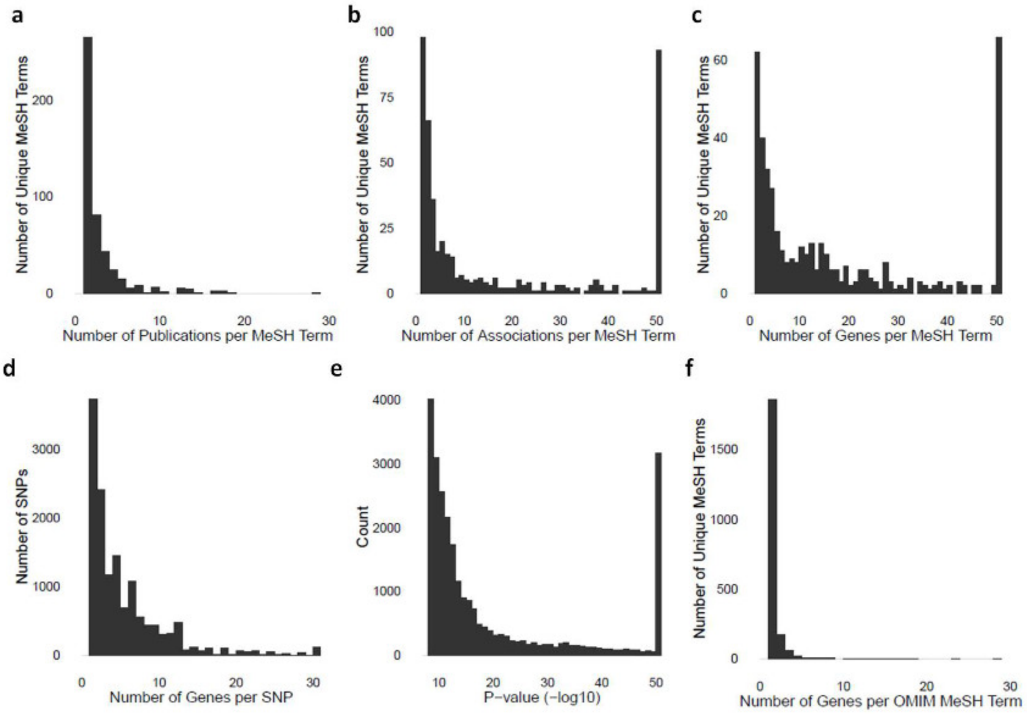
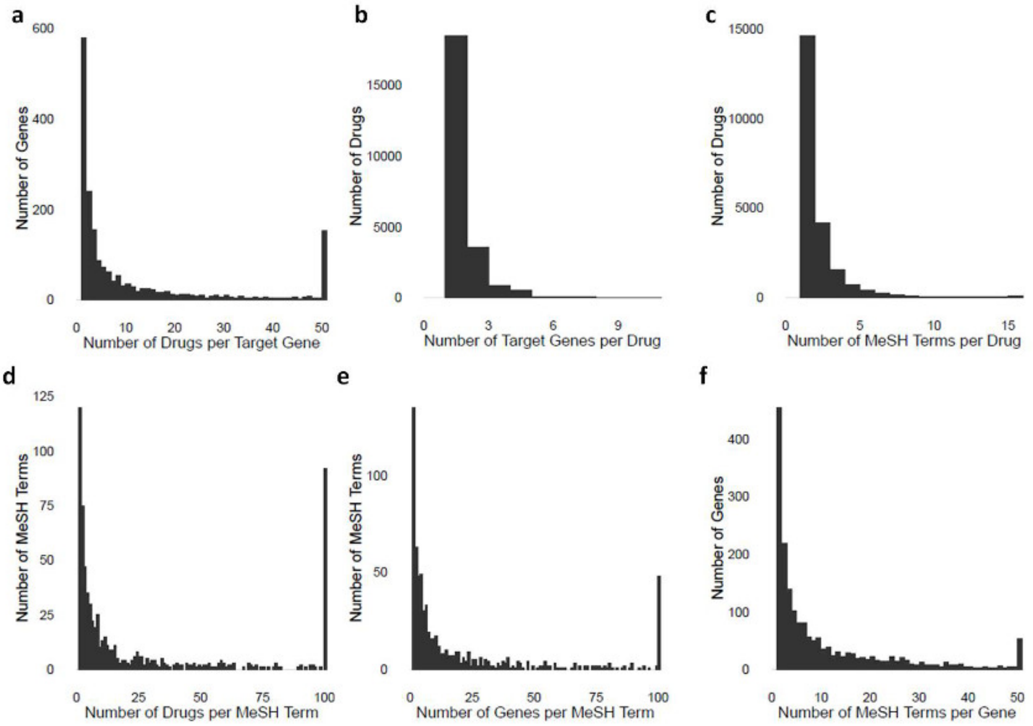


Supplementary Figure 1

Summary of genetic association data and their traits and gene mappings.

