Supplementary Information

Drug Targetor: a web interface to investigate the human druggome for over 500 phenotypes

Supplementary Text 1: Cleaning bioactivity data.

ChEMBL, PHAROS, PDSP K_i database and NCBI PubChem BioAssay were used to mine experimental drug-target bioactivities. PubChem provides an activity score which can be used to assess the strength of the drug-target relationships - it is a score derived from different activity types and normalized between 0 and 100 (higher means that the drug is more active).

PubChem data was filtered based on the PubChem activity score (> 50), using the BioAssayR database from ChemMine tools (Backman, Cao, and Girke 2011), updated April 6th 2016.

ChEMBL, PHAROS, and PDSP K_i database were used to mine binding assays - data measuring binding of a compound to a target. We selected either potency or affinity data. Potency can be broadly defined as the dose range (amount of drug) over which a response is produced, and it is dependent on affinity and efficacy. Affinity is a measure of the ability of a drug and target to form a stable complex (drug + target ⇔ complex), whereas efficacy is the ability of a molecule to trigger a response in a biological system (complex ⇔ response).

In ChEMBL, PDSP K_i database, and PHAROS, we collected measures of potency and affinity: IC50 (concentration of a substance required for half maximal *in vitro* inhibition), EC50 (concentration of a substance required for half maximal *in vitro* activation), AC50 (concentration of a substance required for half maximal activity), potency, K_i (inhibitory or affinity constant), and K_d (drug-target complex dissociation constant). These are activity types that ChEMBL uses to generate its p*ChEMBL* activity, which is equal to -Log(molar IC50, XC50, EC50, AC50, K_i, K_d or Potency) (Bento et al. 2014). Each activity type is converted to p(activity)= $-\log_{10}$(activity), such as pIC50 or pK_i.

Drug-target pairs with an activity range $p(activity)_{max}$ - $p(activity)_{min}$ > 2 were discarded, when multiple entries were reported in one database. We defined a compound as active against a target if $6 \leq$ p(activity) < 14. This threshold is arbitrary - some ligands could be active and induce a response at higher

concentrations, and covalent binding might in some cases be a desirable property (Singh et al. 2011) - but an in-depth analysis of all drug-target interactions is at the moment out of the scope of Drug Tragetor. We performed the quality control on each activity type and database separately, and merged the different data sources.

Supplementary Text 2: Drug Targetor limitations.

Drug Targetor presents four main limitations: lack of power to detect genetic associations, sparsity of drug-target and drug-gene interaction data, absence of protein-protein interactions in the workflow, and simplistic SNP-to-gene mapping approach.

The first limitation is lack of power which can have an impact on gene and drug scores in Drug Targetor networks. The scores derived from GWAS are highly dependent on the GWAS sample size, i.e., the number of individuals in the study, and how heritable and polygenic the trait or disorder is (Wray et al. 2018). The power to detect signal increases with sample size and effect size - a gene can have a low score because the sample size is insufficient or the effect size too small.

The second limitation is linked to the sparsity of drug-target interaction data. Databases such as ChEMBL or PDSP K_i DB, which record drug-target activities, are helpful resources but the available data is still a small fraction of the number of all possible drug-target pairs. Moreover, a drug might have known affinity for a target but its mechanism of action (such as antagonism or agonism) might not be easily retrievable from a database entry if it is not explicitly encoded.

The third limitation is the absence of protein-protein interactions or molecular pathways in the Drug Targetor framework. Given a target *T* with high genetic association with a phenotype, targets upstream or downstream *T* in a protein-protein interaction cascade might actually be more interesting or safer to target - this is also important in the consideration of off-target effects, i.e. avoiding drugs that target unwanted proteins and could produce undesirable side effects.

The fourth limitation is a simplistic SNP-to-gene mapping which does not take into account the complex structure of chromosomes. The next planned version of Drug Targetor (v2.0) would include downstream effects of drug-target interactions, and selection of high-confidence data for drug-target and drug-gene interactions. Other possible improvements to Drug Targetor could include using quantitative

structure-activity relationship (QSAR) to complete the drug-target activity matrix; a better way to score drugs and targets taking into account tissue-specific results, side effects, polyspecificity and polypharmacology; suggesting other targets using protein-protein works; and better SNP-to-gene mapping using information on DNA conformation.

Supplementary Text 3: Gene and drug scores.

