nature genetics

The support of human genetic evidence for approved drug indications

Matthew R Nelson[1], Hannah Tipney[2], Jeffery L Painter[1], Judong Shen[1], Paola Nicoletti[3], Yufeng Shen[3,4], Aris Floratos[3,4], Pak Chung Sham[5,6], Mulin Jun Li[6,7], Junwen Wang[6,7], Lon R Cardon[8], John C Whittaker[2] & Philippe Sanseau[2]

Over a quarter of drugs that enter clinical development fail because they are ineffective. Growing insight into genes that influence human disease may affect how drug targets and indications are selected. However, there is little guidance about how much weight should be given to genetic evidence in making these key decisions. To answer this question, we investigated how well the current archive of genetic evidence predicts drug mechanisms. We found that, among well-studied indications, the proportion of drug mechanisms with direct genetic support increases significantly across the drug development pipeline, from 2.0% at the preclinical stage to 8.2% among mechanisms for approved drugs, and varies dramatically among disease areas. We estimate that selecting genetically supported targets could double the success rate in clinical development. Therefore, using the growing wealth of human genetic data to select the best targets and indications should have a measurable impact on the successful development of new drugs.

Attrition is a major challenge in drug discovery and development, with more than half of clinical studies failing because of lack of efficacy[1–4]. The widespread failure of preclinical model systems to adequately predict efficacy in humans has led drug developers to look for other sources of evidence to inform decisions about which targets to pursue and for which indications (disease or reason for treatment for which a drug is approved). Since the completion of the Human Genome Project and the rise of genome-wide association studies (GWAS) and whole-genome and whole-exome sequencing studies, there has been rapid progress in identifying the genes that influence human health and disease[5]. These genetic insights can potentially transform the process of selecting the best drug targets and indications[6], the key decisions in drug discovery. There are several examples of genes associated with disease traits that have been proven to be effective drug targets. One canonical example is the target for statins, *HMGCR*, which has been associated with serum cholesterol levels[7]. Several other examples were recently highlighted for rheumatoid arthritis[8]. Such examples and the rapidly growing body of human genetic data led us to ask how much weight should be given to genetic associations when choosing which drug targets to pursue for a desired indication.

RESULTS

In this study, we go beyond previous work on drug repositioning[9] to investigate how well clinically successful drug mechanisms (the protein product modulated to elicit a clinical response) are predicted by known genetic associations and how that prediction may change across the drug development pipeline, from preclinical and clinical phases to launched drugs (Drug Approval Process; see URLs). An overview of the data sources, filtering and processing applied is provided in **Figure 1a**. To broadly capture statistically significant ($P \leq 1 \times 10^{-8}$) common variant genetic associations, we used GWASdb[10], which combines data from multiple sources, including the National Human Genome Research Institute (NHGRI) GWAS Catalog, the tables and supplementary materials of manuscripts archived in the NHGRI GWAS Catalog, and the database of Genotypes and Phenotypes (dbGaP), among others. To allow comparisons among all data sources, we manually mapped all traits to the most specific Medical Subject Heading (MeSH) terms applicable. Genetic variants were mapped to potential causal genes using a combination of linkage disequilibrium (LD), position, expression quantitative trait locus (eQTL) and epigenetic data (for example, see **Fig. 1b**). When we observed multiple possible variant-to-gene mappings, these were ranked on the overall strength of evidence. In the final data set, we had 18,566 genetic associations to 434 MeSH traits that mapped to 6,120 genes outside of the extended major histocompatibility complex (xMHC), with a total of 13,855 gene-trait combinations. Genes involved in rare, mendelian traits were derived from Online Mendelian Inheritance in Man (OMIM), providing a data set with 1,898 genes annotated to affect 2,145 traits with MeSH terms, for a total of 2,627 gene-trait combinations. The GWASdb and OMIM gene-MeSH pairs were largely non-overlapping, yielding a combined set of 16,459 gene-trait combinations (**Supplementary Fig. 1** and **Supplementary Table 1**).

[1]Quantitative Sciences, GlaxoSmithKline, Research Triangle Park, North Carolina, USA. [2]Quantitative Sciences, GlaxoSmithKline, Stevenage, UK. [3]Department of Systems Biology, Columbia University, New York, New York, USA. [4]Department of Biomedical Informatics, Columbia University, New York, New York, USA. [5]State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, SAR, China. [6]Centre for Genomics Sciences, Li Ka Shing (LKS) Faculty of Medicine, University of Hong Kong, Hong Kong SAR, China. [7]Department of Biochemistry, LKS Faculty of Medicine, University of Hong Kong, Hong Kong SAR, China. [8]Alternative Discovery and Development, GlaxoSmithKline, Upper Merion, Pennsylvania, USA. Correspondence should be addressed to M.R.N. (matthew.r.nelson@gsk.com).