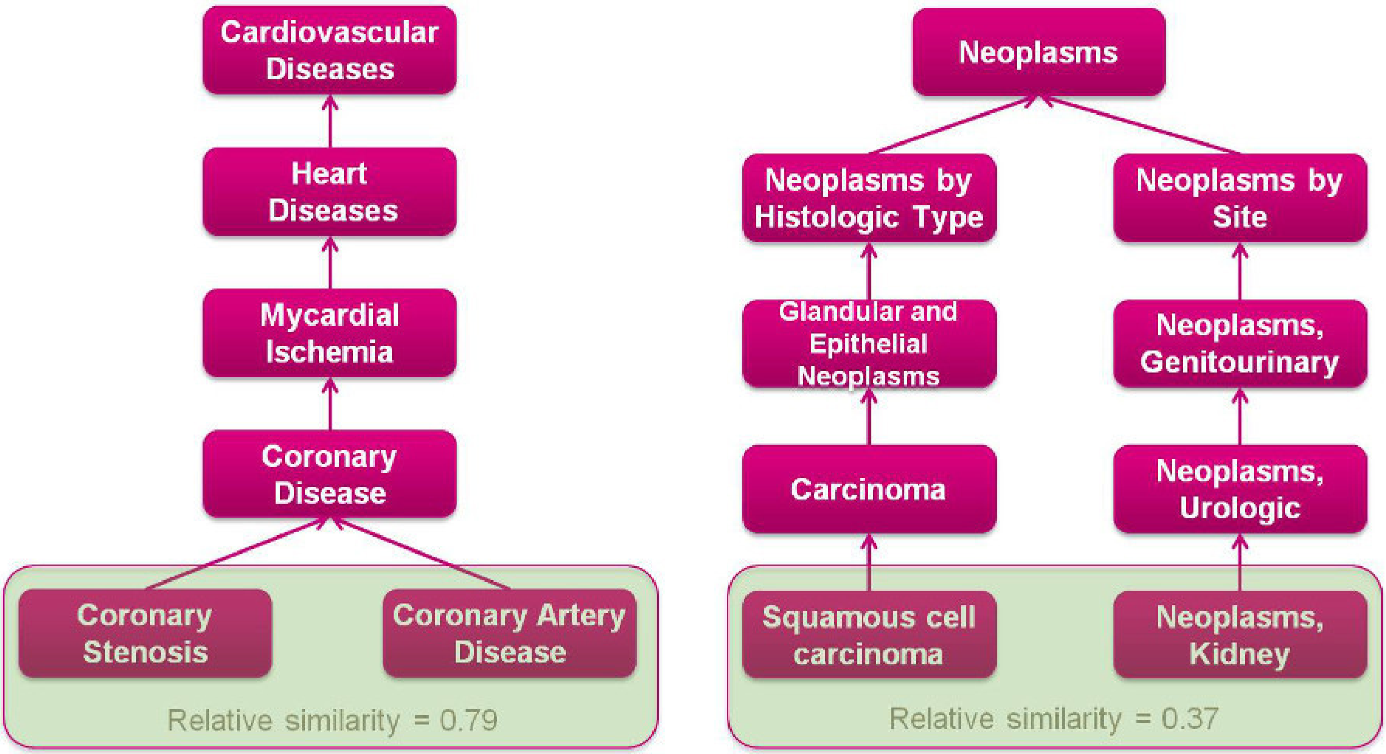
Distribution of the (**a**) number of publications or sources and (**b**) reported associations for each unique MeSH term. (**c**) Distribution of the number of genes mapped for each MeSH term. (**d**) Distribution of the number of genes mapped to each SNP (excluding SNPs with no genes mapped; *n* = 5,272). (**e**) Distribution of *P* values for all unique associations. These summaries are limited to publications and sources with at least one association with a *P* value ≤1 × 10^{-8}. Panels **b** and **c** were truncated at 50, panel **d** was truncated at 30 and panel **e** was truncated at 100. All values over those thresholds are shown at the maximum value. (**f**) Distribution of the number of genes for each unique MeSH term in OMIM.



Supplementary Figure 2

Summary of the drug data and their target gene and indications.

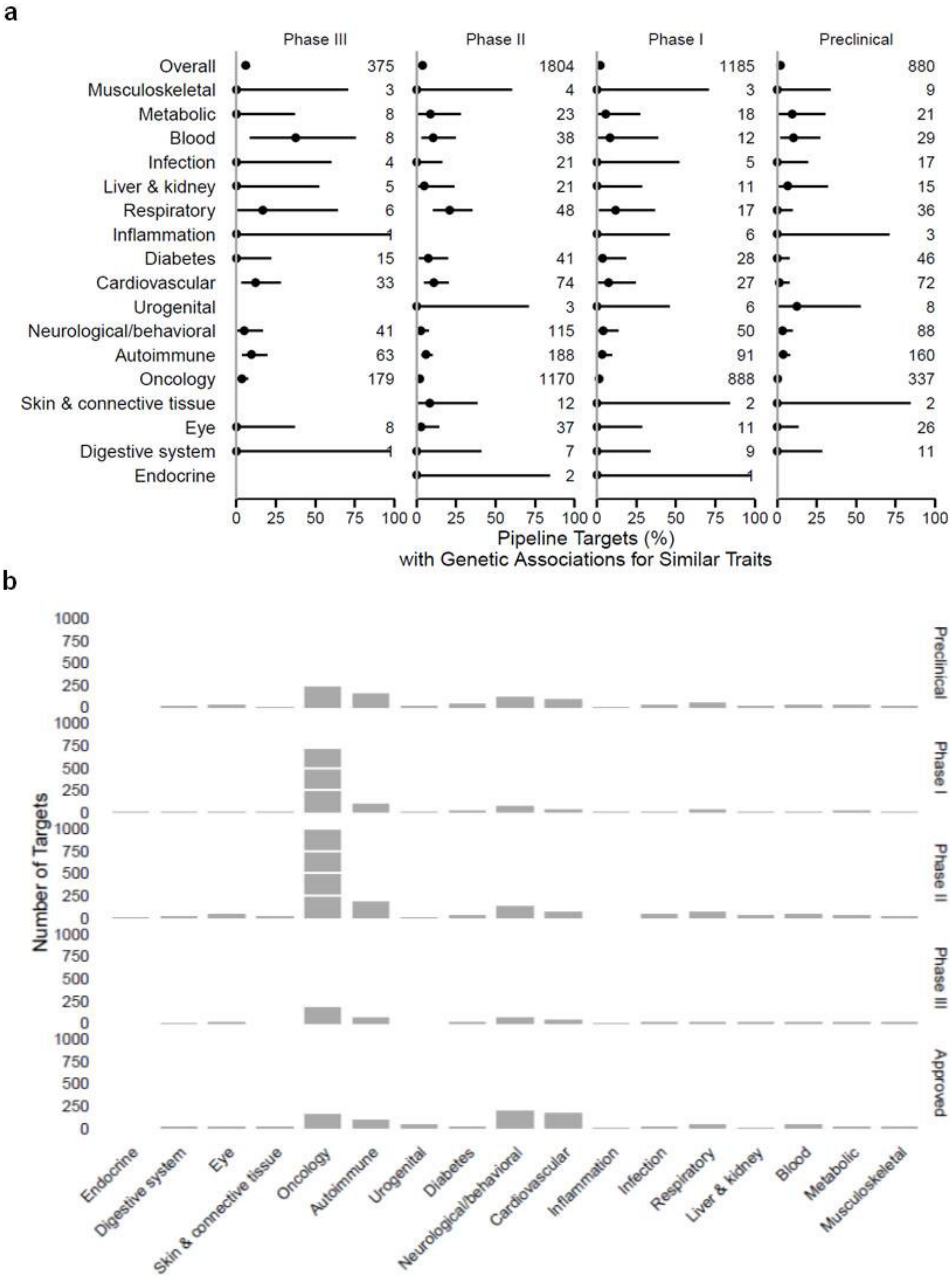
Distribution of the (**a**) number of drugs observed for each target gene in the analysis data set (truncated at 50). (**b**) Distribution of the number of target genes for each drug (i.e., multiple drug targets or combinations of therapeutic agents). (**c**) Distribution of the number of MeSH terms (i.e., unique indications) for each drug (truncated at 15). (**d**) Distribution of the number of drugs listed for each MeSH term (truncated at 100). (**e**) Distribution of the number of target genes for each MeSH term (truncated at 100). (**f**) Distribution of the number of MeSH terms for each target gene (truncated at 50).