Several studies demonstrated previously that GWAS results and new insights from genetics could be used to repurpose drugs, e.g. for schizophrenia (Lencz and Malhotra 2015; Gaspar and Breen 2017; Harrison 2015), inflammatory diseases (Folkersen et al. 2015), and rheumatoid arthritis (Okada et al. 2014). The goal of Drug Targetor is to provide an interface to explore those genetic associations for hundreds of available GWAS, by ranking drugs and genes based on both GWAS summary statistics and eQTL data.

1. Gene scores

Genes (targets) in Drug Targetor networks are ranked using a score ranging from 1 to 7, based on MAGMA v1.06 (de Leeuw et al. 2015) and S-PrediXcan (Barbeira et al. 2018) association tests. The genes with the same score are ordered by MAGMA association p-value.

- Score = 7: significant in both S-PrediXcan and MAGMA (p-value $\leq$ 0.05/20,000)
- Score = 6: significant in S-PrediXcan only (p-value $\leq$ 0.05/20,000)
- Score = 5: significant in MAGMA only (p-value $\leq$ 0.05/20,000)
- Score = 4: nominally significant in both S-PrediXcan and MAGMA (p-value $\leq$ 0.05)
- Score = 3: nominally significant in S-PrediXcan only (p-value $\leq$ 0.05)
- Score = 2: nominally significant in MAGMA only (p-value $\leq$ 0.05)
- Score = 1: p-value > 0.05 in both S-PrediXcan and MAGMA and encoding a target for which drug-target bioactivities are recorded in the Drug Targetor database.

S-PrediXcan provides tissue-specific associations based on both expression data (expression quantitative trait loci, eQTLs) and GWAS summary statistics, whereas MAGMA provides a gene-wise association test only based on the GWAS results. In Drug Targetor, we report S-PrediXcan z-scores for each tissue-gene

pair: positive and negative z-scores can be interpreted as prediction of up- or downregulation for a given tissue and phenotype. For MAGMA results, we report $-\log_{10}(p\text{-}value)$ of the gene association test. MAGMA maps SNPs to genes by their chromosomal position, allowing a 35 kb upstream and 10 kb downstream gene window to account for regulatory regions. MAGMA allows to choose between different gene models, i.e., ways to combine SNP p-values; here, gene scores were computed using the "multi=snp-wise" model: aggregate p-values based on top and mean SNP p-values.

Both S-PrediXcan and MAGMA can therefore be used to derive a gene association p-value, that should be corrected for multiple testing. In Drug Targetor, associations are labelled "significant" if p-value $\leq$ 0.05/20,000, where 20,000 is the approximate number of genes tested (between 17930 and 20188 depending on the phenotype, with a median value of 19834). However, for S-PrediXcan, users should be aware that multiple tissue-wise analyses add a supplementary layer of tests which should be accounted for. In Drug Targetor, we decided to use the same p-value threshold for both MAGMA and S-PrediXcan; the raw p-values (MAGMA) and z-scores (S-PrediXcan) are provided in the interface.

2. Drug scores

In Drug Targetor, each drug is considered as a set of genes, corresponding to targets or genes the drugs interact with. All the drug-gene or drug-target connections are listed in Supplementary Table 1. To assign "genetic" scores to drugs in Drug Targetor networks, we used MAGMA competitive pathway analysis (de Leeuw et al. 2015), as described in one of our previous papers on drug repurposing for schizophrenia (Gaspar and Breen 2017). Competitive pathway analysis tests whether a set of genes is more associated with the phenotype than other genes whilst accounting for linkage disequilibrium (de Leeuw et al. 2015).

Supplementary Text 4: eQTL data in S-PrediXcan.

For each gene, a tissue-dependent association test was conducted for each phenotype using S-PrediXcan (Barbeira et al. 2018) with eQTL data from GTEx (GTEx Consortium 2013) version 7 and Depression Genes and Networks (DGN) whole-blood cohort (Battle et al. 2014).

