feature

Pathway analysis of GWAS loci identifies novel drug targets and repurposing opportunities

Deepali Jhamb, Michal Magid-Slav, Mark R. Hurle and Pankaj Agarwal, Pankaj.agarwal@gsk.com

Genome-wide association studies (GWAS) have made considerable progress and there is emerging evidence that genetics-based targets can lead to 28% more launched drugs. We analyzed 1589 GWAS across 1456 pathways to translate these often imprecise genetic loci into therapeutic hypotheses for 182 diseases. These pathway-based genetic targets were validated by testing whether current drug targets were enriched in the pathway space for the same indication. Remarkably, 30% of diseases had significantly more targets in these pathways than expected by chance; the comparable number for GWAS alone (without pathway analysis) was zero. This study shows that a systematic global pathway analysis can translate genetic findings into therapeutic hypotheses for both new drug discovery and repositioning opportunities for current drugs.

Introduction

Decades of research into the genetic basis of human disease has shown that most common diseases are polygenic in nature. Hence, understanding the pathways or mechanisms by which these genes interact at a pathway or systems level is essential. Previous studies have shown that GWAS genes are enriched for known drug targets [1,2]; however, most GWAS targets are not druggable using either small molecules or biopharmaceuticals. In contrast to a single GWAS locus, single gene perspective, we hypothesized that integrating GWAS loci with biological pathways could add value for drug discovery by identifying drug targets and novel pathways enriched in genetic risk loci.

GWAS has made substantial progress in identifying susceptibility loci associated with complex traits and diseases [3]. Pathways [4,5], networks [6–8], gene expression [9], literature mining [10,11], and gene set enrichment [12] have been used to determine the functional relevance of GWAS variants. We hypothesized that the connection between GWAS hits and drug targets might be more easily discernible in pathway space because it is likely that both disease genetics and drug targets for a disease point to the same pathways. Published pathway analysis has concentrated mainly on individual or a small group of GWAS; the only large-scale analysis published so far [13] aimed at linking GWAS phenotypes based on shared KEGG pathways.

Pathway enrichment for GWAS studies

To begin, we repeated an earlier analysis [1] to see whether all the significant GWAS genes (see Workflow section) in the pathway universe were enriched overall for drug targets. We observed 285 GWAS genes that were drug targets [expected = 217, odds ratio = 1.31 (1.12–1.55), two-sided Fisher's exact test, $P = 6.2e{-}9$]. Thus, GWAS genes within the pathways were indeed enriched for drug targets. However, a major limitation of this calculation was that it did not consider whether the GWAS traits matched the indication for these drug targets. If we limited the analysis to the GWAS trait matching the drug indication, the enrichment was not detectable (Table S1 in the supplemental information online). An earlier study detected enrichment across disease classes if both Mendelian and common disease genes were grouped and related diseases were considered as matched [2].

To investigate our hypothesis, we first obtained 1589 GWAS from STOPGAP, a comprehensive

1359-6446/
https://doi.org/10.1016/j.drudis.2019.03.024

resource of genetic associations [14]. We processed the data using the pipeline summarized in Fig. 1. A crucial step in interpreting GWAS results is to map variants to their effector genes. Causal genes for a GWAS loci are frequently identified using expression trait loci (eQTLs) in disease-relevant tissue or coding variants in high linkage disequilibrium (LD). Unfortunately, only a small percentage of causal genes can be identified with these methods because eQTL data are not available for many cell types and few loci have missense-coding variants [15]. In addition, the causal genes picked by these methods are not always accurate. To avoid the bias of picking a single effector gene, we considered all 'candidate effector genes' [14] associated with all variants in high LD as having an $r^2 \geq 0.7$. Overall, we obtained 13 172 genes associated with 278 traits. During pathway enrichment, for each trait, we adjusted the number of genes mapped from a variant to ensure that at most one gene from each locus was counted (see Workflow section) [16]. This lenient variant to gene mapping assumed that the use of pathway information would implicitly enrich for the correct GWAS target.