Supplementary Figure 3

Illustrated use of the MeSH ontology to estimate relative similarity.

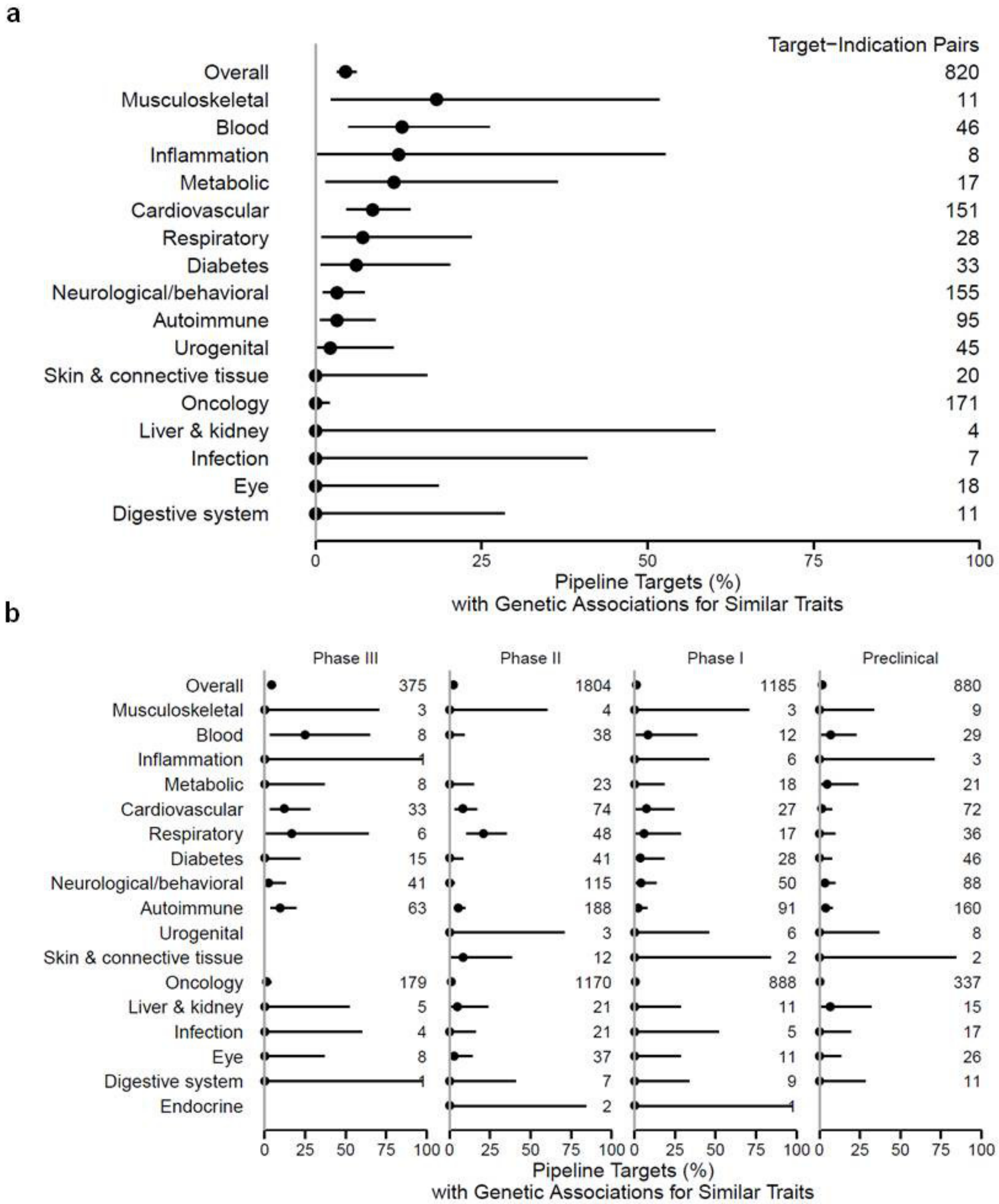
The methods used in this study (lin and resnik, implemented in UMLS::Similarity) combined both path length and information content. See the Online Methods for additional details.



Supplementary Figure 4

Overlap of drug targets with genetic associations by disease category and latest development phase.

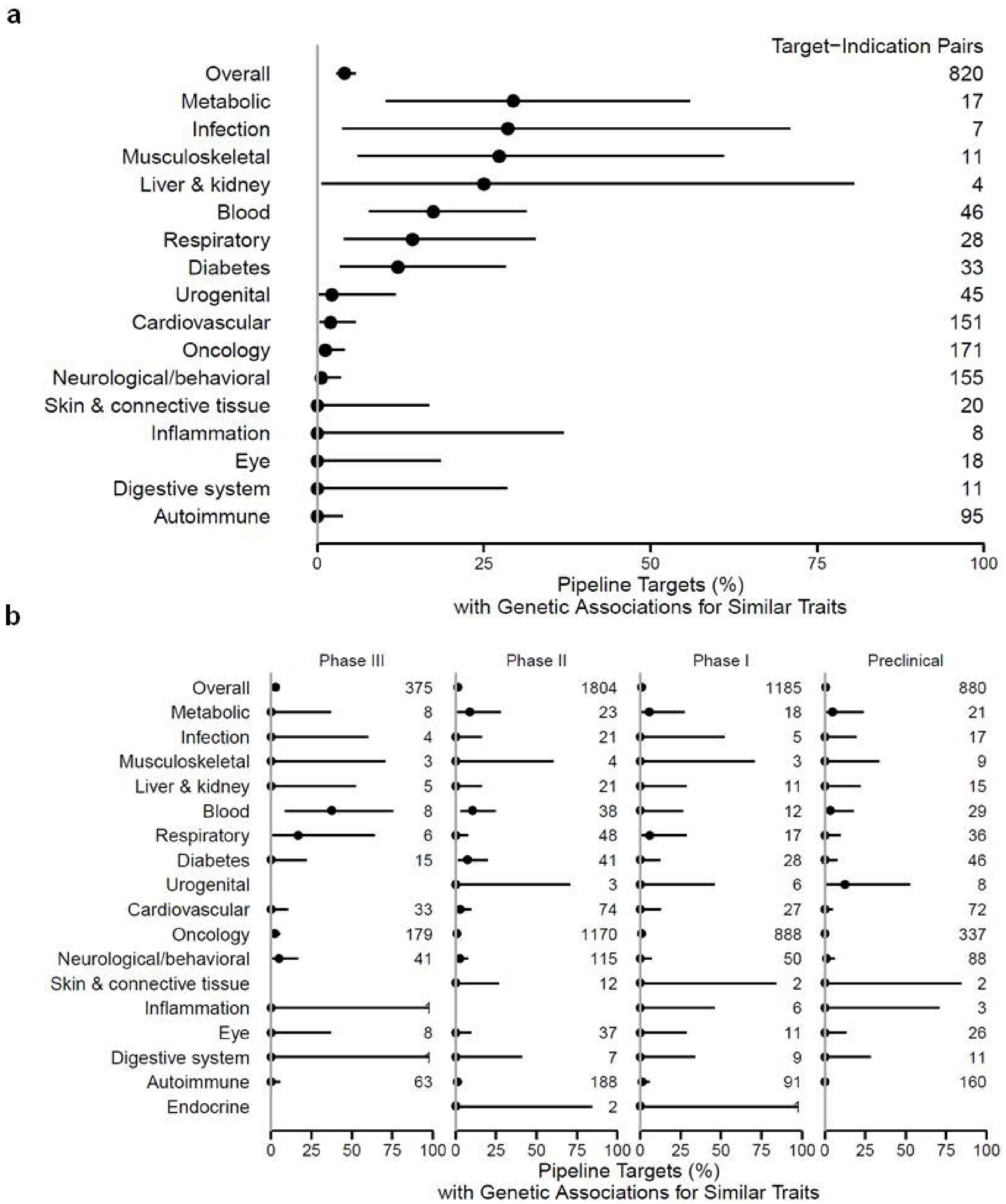
(**a**) Overlap between drug targets and their indications with genetic associations for similar traits. The percentage of target-indication pairs overlapping with gene-trait combinations from GWASdb or OMIM for the latest development phase each pair achieved as recorded in Pharmaprojects. The number of unique target-indication pairs for each category at each phase is shown to the right of each plot. Exact 95% confidence intervals are shown. (**b**) Distribution of the number of target-indication pairs at each phase by category.