GTEx data includes 48 different tissues with eQTL data: Adipose Subcutaneous, Adipose Visceral Omentum, Adrenal Gland, Artery Aorta, Artery Coronary, Artery Tibial, Brain Amygdala, Brain Anterior cingulate cortex BA24, Brain Caudate, Brain Cerebellar Hemisphere, Brain Cerebellum, Brain Cortex, Brain Frontal Cortex BA9, Brain Hippocampus, Brain Hypothalamus, Brain Nucleus accumbens, Brain Putamen, Brain Spinal cord cervical c-1, Brain Substantia nigra, Breast Mammary Tissue, Cells EBV-transformed lymphocytes, Cells Transformed fibroblasts, Colon Sigmoid, Colon Transverse, Esophagus Gastroesophageal Junction, Esophagus Mucosa, Esophagus Muscularis, Heart Atrial Appendage, Heart Left Ventricle, Liver, Lung, Minor Salivary Gland, Muscle Skeletal, Nerve Tibial, Ovary, Pancreas, Pituitary, Prostate, Skin Not Sun Exposed Suprapubic, Skin Sun Exposed Lower leg, Small Intestine Terminal Ileum, Spleen, Stomach, Testis, Thyroid, Uterus, Vagina, Whole Blood.

There are therefore 2 whole blood cohorts (DGN and GTEx) used by S-PrediXcan - the GTEx cohort has a smaller sample size. Sample sizes for each tissue can be accessed on the predictdb.org website: http://predictdb.org/data/sample_size_and_number_of_genes_per_tissue_v7_and_v6p.pdf.

Supplementary Text 5: Gene and drug filtering in Drug Targetor Networks.

Several options are available to filter genes and drugs in the networks - using these filter might require to also change the number of drugs and/or genes to use in the networks. By default, genes with no drug-gene connection in the network are removed, then drugs without connections are removed as well.

1. To visualize top gene associations without connection-based filtering, the user should select ***"show genes without any connection"***.
2. To visualize genes that do have connections with drugs but no connection in the top drugs included in the network, the user should select ***"show genes without connections in top drugs"***. If a drug class was selected as well, this option shows genes that do not have connections with the top drugs used to construct the network but do have connections with other drugs in the drug class.
3. To visualize top drugs without removing drugs which have no connection with top genes, the option ***"show drugs without connections in top genes"*** should be selected.
4. Finally, filters can also be applied to select genes significant or nominally significant in MAGMA and/or S-PrediXcan.

If both options ***"show genes without any connections"*** and ***"show drugs without connections in top genes"*** are selected, top genes and drugs will be shown without connection-based filtering. This might be useful to visualize the whole list of top drugs and top genes. However, for highly polygenic and well-powered studies, hundreds of significant gene associations could be identified.

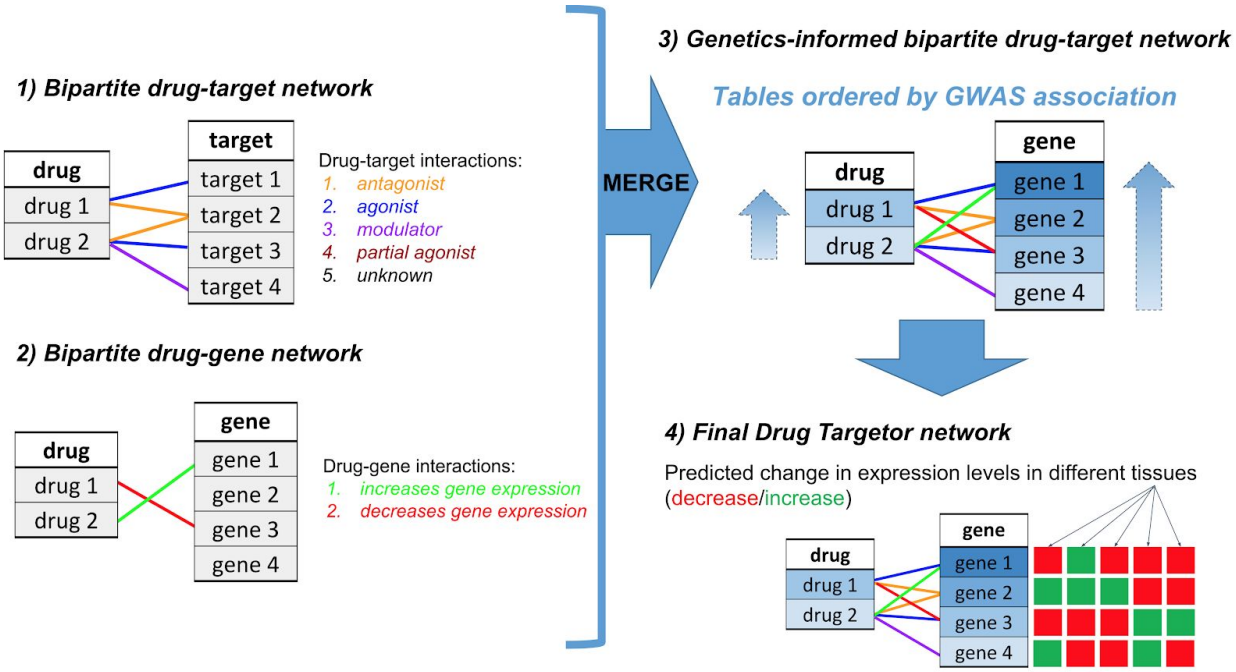
Supplementary Tables

Supplementary Table 1. Data sources used in Drug Targetor.