For the pathway enrichment, MetaBase [17] was used to obtain a collection of 1456 manually curated pathways. These pathways included the major functional categories of cell signaling and human metabolism. A hypergeometric *P* value (as described in the Workflow section) was used to identify enriched pathways for each of the 278 traits. As an example, 103 genes were identified to be associated with Asthma risk loci ($r^2 \geq 0.7$, as described above). These 103 genes were used for pathway enrichment and resulted in 19 enriched pathways at false discovery rate (FDR) $P < 0.1$. Overall, 202 of the 278 GWAS traits were enriched for one or more pathways at FDR $P < 0.1$ (Table S2 in the supplemental information online). Of these 202 traits, 20 were not diseases and were excluded from further analysis. Most of these 182 diseases showed a significant enrichment for ten or less pathways (Fig. S1 in the supplemental information online). However, some of the diseases, related to inflammation, the digestive system, and metabolism, were associated with many significant pathways, possibly highlighting the role of several mechanisms that contribute to the development of these diseases (Fig. S2 in the supplemental information online).

Drug targets in the enriched pathways

We further examined the 182 diseases by asking whether the genetically identified pathways were enriched for drug targets for the same indication (Fig. 1b); for example, whether drug targets in the pathways identified using asthma GWAS genes were indeed drug targets for asthma. However, only 97 of these 182 diseases have one or more known drug targets, as defined in MetaBase [17]. We constructed a 2×2 contingency table for each of these 97 diseases, and tested for drug target enrichment in the pathway space. Remarkably, for 29 (30%) of these 97 diseases, the pathways were enriched in drug targets for the same indication at FDR $P \leq 0.01$ (Fig. 2b, Table S3 in the supplemental information online). By contrast, 0 (0%) of these 97 diseases showed enrichment without the pathway step (i.e., GWAS alone; Fig. 2a). Hence, mapping GWAS target genes into pathways revealed a concordance of GWAS targets within pathways containing already successful drug targets.

Given that the pathway approach adds value, we can pursue new genetically identified biological targets in these 182 diseases, whether (for 97 of them) or not (for 85 of them) there is a drug target against the disease. This suggests that pathway analysis can identify multiple GWAS traits with actionable pathways from both repurposing and novel target discovery perspectives. As an example,