Supplementary Figure 5

Overlap between drug targets and their indications with genetic associations for similar traits with genetic associations restricted to GWASdb only.

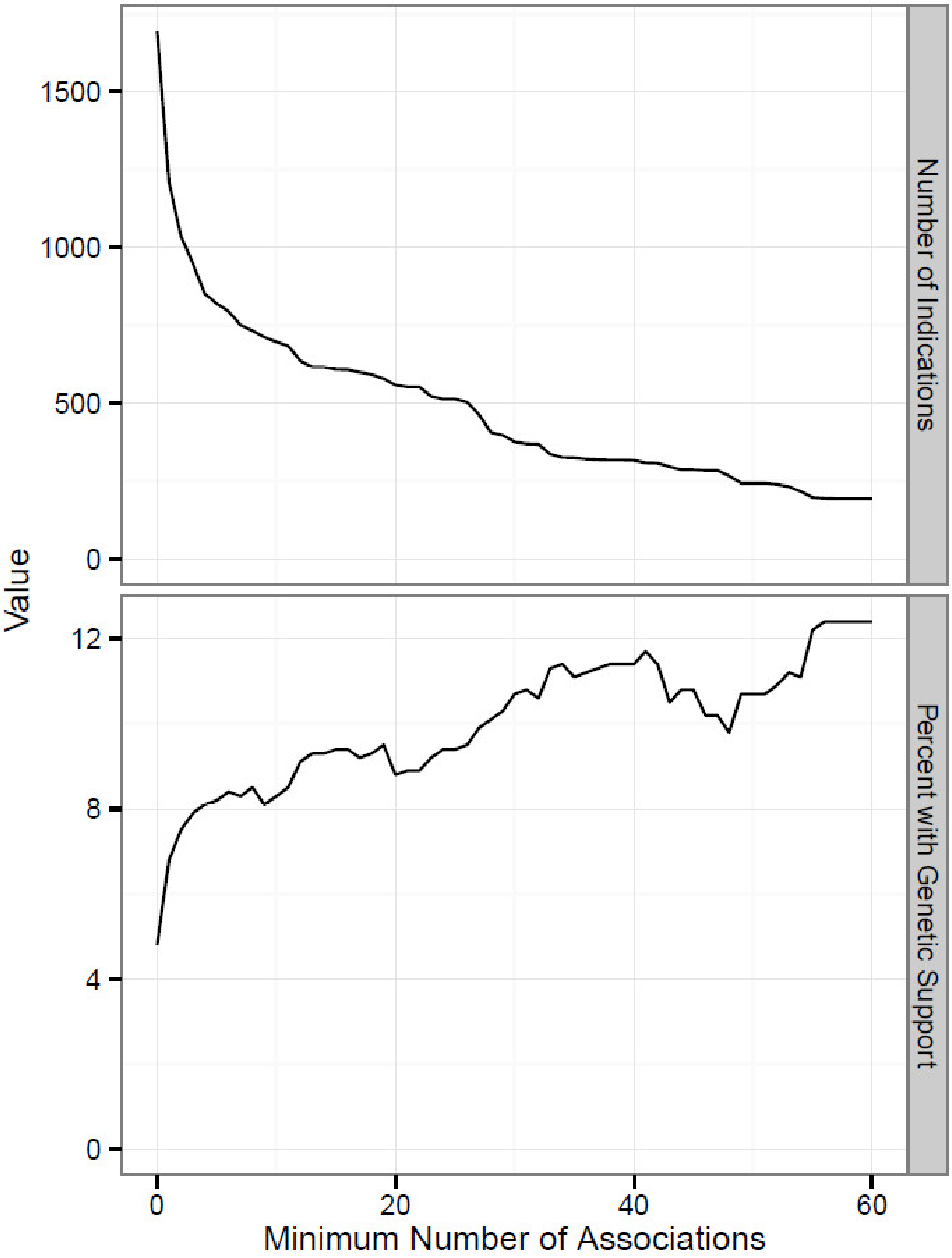
Overlap for (**a**) drugs approved in the United States or European Union and (**b**) the furthest development phase to which each target-indication pair progressed. Exact 95% confidence intervals are shown.



Supplementary Figure 6

Overlap between drug targets and their indications with genetic associations for similar traits with genetic associations restricted to OMIM only.

