Ferrero and Agarwal *BioData Mining* (2018) 11:7
https://doi.org/10.1186/s13040-018-0171-y

BioData Mining

SHORT REPORT Open Access

Connecting genetics and gene expression data for target prioritisation and drug repositioning

Enrico Ferrero[1,3] and Pankaj Agarwal[2*]

* Correspondence:
pankaj.agarwal@gsk.com
[2]Computational Biology, Target Sciences, GSK, 1250 S. Collegeville Road, UP12-100, Collegeville, PA 19426-0989, USA
Full list of author information is available at the end of the article

Abstract

Developing new drugs continues to be a highly inefficient and costly business. By repurposing an existing compound for a different indication, drug repositioning offers an attractive alternative to traditional drug discovery. Most of these approaches work by matching transcriptional disease signatures to anti-correlated gene expression profiles of drug perturbations. Genome-wide association studies (GWASs) are of great interest to researchers in the pharmaceutical industry because drug programmes with supporting genetic evidence are more likely to successfully progress through the drug discovery pipeline.

Here, we present a systematic approach to generate drug repositioning hypothesis based on disease genetics by mining public repositories of GWAS data and drug transcriptomic profiles. We find that genes genetically associated with a certain disease are more likely to be differentially expressed in the same disease (*p*-value = 1.54e-17 and AUC = 0.75) and that, in existing drug – disease combinations, genes significantly up- or down-regulated after drug treatment are enriched for genes genetically associated with that disease (*p*-value = 1.1e-79 and AUC = 0.64). Finally, we use this framework to generate and rank novel GWAS-driven drug repositioning predictions.

Keywords: Drug discovery, Drug repositioning, Genomics, Transcriptomics

Introduction

The discovery, development and commercialisation of a new drug is a long, expensive and often failure-prone process [1–3]. Drug repositioning can be a time- and cost-effective alternative where existing compounds are repurposed for diseases different from the original indication [4, 5]. These approaches can be subdivided into multiple classes, though a majority of recent computational work has focussed on two: drug-based, relying on chemical structure similarity and predictions of drug – target interactions, and disease-based, where transcriptomic readouts of disease samples and drug perturbations are combined [6].

The latter was popularised by the Connectivity Map [7, 8], an in silico pipeline to reverse-match transcriptional disease signatures with gene expression profiles obtained by perturbing cellular systems with a large panel of compounds. The Library of Integrated Network-based Cellular Signatures (LINCS) [9] project greatly expanded the pool of compound profiles, triggering further development of computational methods for drug repositioning [10, 11] as well as approaches for the validation of these in silico predictions [12].

Selecting the right targets is a key decision early in the drug discovery pipeline [13]: a large proportion of the efficacy failures in clinical programmes are due to lack of a clear link between the therapeutic target and the disease under investigation [14]. There is growing recognition that supporting genetic evidence from genome-wide association studies (GWASs) or phenome-wide association studies (PheWASs) linking target and disease can significantly increase the chances of success of drug discovery programmes [15, 16]. The large number of GWASs conducted over the past decade have delivered insights into the causal links of several diseases [17] and more and more genes are expected to be implicated in disease as the size of these studies grow, even though not all associations might be as meaningful as previously thought [18].

GWASs [19], PheWASs [20], Connectivity Map approaches [21–23] and Open Targets [24] have all been used to repurpose drugs. Here, we combine disease data from GWASs with drug perturbation transcriptional profiles and a Connectivity Map-inspired method to generate repositioning hypotheses that, unlike those in standard expression-based repurposing workflows, are supported by genetics evidence.

Methods

Software and code

R 3.4.0 [25] was used for all data processing and analysis. All code was versioned using Git and is hosted at https://github.com/enricoferrero/GCMap.

Data sources

STOPGAP [26] is a database containing associations between DNA mutations occurring in diseases and likely target genes. This includes rare disease associations as well as data from GWASs. For single nucleotide polymorphisms (SNPs) in regulatory regions, associations to target gene are performed on the basis of supporting evidence including eQTL and regulatory genomics data. The complete dataset (294,505 associations between 20,015 genes and 1746 medical terms) was downloaded from https://github.com/StatGenPRD/STOPGAP/blob/master/STOPGAP_data/stopgap.gene.mesh.RData. Open Targets [27] maps diseases to relevant genes using a number of evidence types, including genes differentially expressed in the disease, germline and somatic mutations, curated pathway, animal model and literature data as well as known drugs approved for the treatment of the disease. The Open Targets API at http://api.opentargets.io/v3/platform/docs was accessed on 20th June 2017 and used to download lists of genes differentially expressed in disease (216,942 associations between 22,190 genes and 148 diseases) and links between 595 diseases and 1555 approved drugs (7351 associations). The LINCS [9] L1000 data consists of gene expression profiles obtained by perturbing different cell lines with a large collection of compounds. To obtain the complete genome-wide dataset in a convenient format, we used the Harmonizome [28], a large collection of uniformly processed biological datasets. The file downloaded from http://amp.pharm.mssm.edu/static/hdfs/harmonizome/data/lincscmapchemical/gene_attribute_edges.txt.gz contained 4,189,677 associations between 3924 compounds and 8347 genes differentially expressed after treatment, with a median of 257 genes changing for each compound.