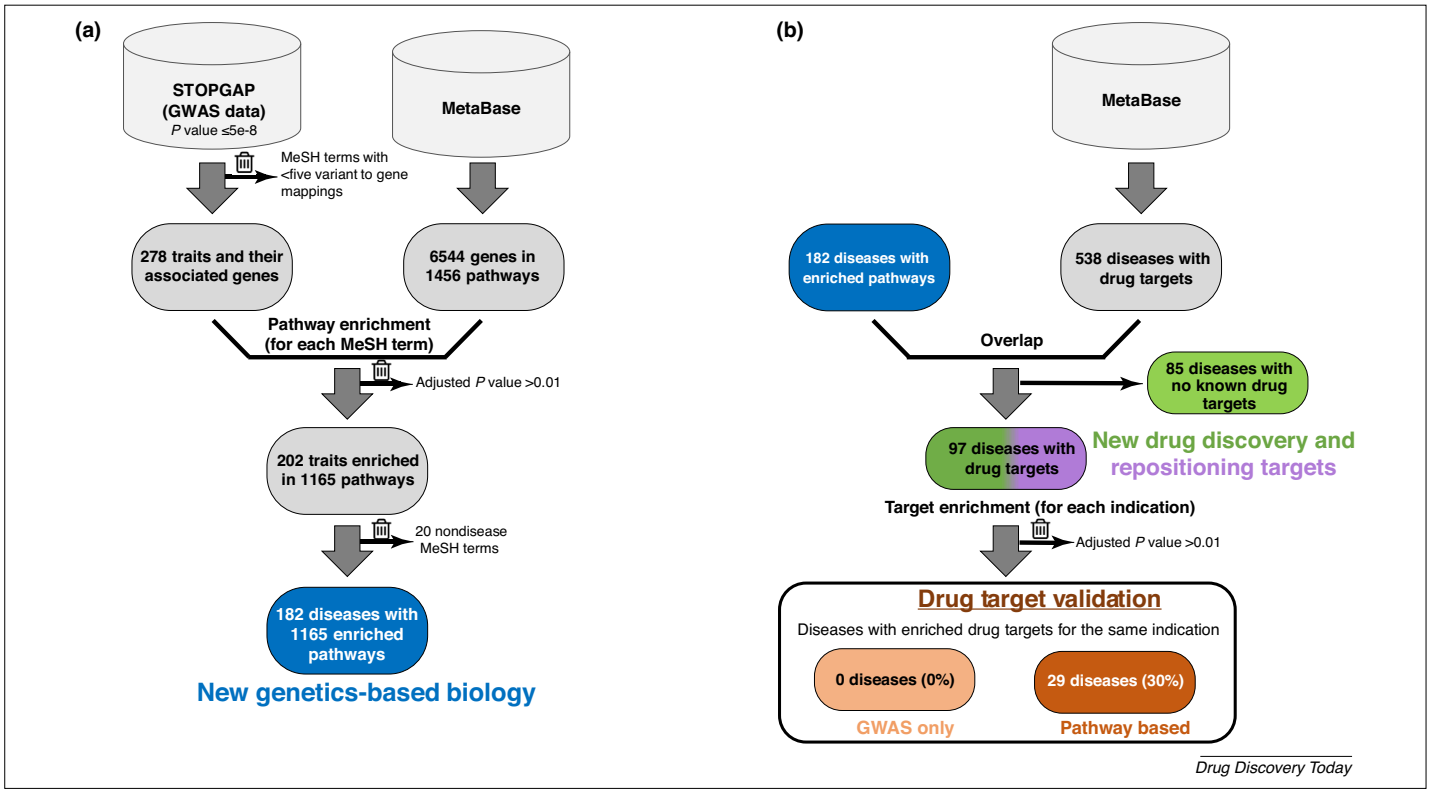

FIGURE 1

Overview of the workflow for data processing. Method used for **(a)** pathway enrichment and **(b)** drug target validation. The trash icon indicates items that were excluded from any further analysis.