Received 23 September 2014; accepted 1 May 2015; published online 29 June 2015; doi:10.1038/ng.3314

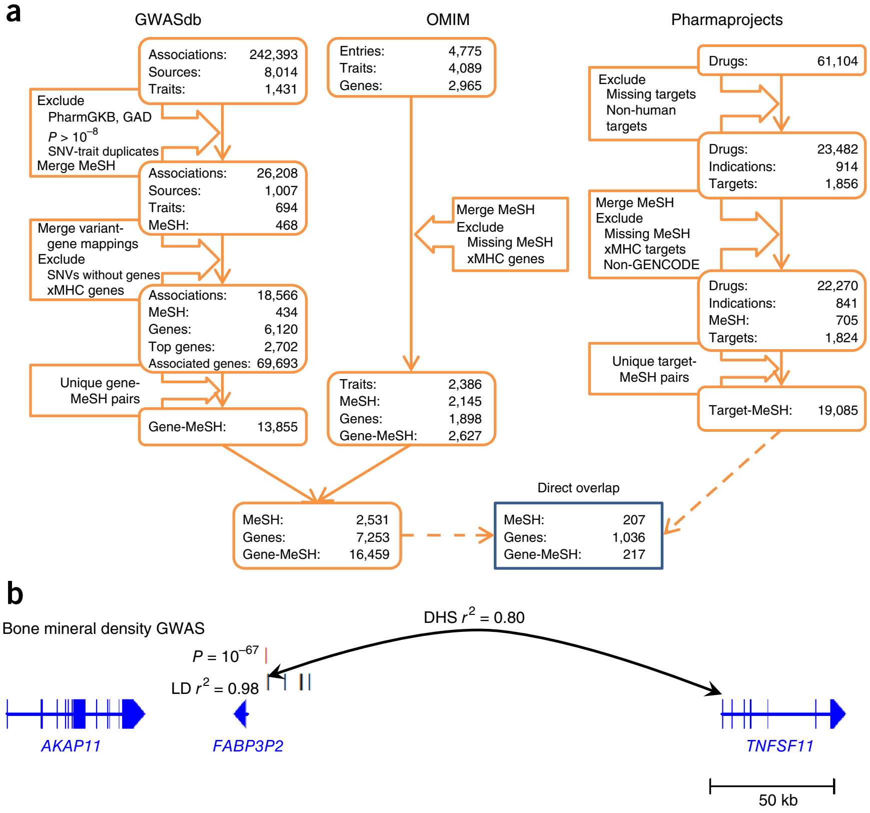

Figure 1 Summary of data resources and mappings between them. (**a**) Summary of each data resource and the key filtering and processing steps applied to create the final set of gene-trait and drug target–indication combinations investigated in this study. GWASdb sources correspond to unique PubMed IDs or other unique data sources given for each association. GAD, Genetic Association Database. (**b**) An example of the approach to map genetically associated variants to genes, illustrated with the bone mineral density GWAS association with rs9533090 (depicted in red). Of five SNPs in strong LD with rs9533090 ($r^2 \geq 0.8$), one falls within a DNase I–hypersensitive site (DHS) that was found to have a sensitivity signal correlated with the DHS of the *TNFSF11* gene transcription start site (TSS).

Information about drugs across the various stages of development was drawn from the commercial Informa Pharmaprojects database. Of a total of 61,104 drugs (including combination therapies; **Supplementary Note**), there were 22,270 drugs known to modulate 1,824 human non-xMHC drug targets for 705 indications, giving a total of 19,085 target-indication pairs (**Supplementary Fig. 2** and **Supplementary Tables 2–4**). Aggregation of the drug information at the target and indication levels eliminated redundancies in drug mechanisms within the database, such as multiple formulations of the same drug or multiple drugs within the same drug class used to treat the same indications.