Data processing

STOPGAP data: gene – disease associations from rare disease sources (OMIM and Orphanet) were excluded. To ensure compatibility with other resources, gene symbols were mapped to Ensembl gene IDs with the EnsDb.Hsapiens.v75 package [29]. MeSH terms were mapped to terms in the Experimental Factor Ontology (EFO) [30] using Zooma [31]. LINCS L1000 data from Harmonizome: Entrez gene IDs were mapped to Ensembl gene IDs. Compound IDs were mapped to ChEMBL IDs with UniChem [32], using PubChem IDs as an intermediate.

Data analysis

EFO IDs and ChEMBL IDs were used to match diseases and drugs across different resources, respectively. A Fisher's exact test [33] was used to perform enrichment tests between gene sets and to generate repositioning hypotheses. Results were corrected for multiple hypothesis testing using the Benjamini – Hochberg correction [34] and only results below a 5% (or lower) false discovery rate (FDR) threshold were considered significant. The Mann – Whitney U test [35] was used to assess whether distributions were significantly different. Receiver operating characteristic (ROC) curves and confidence intervals are calculated using a bootstrap procedure with 1000 iterations performed using the pROC package [36]. The riverplot package [37] was used to draw the Sankey diagram, while all other plots were generated with ggplot2 [38].

Results

We set out to assess whether gene associations from GWASs could be leveraged to formulate drug repositioning hypotheses. First, we asked whether genes that are genetically associated with a disease are more likely to also be differentially expressed in that same disease, when compared to any other disease. As conventional drug repositioning approaches typically rely on transcriptomic readouts as disease representations, if there is a significant overlap between genetic associations and differentially expressed genes (DEGs) in any given disease, then we argue that GWAS data could replace or supplement transcriptional signatures in such workflows. We refer to this as Hypothesis 1: *genes differentially expressed in disease X are enriched for genes genetically associated with disease X, compared to other diseases*. For each disease, we obtained GWAS hits from STOPGAP [26] and lists of genes differentially expressed in both directions from Open Targets [27]. We then calculated the odds ratio and the significance of the overlap between gene sets for each pairwise disease combination using Fisher's exact test. We compared the *p*-values distributions of gene sets from the same disease and from different diseases and observed a statistically significant difference (*p*-value = 1.54e-17), with gene sets from the same disease more likely to show a significant overlap between genetic and transcriptomic hits (Fig. 1a). To quantify the predictive power of our observation, we carried out a receiver operating characteristic (ROC) analysis by considering the negative base 10 logarithm of the adjusted *p*-values as the ranking metric and whether the two gene sets came from the same disease or not as labels (Fig. 1b). We observed a total area under the curve (AUC) of 0.75 (95% confidence interval [0.70, 0.80]) suggesting it is possible to predict whether a genetic and a transcriptomic gene set originate from the same disease based on the significance of their overlap. The list