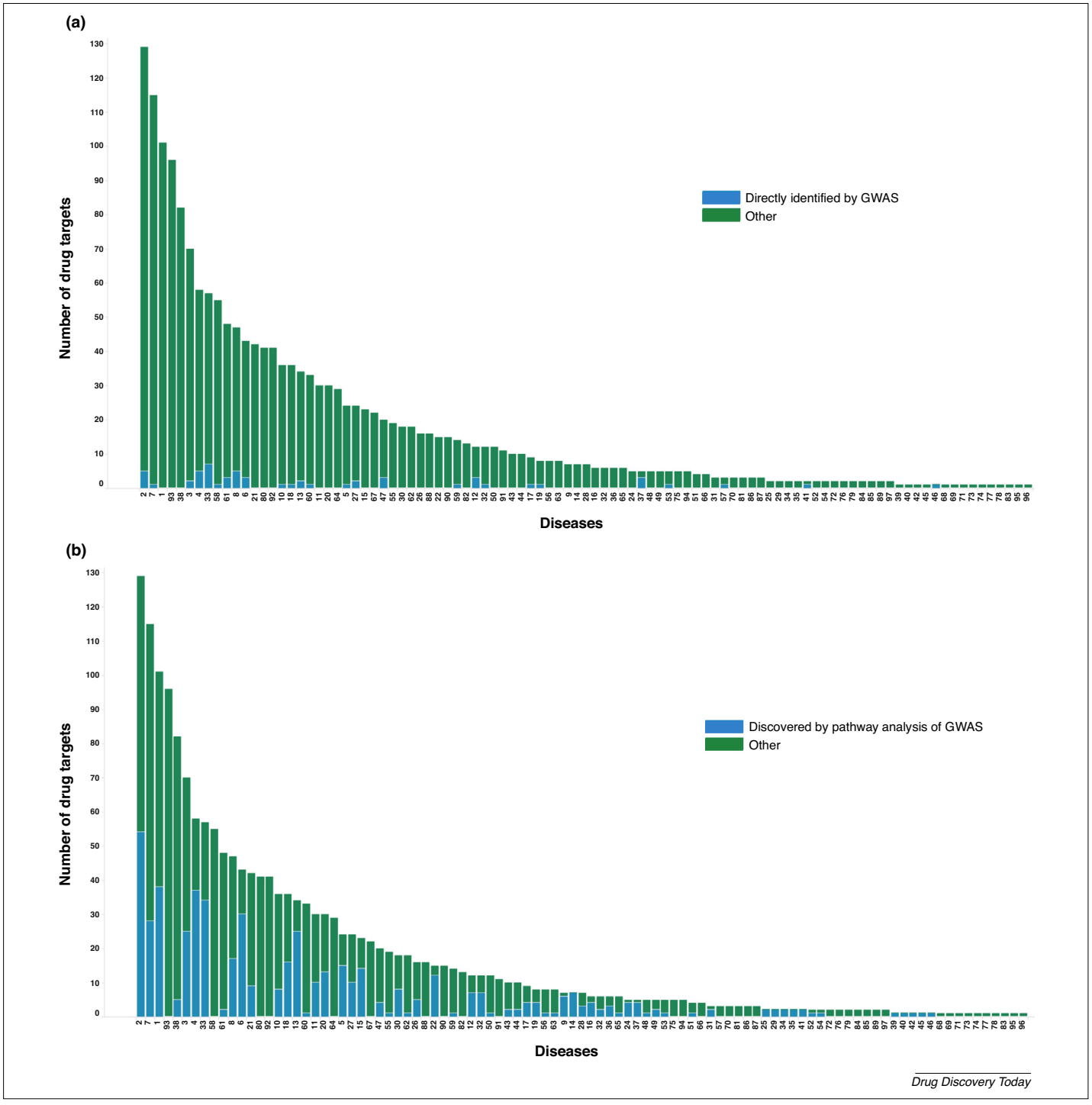

FIGURE 2

Pathway-based drug target enrichment for genome-wide association studies (GWAS). **(a)** Number of drug targets discovered by GWAS alone. **(b)** Number of drug targets discovered by pathway analysis of GWAS. The data are shown for each of the 97 GWAS-probed diseases with at least one drug target, with the value on the x-axis indicating the ranking of the enrichment significance resulting from pathway analysis, with 1 being the most significant. 1, multiple myeloma; 2, prostatic neoplasms; 3, breast neoplasms; 4, arthritis, rheumatoid; 5, lupus erythematosus, systemic; 6, psoriasis; 7, neoplasms; 8, asthma; 9, Behçet's syndrome; 10, glioblastoma; 11, melanoma; 12, dermatitis, atopic; 13, Crohn's disease; 14, inflammation; 15, myocardial infarction; 16, venous thrombosis; 17, postoperative nausea and vomiting; 18, colorectal neoplasms; 19, arthritis, psoriatic; 20, carcinoma, renal cell; 21, pulmonary disease, chronic obstructive; 22, colitis, ulcerative; 23, leukemia, lymphocytic, chronic, b cell; 24, diabetes mellitus, type 1; 25, metabolic diseases; 26, spondylitis, ankylosing; 27, schizophrenia; 28, coronary artery disease; 29, alopecia areata; 30, arthritis, juvenile; 31, lymphoma, follicular; 32, glioma; 33, multiple sclerosis; 34, tobacco-use disorder; 35, nasopharyngeal neoplasms; 36, scleroderma, systemic; 37, hypercholesterolemia; 38, carcinoma, hepatocellular; 39, leprosy; 40, choroidal neovascularization; 41, liver cirrhosis, biliary; 42, mucocutaneous lymph node syndrome; 43, ocular hypertension; 44, glaucoma; 45, hematologic diseases; 46, gout; 47, diabetes mellitus, type 2; 48, myasthenia gravis; 49, cardiovascular diseases; 50, lung neoplasms; 51, urinary bladder neoplasms; 52, anemia; 53, alopecia; 54, Hodgkin's disease; 55, atrial fibrillation; 56, acne vulgaris; 57, macular degeneration; 58, stomach neoplasms; 59, bipolar disorder; 60, Alzheimer's disease; 61, hypertension; 62, obesity; 63,

we describe the simplest significant finding on Behçet's disease for expository reasons. For other diseases with enriched drug targets, pathway-related functional assays can help further validate and prioritize the targets.

Behçet's is a chronic autoimmune and autoinflammatory disease that can affect almost all organ systems and is associated with increased mortality and high morbidity [18]. For example, a survey of the top-ranked GWAS genes from the 19 variants associated with Behçet's syndrome identified two drug targets, IL12A and CCR1, for diseases other than Behçet's, but none of the seven current Behçet's drug targets. However, the pathway approach identified six of these seven Behçet's drug targets in the 12 pathways containing an overrepresentation of genes from the 19 loci (Fig. S3 in the supplemental information online).