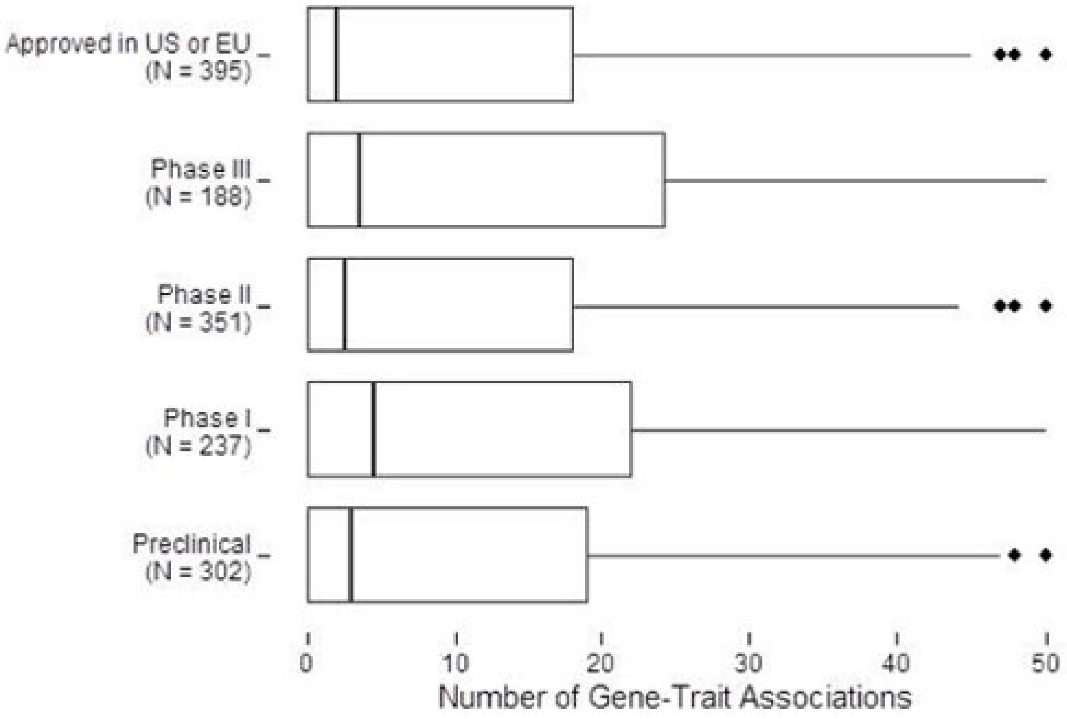
Overlap for (**a**) drugs approved in the United States or European Union and (**b**) the furthest development phase to which each target-indication pair progressed. Exact 95% confidence intervals are shown.



Supplementary Figure 7

Tradeoff between the number of indications studied and overall genetic support.

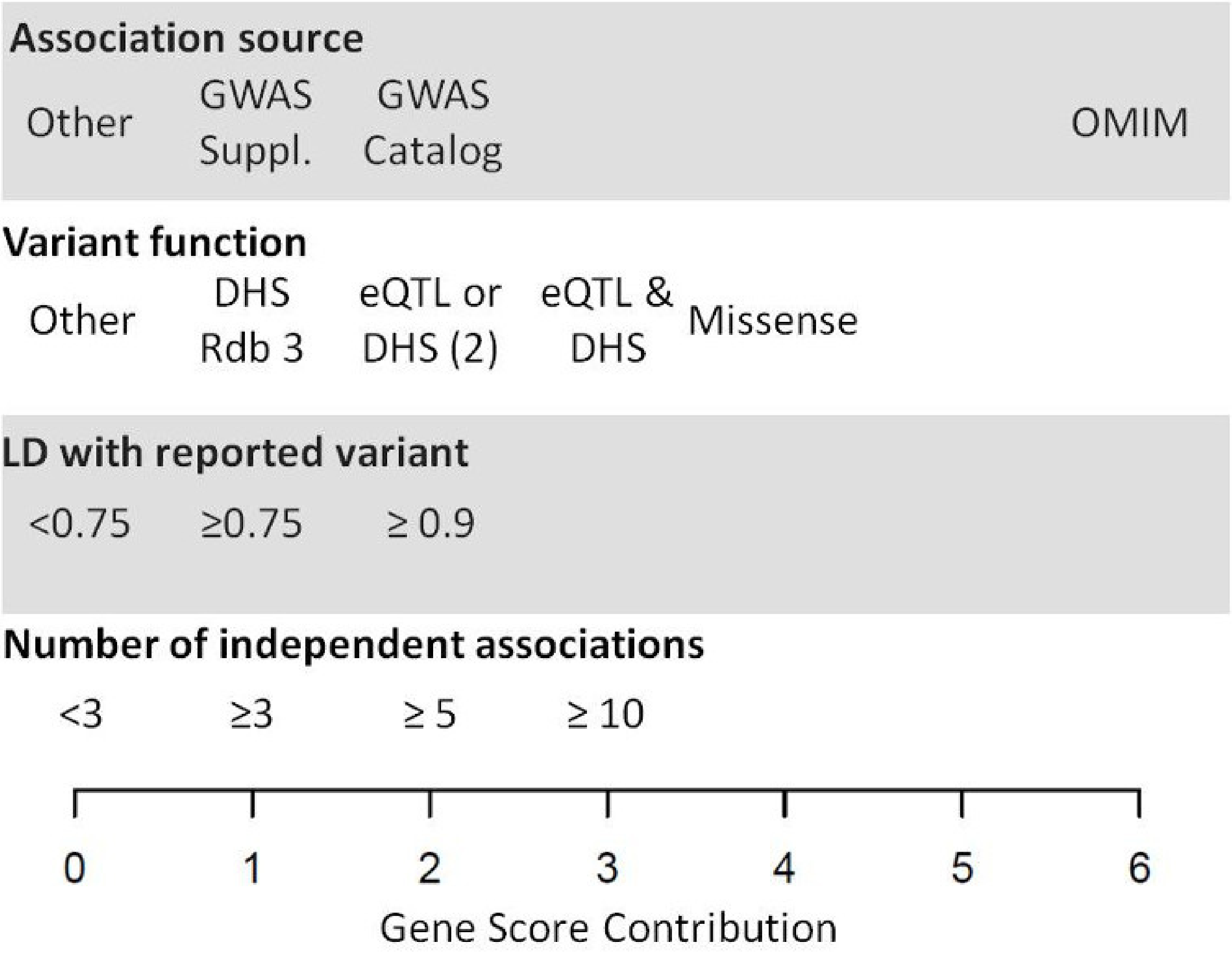
The tradeoff between the number of indications studied and overall genetic support when setting a lower bound on the number of independent genes associated with a trait related to each indication (relative similarity ≥ 0.7), restricted to drugs approved in the United States or European Union. The percentage of target-indication pairs with genetic support increases as indications are restricted to those with the most genetic information available, although at the cost of considering far fewer indications. The analyses reported in **Figure 3** and **Supplementary Figures 4–6** selected five as the threshold, where the first enrichment plateau is observed.