We found that the target genes for drugs approved in the United States or the European Union, our definition of 'successful drug mechanisms', were significantly enriched among genes associated with variation in human traits (**Fig. 2**). The greatest enrichment was for genes identified using OMIM (odds ratio (OR) = 7.2, $P = 8.9 \times 10^{-74}$), where 206 of 389 (53%) target genes for approved drugs were also associated with a mendelian trait, a proportion comparable to that in a previous report[11]. Genes associated with traits through genome-wide associations were also significantly enriched (OR = 2.0, $P = 2.9 \times 10^{-10}$), particularly when genes were limited to the top-ranked gene for each associated variant (OR = 2.7, $P = 1.3 \times 10^{-14}$), with 98 (25%) genes in common. However, we also observed that genes considered to be classically druggable, having binding domains for small molecule drugs[12] (n = 2,639), were also highly enriched among OMIM and GWASdb genes (OR = 1.9 and 1.7, respectively). To account for this relationship, we also assessed the enrichment of genetic associations within the druggable subset of the genome. In this analysis, there was decreased but still highly significant ($P < 1 \times 10^{-3}$) enrichment of the OMIM and top GWASdb genes (OR = 4.5 and 1.6, respectively). There was little added enrichment when considering the combined effects of OMIM and GWASdb. One potential explanation for the correlation between successful drug targets and evidence of genetic effects is that genes that result in notable phenotypic changes when altered are also the most responsive to drug-induced alterations. The greater enrichment among successful targets of genes that give rise to mendelian disorders in comparison to those involved in complex traits supports this explanation. Residual variance intolerance score (RVIS) was recently developed to assess the tolerance of a gene to mutational perturbation[13]. We observed a statistically significant association between genes falling within the lower quartile of the RVIS distribution (most intolerant to change) and approved drug status (OR = 2.1, $P = 7.7 \times 10^{-10}$). However, conditioning on RVIS had little impact on the effect of OMIM and GWASdb association status and hence is an independent predictor of target success and not an explanation for the effect of genetic associations (**Supplementary Note**).

The analysis above did not take into account alignment between the drug indications and the associated traits. Therefore, we next

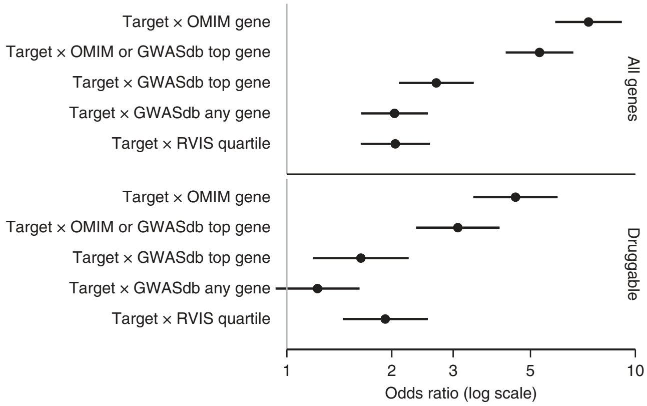

Figure 2 Enrichment of target genes for drugs approved in the United States or the European Union. Associations are shown for all 22,012 coding genes (top half) and for 2,555 classically druggable genes (bottom half). Target enrichment is estimated for genes in OMIM, genes with any connection with a GWASdb association, only the top gene for each GWASdb association, genes in OMIM or the top GWASdb gene, and genes in the lower quartile of the RVIS distribution. Odds ratios with exact 95% confidence intervals were estimated from 2 × 2 tables of the status of each gene as a drug target versus status in the above categories (Online Methods).