The non-Behçet's drug targets found in these 12 pathways represent the repositioning opportunities that are most aligned with the GWAS findings. As an example, IL12B is a member of many of these 12 pathways, but is not in a GWAS locus. Thus, we hypothesized that ustekinumab, a monoclonal antibody that blocks IL12 and IL23, could benefit patients with Behçet's. As it turns out, ustekinumab is currently being clinically tested for Behçet's disease in an open-label trial [19], although this was not recorded in MetaBase. It is also noteworthy that six of the 19 Behçet's GWAS loci were represented in these 12 pathways and that these loci include IL12A, its receptor subunit IL12RB2 and the downstream signaling molecules STAT4 and IL10. Furthermore, pathways that do not include one of the seven current Behçet's targets might represent interesting novel biology.

This analysis also identified correlated diseases with shared biological mechanisms. For example, the 'Altered Ca^{2+} handling in heart failure' pathway is significantly associated with GWAS genes from both tachycardia and autistic disorder, suggesting that the drugs associated with tachycardia could be used for treating autism. A recent review examined several reports on the use of propranolol, a beta-adrenergic antagonist, and indicated that this drug can help manage emotional, behavioral, and autonomic dysregulation in autism spectrum disorder [20].

Workflow for the analysis

GWAS data processing

The GWAS data were obtained using STOPGAP v2.5.1 [14]. STOPGAP includes data from NHGRI, GWASDB, and GRASP, which we filtered to retain only significant variants with *P* value $\leq$5e–8, resulting in a set of 473 traits obtained from 1589 GWA studies. We defined traits based on the MeSH classification available from STOPGAP and the same definition of traits is used throughout the paper. Furthermore, any records containing genes that did not map to Entrez ids were removed and any traits with fewer than five variant-to-gene mappings were also eliminated. This yielded a final set of 278 traits derived from 1398 GWAS. To avoid the biases of picking a single methodology to assign an effector gene, we included all 'candidate effector genes' [14] associated with all variants in high LD having an $r^2 \geq 0.7$. Overall, we obtained 13 172 genes associated with 278 traits (Table S4 in the supplemental information online), but we subsequently corrected for the number of genes mapped to a variant during pathway enrichment to ensure that one gene from each locus was counted at most. As an example, if four genes on a pathway were obtained from the same GWAS locus for that disease, we only included one of those genes for enrichment calculations to avoid erroneous results. This approach takes an optimistic pathway view of each variant-to-gene mapping. All these 'candidate effector genes' associated with these traits were extracted and used for pathway enrichment. To determine whether the GWAS genes in pathways were overall enriched for drug targets (as described earlier), we carried out the pathway enrichment for 5305 'best' GWAS genes (as described in STOPGAP) associated with 261 traits (data not shown). The data for 'best' genes were processed in a similar fashion to that for all 'candidate effector genes'.

Pathway enrichment

Pathway information was obtained from MetaBase version 6.30.68780. MetaBase, by Clarivate Analytics, is the manually curated knowledge base behind MetaCore, a software suite for pathway and network analysis of high-throughput sequencing data [17]. It contains 1456 protein interaction pathways, which are a comprehensive resource of human, mouse, and rat signaling, metabolism, diseases, and stem cells. R programming language was used to query the database and retrieve enriched pathways. The significance of enrichment was assessed by *P* values of hypergeometric distribution [17]. Our null hypothesis was that there was no significant pathway enriched for a given trait. The Benjamini-Hochberg method was used to obtain the adjusted *P* values and a threshold of FDR corrected *P* value $\leq$0.01 was used to determine the significance (Table S2 in the supplemental information online).

Drug data and enrichment

Drug data

Drug information for already launched drugs or those that are in preclinical or clinical development was obtained from MetaBase version 6.30.68780. In total, 672 proteins (gene identifiers) had active projects targeting them. These drugs were for 538 indications, which were mapped to diseases within the MetaBase resource.

Enrichment analysis

All enrichments were computed using a Fisher Exact test using 2×2 contingency tables with a Perl module [21]. In all, 7152 genes were present on at least one of the 1456 MetaBase pathways and, of these, only 6544 mapped to HUGO symbols for protein-coding genes; we used this as our pathway universe to obtain enrichment of drug targets for a matching indication. This is the universe that was used for the calculation described earlier.