Supplementary Figure 8

Distribution of the number of genes associated with traits similar (≥0.7) to the indications included in the analysis of overlap with genetic associations.

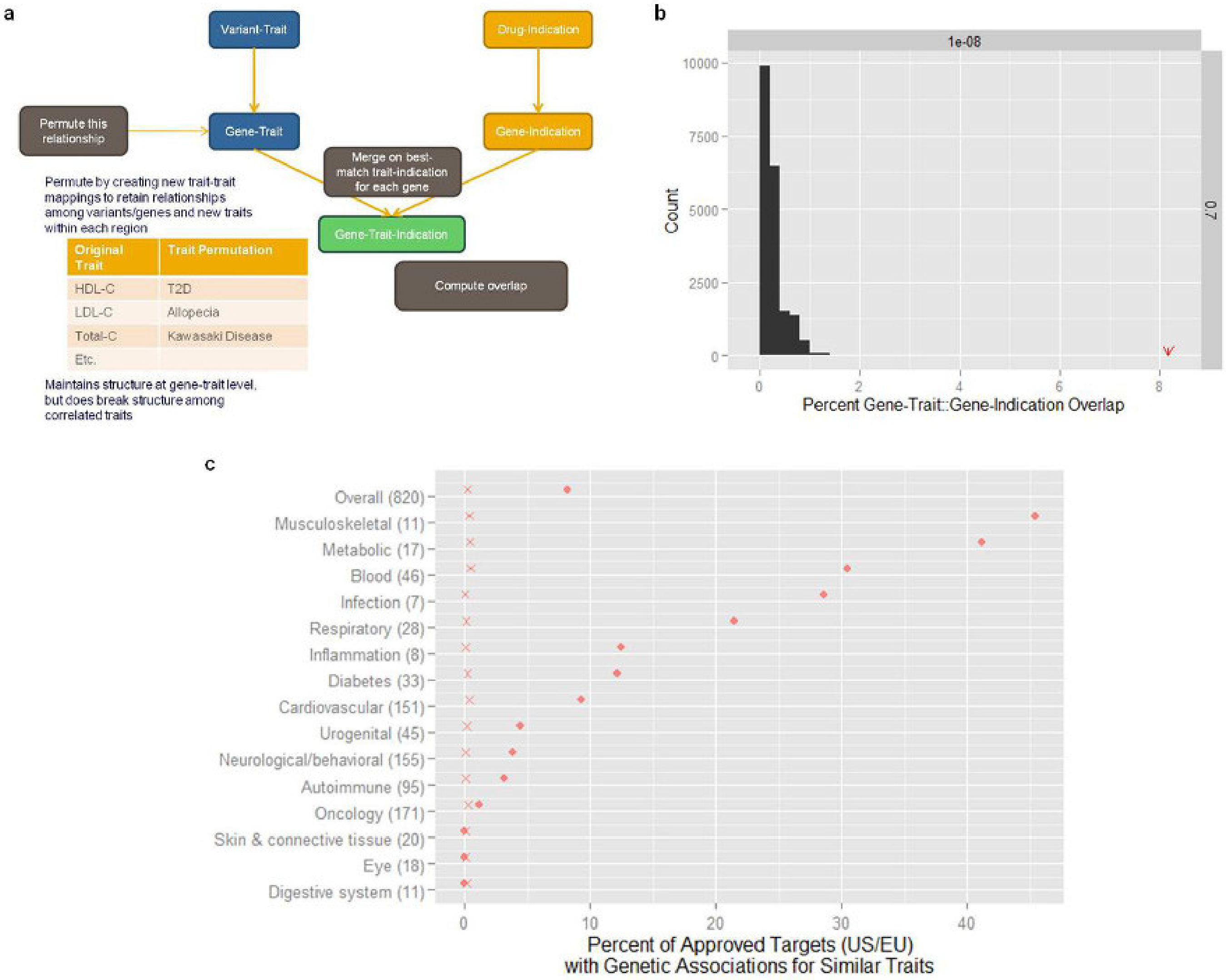
The few indications with very large numbers of genes associated were truncated at 50. (The full range is available in **Supplementary Table 5**.) The box corresponds to the interquartile range, the center line corresponds to the median, the whisker correspond to the maximum or 1.5 times the interquartile range (whichever is largest) and the points identify further outliers. The numbers given on the *y* axis are the number of unique indications observed in each phase. There is no statistically significant variability among phases (*P* = 0.18); analysis of the rank of the number of associations with phase as ordered variable) or with the linear trend (*P* = 0.37; analysis of rank of number of associations with phase as numeric with 1 = preclinical and 5 = approved in the United States or European Union).



Supplementary Figure 9

System for scoring the strength of evidence tying a variant with a phenotypic association to a gene.

For variant function, “DHS Rdb 3” indicates that the variant has a RegulomeDB score of 3 and falls within a proximal or distal DHS site, “eQTL or DHS (2)” indicates that the variant was either identified as an eQTL in the University of Chicago eQTL database or had a RegulomeDB score of 2 and “eQTL & DHS” indicates that the variant was both identified as an eQTL and fell within a DHS site with a RegulomeDB score of 2 or less. LD is in the form of r^2.



Original Trait	Trait Permutation
HDL-C	T2D
LDL-C	Allopecia
Total-C	Kawasaki Disease
Etc.	

Supplementary Figure 10

Permutation test of overlap between approved drug target–indications and genetic evidence (GWASdb or OMIM).

(**a**) The permutation scheme to simulate the null distribution. (**b**) The distribution of the percent of gene-trait and target-indication pairs that overlap over 10,000 permutations and the overlap observed in the original data (red downward arrow). (**c**) The overlap observed in the original data overall and by disease category (red points) and the median percent overlap over 10,000 permutations (red ×).

The support of human genetic evidence for approved drug indications
Nelson et al.

Supplementary Note

Capture and redundancy of drugs in the Pharmaprojects database

We assessed the capture of approved drugs in the Pharmaprojects database by investigating the presence of the top ten selling drugs in the US as reported in http://www.medscape.com/viewarticle/825053 in the data set. These drugs are listed below, along with the number of distinct drugs listed in Pharmaprojects that include the active drug.