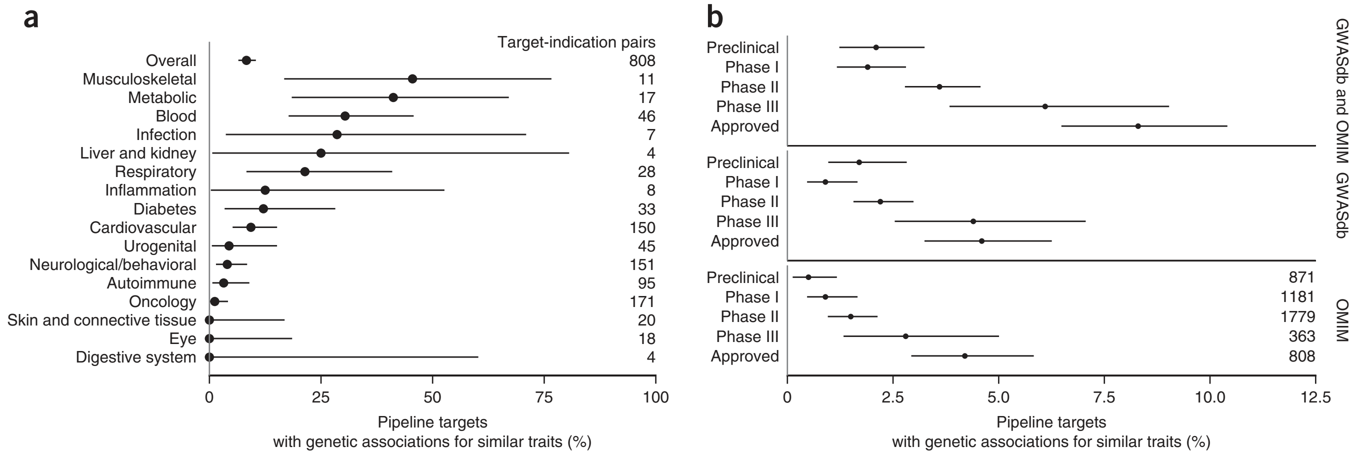

Figure 3 Overlap between drug targets and their indications with genetic associations for similar traits. (**a**) The percentage of target-indication pairs for drugs approved in the United States or the European Union overlapping with gene-trait combinations from GWASdb or OMIM. Indications and traits were considered similar if the relative similarity was at least 0.7. The percent overlap is shown for all 820 pairs ("Overall"; 158 unique indications and 265 targets) and for pairs stratified by the indication therapy category. Indications are limited to those with at least five genes in GWASdb or OMIM associated with a similar trait. Exact 95% confidence intervals are shown, along with the number of target-indication pairs, for each row. The enrichment was further assessed via permutation testing and was found to be highly statistically significant ($P < 1 \times 10^{-5}$; **Supplementary Fig. 9**). (**b**) Percent overlap for all target-indication pairs at the latest stage reached in the development pipeline. Overlap is given for GWASdb, OMIM and GWASdb, and OMIM combined. All other features are as in **Figure 1a**.

investigated the percentage of approved target–indication pairs with a corresponding genetic association tied to the same gene for a similar trait. Using the structure of the MeSH hierarchy to estimate indication-trait similarity[14] (**Supplementary Fig. 3**), we found that 239 of 395 (61%) approved drug indications had at least 1 genetic association (OMIM or GWASdb) with a similar trait (relative similarity ≥ 0.7) and that 158 (40%) approved indications had at least 5 associations reported. The approved drug indications having fewer than five genetic associations—such as anxiety, depression, headache, coronary restenosis and kidney stones—included both diseases where many studies have been done with little success and understudied areas of medical interest currently lacking substantial genetic investigation (**Supplementary Table 5**).

To assess the support that a genetic association provides to drug mechanisms, we focused on the subset of 158 approved drug indications with at least 5 genetic associations for a similar trait, taking this to signify that the indication has been reasonably well studied by genetic approaches (that is, focusing on instances where an opportunity exists for genetic data to support the target indication; **Supplementary Table 5**). Of 820 target-indication pairs, 67 (8.2%) were supported by one or more genetic associations when considering the combined evidence of both OMIM and GWASdb (**Fig. 3a** and **Supplementary Table 6**). Further, we found that there was significant variability among indication categories ($P = 1.1 \times 10^{-16}$; **Fig. 3a**), with the highest degree of genetic support for indications related to musculoskeletal, metabolic and blood categories (percent overlap of greater than 30%) and little or no genetic support for oncology, skin, eye and digestive categories. We observed that there was slightly greater support with GWASdb than with OMIM (4.5% versus 4.1%, respectively; **Fig. 3b**, **Supplementary Figs. 4–6** and **Supplementary Table 7**), although the overlap with OMIM represented a much larger fraction of the total number of OMIM gene-trait associations in comparison to GWASdb (1.2% versus 0.27%, respectively). These results were somewhat sensitive to restricting the indications to those that had varying levels of genetic support, although a cutoff of at least five associations per indication yielded the best tradeoff between the number of indications considered and overall genetic support (**Supplementary Fig. 7**).

If genetic association data are predictive of successful mechanisms of action, then we would expect the percent of target-indication pairs with genetic evidence to increase the further the corresponding drug has progressed in the drug development pipeline, with approval representing a mechanism that has passed the highest evidentiary standards. This is just the pattern that we observed when considering OMIM and GWASdb together or separately (**Fig. 3b**), where in each instance the enrichment of genetic support for target-indication pairs was the lowest in phase I and increased in subsequent phases through drug approval. The genetic support increased from 2.0% for target-indication pairs that had only progressed as far as phase I clinical trials to 8.2% for approved drugs, over a fourfold increase, suggesting that the odds of successful drug mechanisms with genetic support are many times greater than without. For new mechanisms in early development, we cannot rule out the influence that relatively recent GWAS may have had on the choice of targets and indications; however, accounting for such an influence would lead to an upward bias in the estimated overlap at that early stage and a downward bias in the increase in enrichment with progression. It is also possible that the reporting of successful drug mechanisms has influenced some gene-trait annotations that have been added to OMIM, although an informal review of several entries did not find this to be a likely contributor. The enrichment of genetic support we observe here is consistent with a recent AstraZeneca portfolio review[3]. Among 38 phase II programs, an OR of 3.5 (95% confidence interval (CI) = 0.73–20.6, $P = 0.10$) was observed in comparing the genetic support for projects that progressed to that for projects that did not.