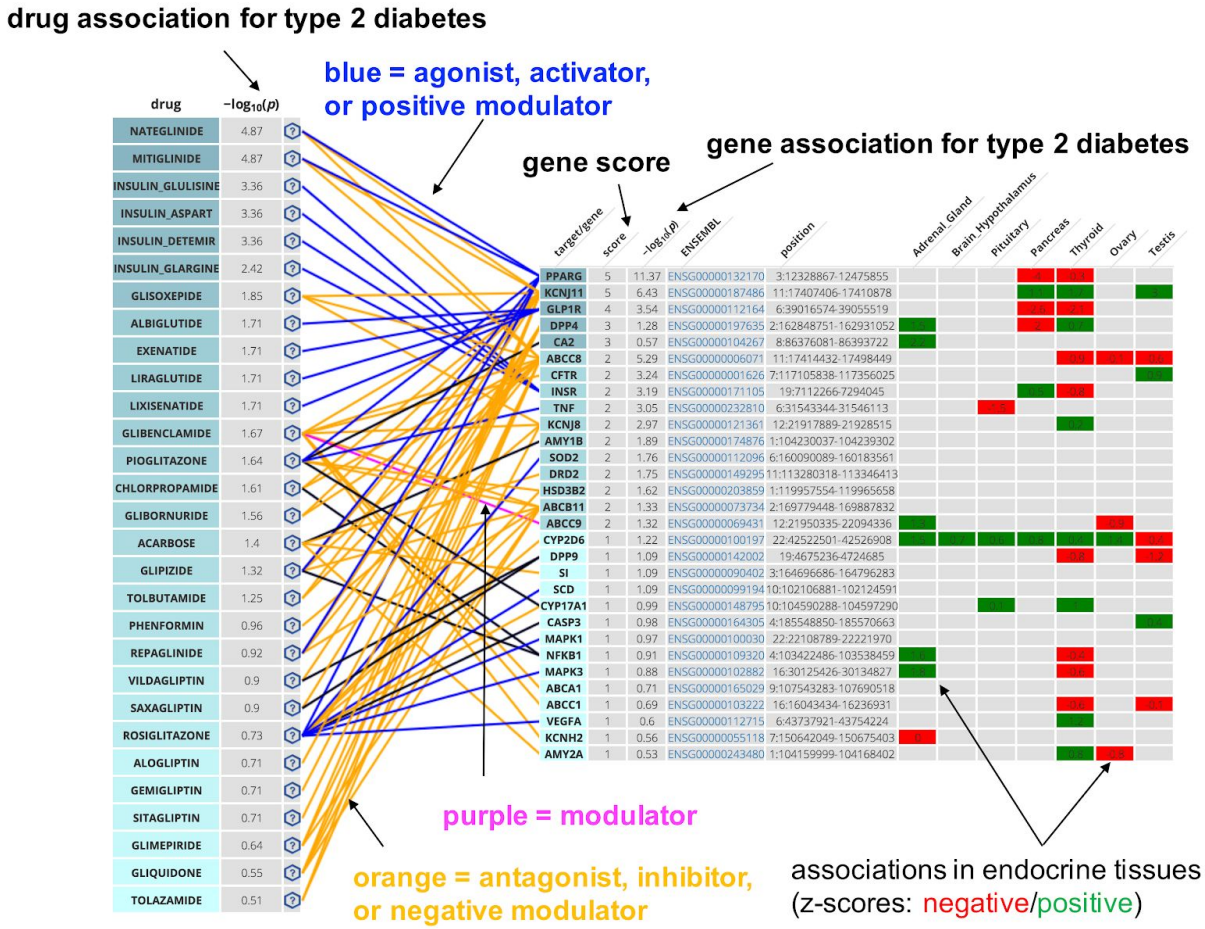
Database	**Interaction type**	**Link**
ChEMBL v23 (Gaulton et al. 2017)	Drug-target	https://www.ebi.ac.uk/chembldb/
PHAROS (Nguyen et al. 2017)	Drug-target	https://pharos.nih.gov/idg/index
PDSP K_i database (Roth et al. 2000)	Drug-target	https://pdsp.unc.edu/databases/kidb.php
NCBI PubChem BioAssay (Wang et al. 2017)	Drug-target	https://pubchem.ncbi.nlm.nih.gov/
DGIdb v2 (Griffith et al. 2013; Wagner et al. 2016)	Drug-gene and drug-target	http://www.dgidb.org/
DSigDB v1.0 (Yoo et al. 2015)	Drug-gene and drug-target	http://tanlab.ucdenver.edu/DSigDB/DSigDBv1.0/