We computed a 2×2 contingency table to check whether all the GWAS genes in pathways were overall enriched for drug targets (Table S5 in the supplemental information online). The expected number of GWAS genes that were drug targets in these pathways was 216.88, whereas the observed number was 285, with an odds ratio of 1.31 (1.12–1.55) with $P = 6.2e{-}9$. Thus, GWAS genes within this pathway universe are enriched for drug targets.

We further checked whether the enrichment (FDR $p < 0.01$) held up for each disease, and built contingency tables for each MeSH disease. Table S6 in the supplemental information online shows an example for asthma. There are two GWAS genes (*IL13* and *PDE4D*) that are explicitly targeted by asthma drugs. However, this enrichment was not significant ($P = 0.7$). We repeated this analysis across 35 MeSH diseases with known drug targets and five or more GWAS genes. Data are shown in Table S1 in the supplemental information online.

thrombocytopenia; 64, epilepsy; 65, hypersensitivity; 66, pulmonary emphysema; 67, glaucoma, open-angle; 68, alcoholism; 69, autistic disorder; 70, carcinoma, basal cell; 71, cerebrovascular disorders; 72, colonic neoplasms; 73, coronary disease; 74, depressive disorder; 75, diabetes mellitus; 76, esophageal neoplasms; 77, glomerulonephritis, iga; 78, Graves' disease; 79, heart diseases; 80, heart failure; 81, hyperlipoproteinemia type 2; 82, irritable bowel syndrome; 83, kidney diseases; 84, kidney failure, chronic; 85, leiomyoma; 86, liver diseases; 87, menopause, premature; 88, migraine disorders; 89, osteitis deformans; 90, osteoarthritis; 91, osteoporosis; 92, ovarian neoplasms; 93, pancreatic neoplasms; 94, precursor cell lymphoblastic leukemia lymphoma; 95, pulmonary fibrosis; 96, thyroid diseases; 97, vascular diseases.

For each GWAS trait, we identified enriched pathways for 'candidate effector genes' mapped to the GWAS loci. We corrected these pathway overlaps to ensure that no more than one gene was used from any loci [16] (Table S2 in the supplemental information online). We then asked whether the genes in these pathways were enriched for drug targets for the disease matching the GWAS trait. Again, we built contingency tables, as shown in Table S7 in the supplemental information online.

For example, Fisher's exact test for asthma provided an enrichment P value = 1.8e–8 (P = 2.18e–7 with Benjamini-Hochberg correction for FDR), suggesting that the pathways enriched for genes in asthma GWAS loci are also enriched for targets of drugs for asthma. The tractable genes from the 472 might in fact make good targets for asthma. These enrichments across 97 diseases are shown in Table S3 in the supplemental information online. The number of diseases with five or more variants and known drug targets increased to 97 (from 35 in Table S1 in the supplemental information online), highlighting the additional value of using all genes in each locus as potentially disease related.

Limitations

There are a few caveats with our analysis. First, the pathway collection included canonical signaling pathways and other useful constructs in terms of disease biology, although these are likely biased in terms of emphasizing known biology, but hopefully not through the excessive inclusion of potential GWAS genes. Second, GWAS, by itself, does not provide the required direction of modulation for the target (increase or decrease), that is, whether we need an agonist or antagonist for therapeutic benefit. Thus, the identified direction for repurposing could be wrong. However, the current indication of the drug will often be helpful. For example, a current anti-inflammatory in Behçet's pathway is also likely to have the correct direction for Behçet's disease. Third, we also took an optimistic view of each genetic locus preferring genes related through pathways as most likely suspects. We also performed the analysis by taking the 'best' gene in each locus (independent of pathway) and observed a significant, although reduced, enrichment for pathways (see Workflow section).