DISCUSSION

On one hand, there are limitations to the ability to identify the genes that are causally related to a genetic association, which, given our inclusive strategy to map all possible causal genes, could inflate our estimate of the proportion of successful drug mechanisms with genetic support. On the other hand, the information available about the functional genomic landscape is incomplete, and there will be many causal relationships left undetected or ascribed to the wrong gene, resulting in a bias of the enrichment estimates toward the null. However, the growing body of functional genomic information will continue to improve the ability to correctly ascribe a molecular pathway by which genetically associated variants influence traits. Such data can also help identify the causal mechanism underlying the association and inform what treatments could lead to a positive outcome in patients. In addition, catalogs of genetic variants that influence human traits are far from complete, which would lead to an underestimation of the proportion of drugs with genetic support. We have identified a number of therapeutic areas where there are large gaps in knowledge about the genetic factors involved, divided evenly across the pipeline (**Supplementary Fig. 8**). We advocate continued support for research on the genetics of these areas to aid in the development of more effective treatments. The availability of a precompetitive genetic resource similar to that produced for the purposes of this analysis that integrates all known genetic associations with measures of statistical confidence, using a common trait ontology, and integrates the most recent sources of functional genomic information to list and rank potential causal pathways would be an invaluable tool for the drug discovery process.

Another potential source of bias is that genetic associations could already be driving decisions on which drugs make it into clinical development and for which indications. Although this would have affected a small subset of the historical drug data, given that drug discovery and development timelines generally extend back well over 10 years, the impact of this bias would be to increase the proportion of drugs with genetic evidence earlier in the pipeline, leading to an underestimation of the relative benefit of genetic support. There may also be instances where known mechanisms for drugs could lead to targeted genetic research that finds supporting information, which would disproportionately affect the overlap with approved drugs. We would not expect these biases to measurably affect the GWAS-based results. However, there is greater potential for the manually curated results in OMIM to influence target selection or for drug targets to influence genetic research. We reviewed the 39 OMIM genes and traits that overlapped approved drug targets and indications (**Supplementary Table 6**) and found several potential instances where genetic information led to the development of therapeutics, including use of the gene product as a therapeutic, as in the case of von Willebrand disease where von Willebrand complex is used in treatment. This finding partially explains the greater overall enrichment of targets associated with traits in OMIM.

Ultimately, we want to know the probability that a therapeutic agent that properly engages the target protein at safe and efficacious doses in the relevant tissues will have the intended effects to prevent or treat disease in patients[3,4]. Several pieces of information required for a thorough analysis are missing from the public domain; most notably, there are relatively few data available on drugs that failed in clinical development and the reasons for these failures (**Supplementary Note**). However, with the historical information available on drug and, hence, target-indication progression through the clinical pipeline, we can derive estimates of the value the support of genetic information brings. Given the observations in our data, we estimated the ratio of the probability of progressing in the drug development pipeline given that the drug mechanism has the support of genetic information to the probability of the drug progressing without genetic support (**Table 1** and **Supplementary Note**), where we considered support from GWASdb and OMIM in combination as well as separately. OMIM support yielded a slightly higher probability of success than GWASdb support. We estimated that genetic support had the largest impact on the probability of progressing from phase II to phase III (ratio = 1.5, combined), with the next largest impact for progression from phase I to phase II (ratio = 1.2, combined); the smallest apparent contribution was for progression from phase III to approved status (ratio = 1.1, combined). We also estimated the converse ratio of the probability of failure to progress in the absence of genetic support versus with support (**Supplementary Note**). As expected, we found that, overall, target-indication pairs that entered clinical development that lacked genetic support were significantly less likely to reach drug approval (ratio = 1.3, 95% confidence interval = 1.2–1.5, combined), and the lack of genetic support in progression had the greatest impact earlier in the drug development process.