ChEMBL, PHAROS, PDSP K_i database and PubChem can be mined for drug-target binding data (for cleaning procedure, cf. Supplementary Text 1), whereas DGIdb and DSigDB provide more diverse sources of information already encoded as drug-gene pairs - DSigDB is the largest database, including notably CMAP (Lamb et al. 2006) drug perturbagen signatures, which describe how drugs influence gene expression, and interactions obtained using text mining. Subsets of drug-target and drug-gene interactions ensembles can be selected in Drug Targetor.

Supplementary Figures



Supplementary Figure 1. Drug Targetor Networks. Regular bipartite drug-target networks are merged with drug-gene interaction networks to obtain genetics-informed networks, where drugs and genes are ordered by GWAS association; the final Drug Targetor network also shows the predicted change in expression levels in different tissues. GWAS = genome-wide association study.



Supplementary Figure 2. Drug Targetor network for diabetes drugs using type 2 diabetes genetic evidence. Diabetes drugs and their targets with genetic scores ($-\log_{10}(p)$) derived from type 2 diabetes genome-wide association study (Scott et al. 2017). Blue connections = agonist, orange = antagonist, black = unknown mechanism of action. Drug and gene associations were computed using MAGMA. Gene scores, on the other hand are based on both MAGMA and S-PrediXcan results (cf. Supplementary Text 3). Drugs and targets are ordered using drug and gene scores. Tissue-wise expression levels (z-scores) highlighted in green and red in the target table were computed using S-PrediXcan.

Supplementary note 1 - Acknowledgments.

We thank our colleagues for allowing us to use their GWAS results conducted using UK Biobank data for different phenotypes: Jonathan Coleman for neuroticism (phenotype codes in the interface: NEUR02B, NEUR02F, NEUR02M, NEUR03B, NEUR03F, NEUR03M), Kirstin Purves for anxiety (ANXI03, ANXI03F, ANXI03M), and Ken Hanscombe for physical activity (PHYS01, PHYS01F and PHYS01M).

Supplementary note 2 - Author contributions.

HG designed and implemented Drug Targetor, collected and curated the GWAS and cheminformatics data, performed the analyses, and wrote the first draft. CH contributed to the GWAS data collection and shared his own body composition GWASs (from UK biobank data) which are now available in the interface. CH and GB contributed to the writing of the paper.

References

Backman, Tyler W. H., Yiqun Cao, and Thomas Girke. 2011. "ChemMine Tools: An Online Service for Analyzing and Clustering Small Molecules." *Nucleic Acids Research* 39 (Web Server issue): W486–91.

Barbeira, Alvaro N., Scott P. Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E. Wheeler, Jason M. Torres, Eric S. Torstenson, et al. 2018. "Exploring the Phenotypic Consequences of Tissue Specific Gene Expression Variation Inferred from GWAS Summary Statistics." *Nature Communications* 9 (1): 1825.

Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, et al. 2014. "Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals." *Genome Research* 24 (1): 14–24.

Bento, A. Patrícia, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, Felix A. Krüger, et al. 2014. "The ChEMBL Bioactivity Database: An Update." *Nucleic Acids Research* 42 (Database issue): D1083–90.

Folkersen, Lasse, Shameek Biswas, Klaus Stensgaard Frederiksen, Pernille Keller, Brian Fox, and Jan Fleckner. 2015. "Applying Genetics in Inflammatory Disease Drug Discovery." *Drug Discovery Today* 20 (10): 1176–81.

Gaspar, H. A., and G. Breen. 2017. "Drug Enrichment and Discovery from Schizophrenia Genome-Wide Association Results: An Analysis and Visualisation Approach." *Scientific Reports* 7 (1): 12460.