Concluding remarks

Previously published studies showed that progressing genetically validated targets can significantly increase the chances for clinical success [2,22] and that genetics-based targets can lead to 28 ± 18% more launched drugs [23]. This global pathway analysis of GWAS provides a framework to translate common disease genetics into therapeutic hypotheses and, thus, realize that 28% gain in productivity. Overall, our systematic analysis of GWAS-enriched pathways identified known and novel pathways with drug targets for 182 diseases. This provides mechanistic pathway hypotheses for each of these diseases with multiple tractable targets in each pathway and drug-repositioning opportunities for many diseases. We validated our findings by showing that 30% of the diseases had significantly more targets in these pathways than expected by chance, which is remarkable considering that none of these diseases were enriched for targets using GWAS alone. This highlights the potential for drug discovery to focus the search for novel targets and repositioning opportunities within biological pathways enriched for GWAS targets.

Acknowledgments

Authors would like to thank Andrew D. Rouillard, Daniel F. Simola, David N. Mayhew, and Philippe Sanseau for their valuable insights and discussions on this manuscript. D.J., M.M.S., and P.A. designed the study. D.J. and P.A. performed the analysis. D.J., M.M.S., M.H., and P.A. interpreted the results and wrote the manuscript.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi: https://doi.org/10.1016/j.drudis.2019.03.024.

References

1 Sanseau, P. *et al.* (2012) Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* 30, 317–320
2 Nelson, M.R. *et al.* (2015) The support of human genetic evidence for approved drug indications. *Nat. Genet.* 47, 856–860
3 Visscher, P.M. *et al.* (2017) 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22
4 Hirschhorn, J.N. (2009) Genomewide association studies—illuminating biologic pathways. *N. Engl. J. Med.* 360, 1699–1701
5 Wang, K. *et al.* (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11, 843–854
6 Sun, J. *et al.* (2015) A comparative study of disease genes and drug targets in the human protein interactome. *BMC Bioinf.* 16 (Suppl. 5), S1
7 Cao, C. and Moult, J. (2014) GWAS and drug targets. *BMC Genomics* 15 (Suppl. 4), S5
8 Li, J. and Lu, Z. (2013) Pathway-based drug repositioning using causal inference. *BMC Bioinf.* 14 (Suppl. 16), S3
9 Pers, T.H. *et al.* (2015) Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* 6, 5890
10 Ailem, M. *et al.* (2016) Unsupervised text mining for assessing and augmenting GWAS results. *J. Biomed. Inform.* 60, 252–259
11 Raychaudhuri, S. *et al.* (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* 5, e1000534
12 Mooney, M.A. *et al.* (2014) Functional and genomic context in pathway analysis of GWAS data. *Trends Genet.* 30, 390–400
13 Brodie, A. *et al.* (2014) Large scale analysis of phenotype–pathway relationships based on GWAS results. *PLoS One* 9, e100887
14 Shen, J. *et al.* (2017) STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. *Bioinformatics* 33, 2784–2786
15 Spain, S.L. and Barrett, J.C. (2015) Strategies for fine-mapping complex traits. *Hum. Mol. Genet.* 24, R111–R119
16 Agarwal, P. *et al.* SmithKline Beecham. Biological data set comparison method. US20070168135 A1.
17 Nikolsky, Y. *et al.* (2009) Functional analysis of OMICs data and small molecule compounds in an integrated 'knowledge-based' platform. *Methods Mol. Biol.* 563, 177–196
18 Nair, J.R. and Moots, R.J. (2017) Behçet's disease. *Clin. Med.* 17, 71–77
19 ClinicalTrials.Gov (2016) *Efficacy and Safety of Ustekinumab, a Human Monoclonal Anti-IL-12/IL-23 Antibody, in Patients with Behçet Disease (STELABEC).* ClinicalTrials.Gov
20 Sagar-Ouriaghli, I. *et al.* (2018) Propranolol for treating emotional, behavioural, autonomic dysregulation in children and adolescents with autism spectrum disorders. *J. Psychopharmacol.* 32, 641–653
21 Lanczos, C. (1964) A precision approximation of the gamma function. *J. Soc. Ind. Appl. Math. Ser. B: Num. Anal.* 1, 86–96
22 Cook, D. *et al.* (2014) Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* 13, 419–431
23 Hurle, M.R. *et al.* (2016) Trial watch: impact of genetically supported target selection on R&D productivity. *Nat. Rev. Drug Discov.* 15, 596–597

Deepali Jhamb
Michal Magid-Slav
Mark R. Hurle
Pankaj Agarwal*

Computational Biology, GSK R&D, Collegeville, PA, USA

*Corresponding author.