Table 1 The relative value of genetic support for the probability that a target-indication pair progresses along the drug development pipeline, based on historical drug trial information

	p(progress | genetic support)/(progress | no genetic support)		
Progression	GWASdb and OMIM	GWASdb	OMIM
Phase I to phase II	1.2 (1.1–1.3)	1.2 (1.1–1.3)	1.2 (1.1–1.3)
Phase II to phase III	1.5 (1.3–1.7)	1.4 (1.2–1.7)	1.6 (1.3–1.9)
Phase III to approval	1.1 (1.0–1.2)	1.0 (0.8–1.2)	1.1 (0.9–1.3)
Phase I to phase III	1.8 (1.5–2.1)	1.8 (1.4–2.1)	1.9 (1.5–2.3)
Phase I to approval	2.0 (1.6–2.4)	1.8 (1.3–2.3)	2.2 (1.6–2.8)

Values give the ratio of the probability of a target-indication pair progressing given genetic support to the probability of progressing without genetic support; 95% confidence intervals are given in parentheses.

The relatively low impact of genetic support on success in phase III is surprising, given that attrition rate estimates attribute most phase III failures to lack of efficacy[2]. It may be that failures in phase III are different in nature from those in earlier stages, for example, because they may reflect a failure to improve over standard of care rather than failure of the targeted biological mechanism to be causal for disease at all. Or it may be that, in phase III, study endpoints are more complex and less closely related to specific biological mechanisms, including the use of broad endpoints such as all major coronary events in cardiovascular outcome studies. In addition, we note the limitations of the available data. We rely on the latest stage to which a target-indication pair was reported to have progressed as a proxy for success and failure, although such data may be incomplete or even inaccurate in some cases. Furthermore, the interpretation of risk ratios is dependent on the absolute risk, which varies substantially by phase.

Overall, we estimate that drug mechanisms with genetic support would succeed twice as often as those without it (from phase I to approval). Therefore, increasing the proportion of discovery and development activities focused on targets with genetic support and allowing genetic data to guide selection of the most appropriate indications should lead to lower rates of failure due to lack of efficacy in clinical development.

URLs. Drug Development Process, http://www.fda.gov/downloads/Drugs/ResourcesForYou/Consumers/UCM284393.pdf; GWASdb, http://jjwanglab.org/gwasdb; Online Mendelian Inheritance in Man (OMIM), http://www.omim.org/; MeSH browser, https://www.nlm.nih.gov/mesh/MBrowser.html; UMLS::Similarity, http://www.d.umn.edu/~tpederse/umls-similarity.html; PharmGKB, https://www.pharmgkb.org/; Genetic Association Database, http://geneticassociationdb.nih.gov/; Informa Pharmaprojects database, http://www.citeline.com/; MeSH thesaurus, http://www.nlm.nih.gov/mesh.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank N. Srivastava for much of the manual mapping of GWASdb traits and Pharmaprojects indications to MeSH terms and P. Agarwal for many helpful conversations. GWASdb-related work was supported by the Research Grants Council, Hong Kong SAR, China (781511M, 17121414M) and the National Natural Science Foundation of China (91229105).

AUTHOR CONTRIBUTIONS

This work was conceived by M.R.N., H.T., L.R.C., J.C.W. and P.S. The primary analyses were designed and conducted by M.R.N. Supporting analyses were provided by J.L.P. and J.S. The mapping of variants to genes was conducted by M.R.N., P.N., Y.S. and A.F. The GWASdb data were created and provided by P.C.S., M.J.L. and J.W. The manuscript was written by M.R.N. with contributions from H.T., J.C.W. and P.S.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. 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).
2. Arrowsmith, J. & Miller, P. Trial watch: phase II and phase III attrition rates 2011–2012. *Nat. Rev. Drug Discov.* **12**, 569 (2013).
3. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–431 (2014).
4. Morgan, P. *et al.* Can the flow of medicines be improved? Fundamental pharmacokinetic and pharmacological principles toward improving Phase II survival. *Drug Discov. Today* **17**, 419–424 (2012).
5. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
6. Plenge, R.M., Scolnick, E.M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
7. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
8. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
9. Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30**, 317–320 (2012).
10. Li, M.J. *et al.* GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* **40**, D1047–D1054 (2012).
11. Wang, Z.Y. & Zhang, H.Y. Rational drug repositioning by medical genetics. *Nat. Biotechnol.* **31**, 1080–1082 (2013).
12. Hopkins, A.L. & Groom, C.R. The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730 (2002).
13. Petrovski, S., Wang, Q., Heinzen, E.L., Allen, A.S. & Goldstein, D.B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9**, e1003709 (2013).
14. 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).