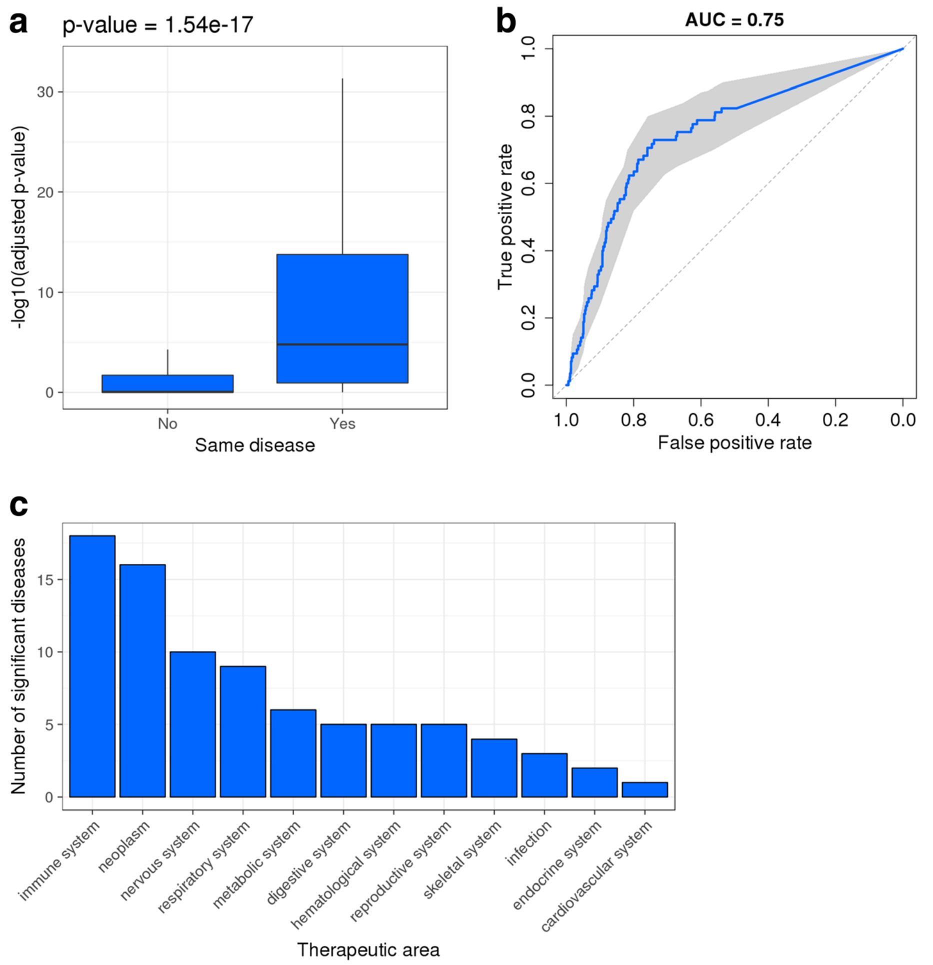

Fig. 1 Genes differentially expressed in disease are enriched for genes genetically associated with the same disease. **a** Boxplots showing distributions of Fisher's exact test *p*-values for genetic and transcriptomic gene sets overlaps from the same or different diseases. **b** ROC curve with 95% confidence intervals obtained using same or different disease as labels and significance of enrichment between genetic and transcriptomic gene sets as the ranking metric. **c** Barplot showing breakdown of diseases by therapeutic area for which there is a significant overlap between GWAS associations and genes differentially expressed

of diseases with significant overlap between DEGs and GWAS associations (adjusted *p*-value < 0.05) includes several immune diseases as well as a proportion of oncology, neurological and respiratory indications (Fig. 1c and Additional file 1: Table S1). These results highlight the confluence of genetic association and gene expression changes in disease and suggest that GWAS genes with dysregulated expression might be better initial candidates for drug target discovery.

We then asked whether, for indications with commercially available drugs, genes transcriptionally modulated by the drug are enriched for genes genetically associated with the disease. For indications showing enrichment, we propose that the drug causing significant expression changes in this set of GWAS genes could constitute a potential repositioning option. We refer to this as Hypothesis 2: *genes differentially expressed after treatment with drug Z for disease X are enriched for genes genetically associated with disease X*. We retrieved all 595 indications of 1555 current drugs from Open Targets [27] and calculated the significance of the overlap between genes differentially expressed after drug treatment and genes genetically associated with disease for all drug – disease combinations using Fisher's exact test. We then split the dataset into two groups according to whether the drug was already approved for that indication and assessed whether the corresponding distributions of adjusted *p*-values were different. We found that the *p*-values of the 7351 approved drug – indication pairs were considerably lower than the rest (*p*-value = 1.10e-79, Fig. 2a). This shows that, in several cases, approves drugs do indeed regulate the expression of genes genetically associated with the disease, and suggests that drugs can be repositioned based on the overlap between the genes they modulate and the genetic hits in target diseases. We generated ROC curves by using the significance of the overlap between the two gene sets as predictions and whether the drug – disease association was an approved one or not as labels and observed an AUC of 0.64 (95% confidence interval [0.63, 0.65], Fig. 2b), highlighting that this approach can classify existing and non-existing drug – indication pairs based on the overlap between genetic hits from the disease and genes modulated by the drug.