Rank	Drug (Brand Name)	Drug Name	Drug Entries
1	Abilify	Aripiprazole	7
2	Nexium	Esomeprazole	8
3	Humira	Adalimumab	10
4	Crestor	Rosuvastatin	9
5	Advair Diskus	Fluticasone/salmeterol	6
6	Enbrel	Etanercept	23
7	Remicade	Infliximab	9
8	Cymbalta	Duloxetine	2
9	Copaxone	Glatiramer acetate	5
10	Neulasta	Pegfilgrastim	14

As shown, all of the top ten drugs are included in the final Pharmaprojects analysis data set in multiple forms. This full coverage confirms our high confidence that there are no major gaps in the target–indication content of our final data set. The full listing of drugs in our data set is included in Supplementary Table 4 for further inspection for any drug of interest. Based on the data selection criteria, any drug with a non–human target or an unknown target gene is not included.

Conditional analysis of OMIM, GWASdb, and RVIS lower quartile on successful drug targets

We observed a highly significant enrichment of genes with Mendelian and complex trait associations among successful drug targets. One potential explanation for this overlap is that the genes that have greater phenotypic effects when subjected to genetic modification may similarly be more likely to yield important physiological changes when their protein products are modulated therapeutically. We explored this hypothesis by investigating the relationship of these genes with residual variance intolerance score (RVIS), a measure of how tolerant a gene is to mutation (see reference in main manuscript). To address this, we fit the approval status of each coding gene with a logistic model with gene presence in OMIM, GWASdb (top), and RVIS percentile as independent variable. We restricted this analysis to only 16,701 GENCODE/RefSeq genes with observed RVIS values. We show below the summary statistics for complete and reduced general linear models and the conditional likelihood tests of significance of the RVIS percentile and genetic information in the model. From the complete model, we estimate the odds ratio (OR) for presence in OMIM to be 5.4, a top GWAS gene to be 1.6, and 0.99 for each RVIS percentile.

Complete model: Approved ~ OMIM + GWASdb + RVIS

```
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -4.181963   0.124087 -33.702  < 2e-16 ***
```

```
OMIMTRUE        1.677840   0.113516  14.781  < 2e-16 ***
GWASdb.TopTRUE  0.484017   0.128614   3.763 0.000168 ***
RVIS.Ptile     -0.008208   0.001989  -4.127 3.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 3266.8  on 16700  degrees of freedom
Residual deviance: 2997.7  on 16697  degrees of freedom
AIC: 3005.7
```

Reduced model 1: Approved ~ RVIS

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.360310   0.098044 -34.274  < 2e-16 ***
RVIS.Ptile  -0.011796   0.001984  -5.945 2.76e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 3266.8  on 16700  degrees of freedom
Residual deviance: 3230.2  on 16699  degrees of freedom
AIC: 3234.2
```

Reduced model 2: Approved ~ OMIM + GWASdb

```
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.57907    0.08553 -53.538  < 2e-16 ***
OMIMTRUE        1.73036    0.11284  15.335  < 2e-16 ***
GWASdb.TopTRUE  0.51548    0.12827   4.019 5.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 3266.8  on 16700  degrees of freedom
Residual deviance: 3015.1  on 16698  degrees of freedom
AIC: 3021.1
```

Significance of OMIM and GWASdb, conditioned on RVIS:

```
Model 1: Approved.US.EU ~ RVIS.Ptile
Model 2: Approved.US.EU ~ OMIM + GWASdb.Top + RVIS.Ptile
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     16699     3230.2
2     16697     2997.7  2   232.54 < 2.2e-16
```

Significance of RVIS, conditioned on OMIM and GWASdb:

```
Model 1: Approved.US.EU ~ OMIM + GWASdb.Top
Model 2: Approved.US.EU ~ OMIM + GWASdb.Top + RVIS.Ptile
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     16698     3015.1
2     16697     2997.7  1   17.442 2.962e-05
```

These reciprocal conditional analyses demonstrate that the explanatory ability of the genetic and RVIS percentile data are largely independent, with the genetic information, particularly OMIM, explaining much more of the variability in targets success.

Estimating the impact of genetic support on the probability of success and failure

Ultimately, we are interested in what the expected drug development success rate would be for drug mechanisms with genetic support versus those without it. The preferred approach to deriving this estimate would be through a statistical model that included a list of drugs that failed in development due to lack of efficacy and those that succeeded as the response variable and the presence or absence of genetic evidence as the primary independent variable of interest. However, there is a lack of sufficient historical information about trial failures due to lack of efficacy. We attempted to derive this data from Citeline Trialtrove® (http://www.citeline.com/products/trialtrove), a commercial database containing information on more than 127,000 drug trials. To identify trials that failed due to lack of efficacy, we queried Trialtrove for phase 2 or phase 3 trials where trial outcome was recorded, and the outcome was listed as completed, but primary outcome not met or terminated due to lack of efficacy. This resulted in only 224 trials indicated as having failed due to lack of efficacy, 68% of which were for oncology studies, followed by digestive system (8%) and musculoskeletal (7%). By comparison, only 15% of target–indication pairs for drugs approved in the US or EU are for oncology indications. Given such a small pool of negative studies (compared to 820 approved target–indication pairs), we abandoned this approach to estimating the probability of success given a genetic association.

An alternative approach is to take the information we have, the probability of a genetic association given a successful target–indication (P(G | Success)), and apply Bayes theorem to estimate the probability of a successful target–indication given a genetic association (P(Success | G)). We used the observed proportion of successful mechanisms with a genetic association (P(G | Success) = 0.082, as reported in the main text) and the overall proportion of unique target–indications from phase 1 through approved as our estimate of the P(G), which is 0.043. To estimate the overall probability of success, we used the results from two historical data analysis recently published[1, 2]. Based on these data, the overall probability that a drug will be approved is 0.19 for drugs that enter phase 1, 0.27 for phase 2, and 0.60 for phase 3. Given that efficacy is not generally a primary purpose of phase 1 studies and hence not a significant factor in determining which drugs will progress to phase 2, we used the phase 2 estimate of

0.27 for the purposes of this analysis. (Although, the marked increase in genetic associations supporting target–indications between phase 1 and phase 2 suggests that efficacy is an important factor in progression between those phases.)

$$\begin{aligned}\mathrm{P(Success\ |G)} &= \frac{\mathrm{P(G\ |Success)P(Success)}}{\mathrm{P(G)}}\\ &= \frac{(0.082)(0.27)}{0.043}\\ &= 0.517\end{aligned}$$

Hence, we estimate the overall probability of success to be approximately 52% under these assumed estimates of the marginal probabilities. Perhaps the measure of greatest interest is not so much the overall probability of success given genetic association, which is intrinsically sensitive to our estimate of the overall success rate, but rather the relative advantage of genetic support, versus without genetic support i.e.

$$\frac{\mathrm{P(Success\ |G)}}{\mathrm{P(Success|!\,G)}}.$$

We can estimate this probability just as we did above.

$$\begin{aligned}\mathrm{P(Success|!\,G)} &= \frac{\mathrm{P(!\,G|Success)P(Success)}}{\mathrm{P(!\,G)}}\\ &= \frac{(1-\mathrm{P(G|Success))P(Success)}}{1-\mathrm{P(G)}}\\ &= 0.259\end{aligned}$$

So the estimate of the ratio is 0.517/0.259 = 1.99 and we therefore estimate that the support of genetic information roughly doubles the overall success rate. It is important to note that this ratio is not influenced by our marginal estimate of the probability of success, since that term cancels out in the estimate of the ratio. We estimated the 95% confidence interval of this ratio using the `riskratio` function of the `epitools` R package with the bootstrap method and 50,000 replicates of 1.60 and 2.40. Hence, we estimate that the presence of genetic support doubles the probability of successful drug development with a small 95% confidence interval. The contingency table for this calculation, derived from Supplementary Table 7, is shown here.