ONLINE METHODS

Genetic data. Genetic association data were drawn from the data available in GWASdb[10] (version dated 21 May 2013), a manually curated database that brings together information from eight sources. We excluded all data from PharmGKB and the Genetic Association Database. Genetic associations reported from these two sources contained no supporting statistical association evidence (with most P values equal to zero) to accompany the entries, and the new associations included were largely drawn from candidate gene association studies that lacked rigorous criteria for reporting a statistical association. In particular, we found that there were a large number of candidate gene associations in PharmGKB for drug target genes, which would result in an upward bias in the number of drug targets with supposed genetic associations. We also excluded a few large metabolomic studies with numerous traits screened that had very large numbers of associations reported. Finally, we identified one study[15] where a supplementary table was misinterpreted, leading to many falsely identified associations that were also excluded. For the variants, traits and P values reported, we removed any duplicate entries found across the various GWASdb data sources. For the purposes of this study, we set a P-value threshold of 1×10^{-8} to limit associations to those with relatively strong evidence. The OMIM database (accessed 3 October 2013) was used to provide additional information on the effects of genetic variants and mutations on human traits. Only entries with valid MeSH terms were included in the analyses reported here.

Genetic variant-to-gene mapping. Variants with phenotypic associations were mapped to the genes that they could be causally affecting through a combination of approaches. First, all variants in LD having $r^2 \geq 0.5$ with each associated variant were identified on the basis of the 1000 Genomes Project pilot sequence genotypes for the European-ancestry (CEU) population[16]. No effort was made to conduct LD pruning to represent independent associations as the purpose of our study was to identify all possible genes that could be responsible for the observed effect. For each variant in LD, the plausible mapping of a variant to a particular gene was performed using a combination of physical proximity to the gene, evidence for association of the variant with the expression of the gene and determination of whether the variant fell within a regulatory element predicted to affect the expression of the gene. The variant was mapped to the physical location of the gene plus or minus 5 kb on the basis of the longest gene transcript to define the gene boundaries plus 1.5 kb in UCSC-distributed RefSeq (v37.1) annotation. Gene eQTLs were drawn from eqtl.chicago.edu (accessed 21 May 2013), which includes eQTLs from several studies of several cell lines and primary tissues as well as the results from primary liver tissue[17] at false discovery rate (FDR) ≤ 0.1, computed by Kruskal-Wallis test. To map variants to genes on the basis of regulatory evidence, we identified all variants that fell within a predicted transcription factor binding site located within a DHS peak using RegulomeDB[18] (accessed 7 February 2013). For variants with a RegulomeDB score ≤4, we determined whether the genomic location overlapped a DHS peak that was either located with a gene TSS or a distal DHS peak that was correlated with a TSS DHS across cell lines, as described[19] (data courtesy of J. Stamatoyannopoulos, University of Washington). Variants that affected the amino acid sequence of any gene transcripts were identified via the Ensembl Variant Effect Predictor from the European Bioinformatics Institute (EBI; accessed 27 February 2014). We restricted our analyses to genes reported in GENCODE (v17) or RefSeq (v37.1).

In many instances, a variant with a phenotypic association could be mapped to more than one gene using this combination of approaches. We devised an ad hoc scoring scheme to assess the relative weight of evidence for a causal relationship between the variant reported to be associated and each gene to which it was mapped (**Supplementary Fig. 9**), including the source of the association, the LD between the associated variant and the variant mapped to the gene, the nature of the mapping information and the number of times that the variant in LD had been associated with the trait. This scheme yielded a potential gene score between 0 and 11, with 11 reflecting the strongest evidence. The factors included in the gene scoring scheme were also used to rank the variant-to-gene mappings, such that the top-ranked gene for a particular variant presumably had the strongest evidence (**Fig. 2**). When two gene mappings had equal support, the ranking was arbitrarily decided.

Drug data. Information about drugs, their gene targets, the indications for which they have been investigated and the latest stage of development to which they have progressed was derived from the commercial Informa Pharmaprojects database. Drugs were retained for analysis if (i) they were annotated to have human gene targets (on the basis of GENCODE v16), (ii) the gene did not map to the xMHC and (iii) the indication could be mapped to a MeSH term. Most analyses using Pharmaprojects were conducted using a transformation of the data into a single entry per gene target and indication with the latest phase in development to which that unique combination progressed for any drug. A target was defined as successful in treating an indication if a drug targeting that gene product was approved for the corresponding indication in the United States or the European Union, as annotated in Pharmaprojects.