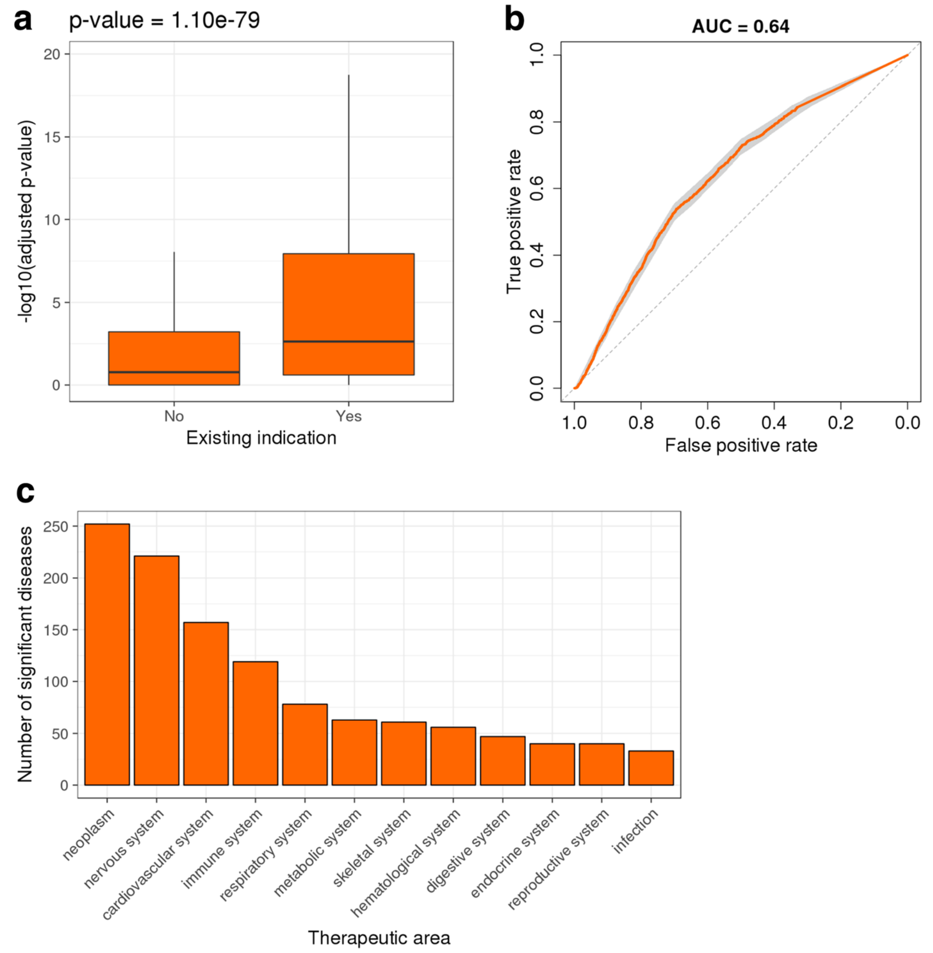

Fig. 2 Genes differentially expressed after treatment with drug approved for a disease are enriched for genes genetically associated with the same disease. **a** Boxplots showing distributions of Fisher's exact test *p*-values between GWAS disease associations and DEGs after drug treatment for current and other indications. **b** ROC curve with 95% confidence interval obtained using current indications as positive labels and significance of enrichment between genetic hits in disease and drug transcriptomic profiles as the ranking metric. **c** Barplot showing breakdown of current drug indications by therapeutic area where there is a significant overlap between GWAS associations and DEGs after drug treatment

We recover 911 significant existing drug – disease associations (adjusted *p*-value < 0.05), particularly for diseases in oncology, immunology and neurological and cardiovascular therapeutic areas (Fig. 2c and Additional file 2: Table S2).

Overall, these results show that successful drug – disease combinations tend to display a significant overlap between the genetic background of the disease and the transcriptional response to the drug used to treat the disease. Hence, we propose to utilise the most significant, though not yet approved and perhaps not even tested, drug – indication pairs resulting from this analysis as drug repositioning hypotheses. To limit false positives, we filtered our results using a stringent adjusted *p*-value threshold (1e-10) and identified nearly 9000 such opportunities which could be prioritized and tested (Additional file 3: Table S3). Visualisation of the entire repurposing space (Fig. 3) reveals oncology and neurology as the two therapeutic areas with the largest pool of