```
            NotSuccess Success Total
NotGenetics       3252     753  4005
Genetics           112      67   179
Total             3364     820  4184
```

In addition to combined contribution of GWAS and OMIM to drug development success, we estimated the ratio of successful given genetic association versus successful in the absence of genetic information considering the GWASdb and OMIM information separately (Supplementary Table 7, see contingency tables below) in the same manner, shown in Table 1 of the main text.

Similarly, we can use this same approach to gain insight into the impact of the absence of genetic support on the failure of target–indication pairs to progress from phase 1 to approved. The ratio we estimate here using the probability estimates above is

$$\frac{\mathrm{P(!\,Success|!\,G)}}{\mathrm{P(!\,Success|G)}} = \frac{1 - \mathrm{P(Success|!\,G)}}{1 - \mathrm{P(Success|G)}}.$$

We computed these estimates and 95% confidence intervals by reversing the rows and columns of the shown contingency tables with the `riskratio` function in R with the GWASdb and OMIM data combined and separately.

Finally, we extended these calculations to estimate the relative contribution of genetic support to the progression of a target–indication pair from one drug development phase to another from the contingency tables shown below, reported in Table 1. Some caution is needed when interpreting these results, as we rely on the latest stage that a target–indication pair were reported to have progressed as a proxy for success and failure, and that efficacy is only one of many reasons why a drug could fail to progress.

The estimates and 95% confidence intervals are summarized in the following table:

	P(Progress | Genetic Support)/ (Progress | No Genetic Support)			**P(No Progress | No Genetic Support)/ P(No Progress | Genetic Support)**		
	GWASdb & OMIM	**GWASdb**	**OMIM**	**GWASdb & OMIM**	**GWASdb**	**OMIM**
Phase 1 to Phase 2	1.2 (1.1–1.3)	1.2 (1.1–1.3)	1.2 (1.1–1.3)	2.2 (1.5–3.4)	2.4 (1.5–4.5)	2.1 (1.3–4.5)
Phase 2 to Phase 3	1.5 (1.3–1.7)	1.4 (1.2–1.7)	1.6 (1.3–1.9)	1.4 (1.2–1.8)	1.4 (1.1–1.8)	1.6 (1.2–2.3)
Phase 3 to Approved	1.1 (1.0–1.2)	1.0 (0.8–1.2)	1.1 (0.9–1.3)	1.3 (0.9–2.0)	1.0 (0.7–1.8)	1.4 (0.9–2.9)
Phase 1 to Phase 3	1.8 (1.5–2.1)	1.8 (1.4–2.1)	1.9 (1.5–2.3)	1.4 (1.3–1.7)	1.4 (1.2–1.8)	1.6 (1.3–2.1)
Phase 1 to Approved	2.0 (1.6–2.4)	1.8 (1.3–2.3)	2.2 (1.6–2.8)	1.3 (1.2–1.5)	1.2 (1.1–1.4)	1.4 (1.2–1.7)

Phase 1 to Phase 2

```
GWASdb & OMIM
            NotSuccess Success Total
NotGenetics       1161    2844  4005
Genetics            24     155   179
Total             1185    2999  4184

GWASdb
            NotSuccess Success Total
NotGenetics       1172    2905  4077
Genetics            13      94   107
Total             1185    2999  4184

OMIM
```

```
            NotSuccess Success Total
NotGenetics       1174    2929  4103
Genetics            11      70    81
Total             1185    2999  4184
```

Phase 2 to Phase 3

```
GWASdb & OMIM
            NotSuccess Success Total
NotGenetics       1738    1106  2844
Genetics            66      89   155
Total             1804    1195  2999

GWASdb
            NotSuccess Success Total
NotGenetics       1763    1142  2905
Genetics            41      53    94
Total             1804    1195  2999

OMIM
            NotSuccess Success Total
NotGenetics       1778    1151  2929
Genetics            26      44    70
Total             1804    1195  2999
```

Phase 3 to Approved

```
GWASdb & OMIM
            NotSuccess Success Total
NotGenetics        353     753  1106
Genetics            22      67    89
Total              375     820  1195

GWASdb
            NotSuccess Success Total
NotGenetics        359     783  1142
Genetics            16      37    53
Total              375     820  1195

OMIM
            NotSuccess Success Total
NotGenetics        365     786  1151
Genetics            10      34    44
Total              375     820  1195
```

Phase 1 to Phase 3

```
GWASdb & OMIM
            NotSuccess Success Total
NotGenetics       2899    1106  4005
Genetics            90      89   179
Total             2989    1195  4184

GWASdb
            NotSuccess Success Total
NotGenetics       2935    1142  4077
Genetics            54      53   107
Total             2989    1195  4184

OMIM
            NotSuccess Success Total
NotGenetics       2952    1151  4103
Genetics            37      44    81
Total             2989    1195  4184
```

Phase 1 to Approved

```
GWASdb & OMIM
            NotSuccess Success Total
NotGenetics       3252     753  4005
Genetics           112      67   179
Total             3364     820  4184

GWASdb
            NotSuccess Success Total
NotGenetics       3294     783  4077
Genetics            70      37   107
Total             3364     820  4184

OMIM
            NotSuccess Success Total
NotGenetics       3317     786  4103
Genetics            47      34    81
Total             3364     820  4184


$OMIM.PhaseII
            NotSuccess Success Total
NotGenetics       1174    2929  4103
Genetics            11      70    81
Total             1185    2999  4184
```

Reference List

1. Arrowsmith,J. & Miller,P. Trial watch: phase II and phase III attrition rates 2011-2012. *Nat. Rev. Drug Discov.* **12**, 569 (2013).

2. Dimasi,J.A., Feldman,L., Seckler,A., & Wilson,A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin. Pharmacol. Ther.* **87**, 272-277 (2010).