Medical Subject Heading term mapping and use. We used the MeSH thesaurus to provide a common vocabulary among traits from GWASdb and OMIM and indications from Pharmaprojects. MeSH term mappings to OMIM traits was derived from Comparative Toxicogenomics Database mapping[20]. Mappings for GWASdb and Pharmaprojects were performed manually using the MeSH Browser by searching with each of the unique original terms listed in the respective database and identifying the overall best match. Some traits did not yield a satisfactory MeSH term. Any data entries missing MeSH terms were excluded from the primary analyses described in this study.

When comparing the overlap between traits with respect to evidence for genetic association and drug indications, we recognized that there could be many instances where the genetic evidence was for a trait very closely related to the indication but not an exact match. To allow for such near misses, we used similarity measures based on the MeSH ontology, implemented in the UMLS::Similarity Perl module[14]. Several measures of similarity and relationships are implemented in this package. We evaluated all of these measures on a subset of 50 randomly selected MeSH entries from our combined data set to assess how well the subsequent trait clustering reflected expert interpretation. On the basis of this evaluation, we selected two similarity measures that incorporated both path distance and information content, Resnik[21] and Lin[22]. The measures were standardized to a measure of relative similarity between zero and one and averaged together to yield a final relative similarity measure for subsequent analysis. We noted that in some instances, because of the structure of the MeSH ontology, very closely related traits resulted in very low measures of similarity. Two examples are systolic or diastolic blood pressure with hypertension and bone mineral density with osteoporosis. To address this, we reviewed the laboratory-based MeSH terms and manually assigned relative similarity scores of 0.5, 0.7 and 0.9 on the basis of the known relationships between traits. The two examples above were assigned a relative similarity of 0.9. The manually assigned relative similarities are given in **Supplementary Table 8**. The relative similarity matrix used for the analyses is available in **Supplementary Data Set 1**. Each MeSH term was subsequently manually mapped to 1 of 20 disease categories (**Supplementary Table 9**).

Genetic association enrichment. We assessed enrichment of genetic associations both without and with respect to the trait underlying the association. We assessed enrichment without respect to trait or indication as presented in **Figure 2** by constructing a 2 × 2 table of genes in GENCODE (v17) or RefSeq (v37.1) and counts corresponding to the presence or absence of the gene as a target for a drug approved in the United States or the European Union versus the presence or absence of evidence for genetic association for each gene. Evidence of genetic association was further stratified by OMIM, any possible gene for each GWASdb association and the top gene (top ranked, as described above) for each GWASdb association. Enrichment for RVIS was based on published scores[13], with stratification for the lowest quartile. The druggable genome was based on the description of Hopkins and Groom[12]. Odds ratios and 95% confidence intervals were estimated using the exact method implemented in fisher.exact in R.

The overlap between genetic evidence and drug targets presented in **Figure 3**, taking traits and indications into account, was based on the direct overlap of gene and target names with a relative trait-indication similarity of at least 0.7. The confidence intervals presented were computed using the Pearson-Klopper exact method implemented in the binom package in R. A permutation test (**Supplementary Fig. 10**) was performed to assess the significance of the

doi:10.1038/ng.3314

observed overlap given the high degree of correlation among genes and traits in the data. In the permutation test, the null distribution was simulated by breaking the relationships between traits and genes in the genetic association data. This was done in a manner to maintain the relationships among genes associated with the same trait by permuting the traits and replacing all associations for the observed trait with the same permuted trait (for example, by replacing all genes originally associated with alopecia with those associated with type 2 diabetes in permutation 1, with those associated with Kawasaki disease in permutation 2, etc.). We conducted 10,000 replicates.

All statistical analyses were conducted using R version 3.1.0 (ref. 23). Most figures were created using the R package ggplot2 (ref. 24).

Code availability. The R scripts and Sweave files used to process the data and conduct the analyses described herein are available from the authors by request. All key analyses can be reproduced from **Supplementary Data Sets 1–4** and the supplementary tables included.

15. Patsopoulos, N.A. *et al.* Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.* **70**, 897–912 (2011).
16. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
17. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
18. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
19. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
20. Davis, A.P. *et al.* The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Res.* **41**, D1104–D1114 (2013).
21. Resnik, P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* **11**, 95–130 (1999).
22. Lin, D. in *Proc. Int. Conf. Machine Learning* 296–304 (Morgan Kaufmann Publishers, 1998).
23. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
24. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).

doi:10.1038/ng.3314