Gaspar, H. A., Z. Gerring, C. Hübel, and C. M. Middeldorp. 2018. "Using Genetic Drug-Target Networks to Develop New Drug Hypotheses for Major Depressive Disorder." *bioRxiv*. https://www.biorxiv.org/content/early/2018/04/18/304113.abstract.

Gaulton, Anna, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, et al. 2017. "The ChEMBL Database in 2017." *Nucleic Acids Research* 45 (D1): D945–54.

Griffith, Malachi, Obi L. Griffith, Adam C. Coffman, James V. Weible, Josh F. McMichael, Nicholas C. Spies, James Koval, et al. 2013. "DGIdb: Mining the Druggable Genome." *Nature Methods* 10 (12): 1209–10.

GTEx Consortium. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nature Genetics* 45 (6): 580–85.

Harrison, Paul J. 2015. "Recent Genetic Findings in Schizophrenia and Their Therapeutic Relevance." *Journal of Psychopharmacology* 29 (2): 85–96.

Lamb, Justin, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, et al. 2006. "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease." *Science* 313 (5795): 1929–35.

Leeuw, Christiaan A. de, Joris M. Mooij, Tom Heskes, and Danielle Posthuma. 2015. "MAGMA: Generalized Gene-Set Analysis of GWAS Data." *PLoS Computational Biology* 11 (4): e1004219.

Lencz, T., and A. K. Malhotra. 2015. "Targeting the Schizophrenia Genome: A Fast Track Strategy from GWAS to Clinic." *Molecular Psychiatry* 20 (7): 820–26.

Morris, Andrew P., Benjamin F. Voight, Tanya M. Teslovich, Teresa Ferreira, Ayellet V. Segrè, Valgerdur Steinthorsdottir, Rona J. Strawbridge, et al. 2012. "Large-Scale Association Analysis Provides Insights into the Genetic Architecture and Pathophysiology of Type 2 Diabetes." *Nature Genetics* 44 (9): 981–90.

Nguyen, Dac-Trung, Stephen Mathias, Cristian Bologa, Soren Brunak, Nicolas Fernandez, Anna Gaulton, Anne Hersey, et al. 2017. "Pharos: Collating Protein Information to Shed Light on the Druggable

Genome." *Nucleic Acids Research* 45 (D1): D995–1002.

Okada, Yukinori, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, et al. 2014. "Genetics of Rheumatoid Arthritis Contributes to Biology and Drug Discovery." *Nature* 506 (7488): 376–81.

Roth, Bryan L., Estelle Lopez, Shamil Patel, and Wesley K. Kroeze. 2000. "The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches?" *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry* 6 (4): 252–62.

Singh, Juswinder, Russell C. Petter, Thomas A. Baillie, and Adrian Whitty. 2011. "The Resurgence of Covalent Drugs." *Nature Reviews. Drug Discovery* 10 (4): 307–17.

So, Hon-Cheong. 2017. "Translating GWAS Findings Into Therapies For Depression And Anxiety Disorders: Drug Repositioning Using Gene-Set Analyses Reveals Enrichment Of Psychiatric Drug Classes." *bioRxiv*. https://doi.org/10.1101/132563.

Wagner, Alex H., Adam C. Coffman, Benjamin J. Ainscough, Nicholas C. Spies, Zachary L. Skidmore, Katie M. Campbell, Kilannin Krysiak, et al. 2016. "DGIdb 2.0: Mining Clinically Relevant Drug-Gene Interactions." *Nucleic Acids Research* 44 (D1): D1036–44.

Wang, Yanli, Stephen H. Bryant, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A. Shoemaker, Paul A. Thiessen, Siqian He, and Jian Zhang. 2017. "PubChem BioAssay: 2017 Update." *Nucleic Acids Research* 45 (D1): D955–63.

Wray, Naomi R., Cisca Wijmenga, Patrick F. Sullivan, Jian Yang, and Peter M. Visscher. 2018. "Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model." *Cell* 173 (7): 1573–80.

Yoo, Minjae, Jimin Shin, Jihye Kim, Karen A. Ryall, Kyubum Lee, Sunwon Lee, Minji Jeon, Jaewoo Kang, and Aik Choon Tan. 2015. "DSigDB: Drug Signatures Database for Gene Set Analysis." *Bioinformatics* 31 (18): 3069–71.