46 - 0.55	3.089E-111
CVsample5	ClinicalVariant	0.56	0.55 - 0.56	2.225E-308
CVsample5	GWA trait	0.39	0.37 - 0.4	2.225E-308
CVsample5	GeneBurden	0.3	0.27 - 0.32	2.486E-168
CVsample5	SingleVar	0.67	0.62 - 0.71	1.754E-186

Abbreviations: CV, cross-validated; CI, confidence interval

Table S4: Association of the SE-GPS DOE with drug side effects in the Open Target dataset by CV sample.

CV	OR	95% CI	P-value
CVsample5	2.18	1.82 - 2.6	1.3623E-17
CVsample3	2.03	1.64 - 2.52	9.1885E-11
CVsample2	1.91	1.59 - 2.29	2.7363E-12
CVsample1	1.6	1.29 - 1.98	1.7667E-05
CVsample4	1.6	1.32 - 1.93	1.1279E-06

Abbreviations: CV, cross-validated; OR, odds ratio; CI, confidence interval; SE-GPS DOE, side effect genetic priority score with direction of effect

Table S5: Mixed effect regression beta coefficients used to create the SE-GPS and SE-GPS-DOE in OnSIDES and across all genes

Predictor	SE-GPS	SE-GPS-DOE
ClinicalVariant	0.53	0.56
GWA trait	0.36	0.39
GeneBurden	0.3	0.3
SingleVar	0.44	0.67

Abbreviations: SE-GPS, side effect genetic priority score; SE-GPS DOE, side effect genetic priority score with direction of effect

Table S6: Table of toxicity class mappings to Phecode categories.

Toxicity class	**Phecode category**
Carcinogenicity	Neoplasms
Cardiotoxicity	Cardiovascular
Dermatological toxicity	Dermatological
Gastrointestinal toxicity	Gastrointestinal
Hematological toxicity	Blood/immune
Hepatotoxicity	Gastrointestinal
Immune system toxicity	Blood/immune
Infectious disease	Infections
Metabolic toxicity	Endocrine/metabolic
Musculoskeletal toxicity	Musculoskeletal
Nephrotoxicity	Genitourinary
Neurotoxicity	Neurological
Psychiatric toxicity	Mental
Respiratory toxicity	Respiratory
Vascular toxicity	Cardiovascular

References

1. Gottlieb, A., Hoehndorf, R., Dumontier, M. & Altman, R.B. Ranking adverse drug reactions with crowdsourcing. *J. Med. Internet Res.* **17**, e80 (2015).