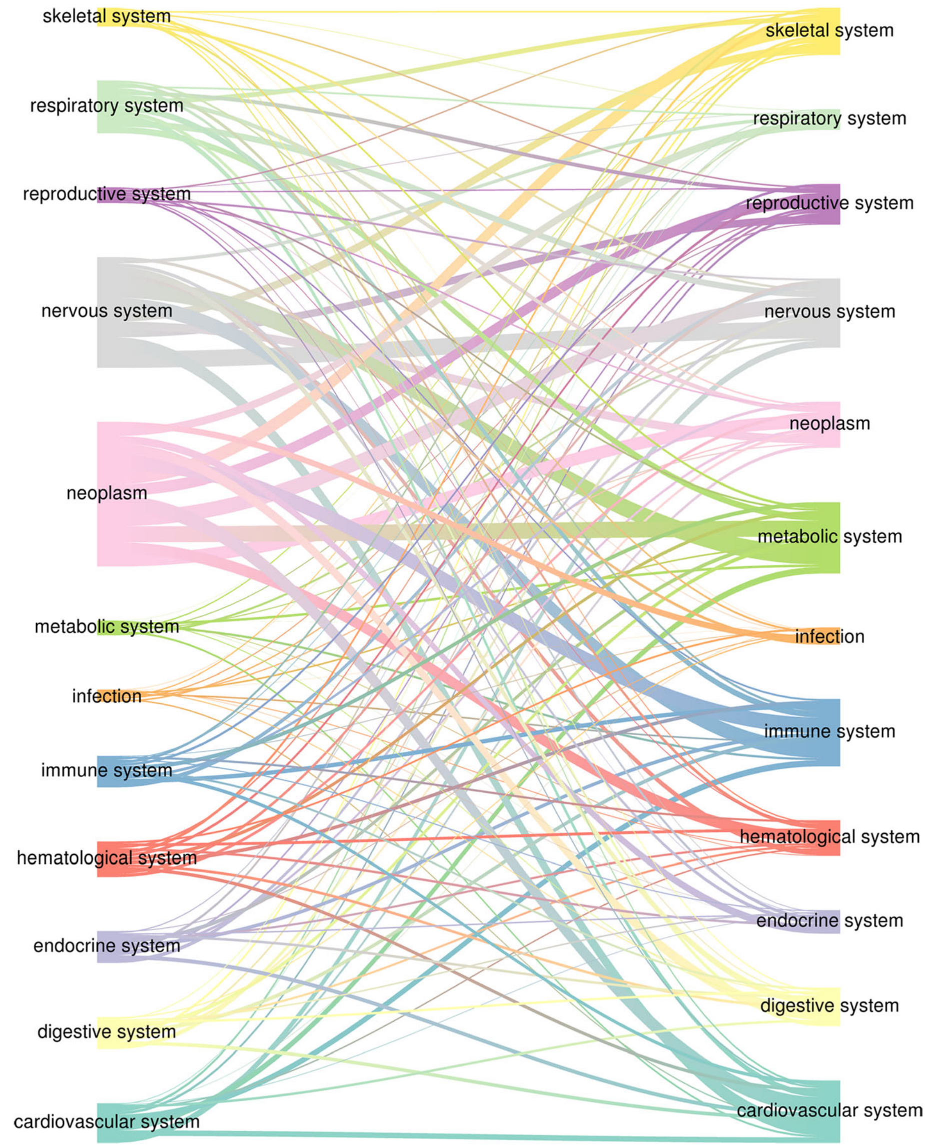

Fig. 3 GWAS-driven drug repositioning hypotheses by therapeutic area. Sankey diagram showing all significant drug repurposing trajectories across different therapeutic areas

approved drugs that could be repositioned elsewhere (1093 and 926 compounds, respectively), followed by respiratory (654). However, many oncology drugs are not suitable for other indications because of their toxicity profiles. Nervous system indications could also be among the largest recipient of repurposed drugs (709), together with metabolic (719) and immune diseases (703). More specifically, the most promising trajectories appear to be nervous system to metabolic system (113 drugs), nervous system to nervous system (108), neoplasm to metabolic system (106), neoplasm to immune system (105) and nervous system to immune system (104).

Conclusions

We showed that genes genetically associated with a disease often significantly overlap with genes differentially expressed in the same disease, as well as with genes induced or repressed by drugs used to treat that disease. To our knowledge, this is the first report to test and validate these hypotheses. We presented a simple approach to generate target prioritisation and drug repositioning hypotheses that are driven by the genetic background of the disease. Unlike more conventional repurposing approaches that rely on reverse matching of drug and disease transcriptomic signatures, we have taken advantage of the notion that genetic evidence is crucial to maximise the chances of success of drug discovery programmes [15, 16].

Our in silico framework returns a large number of statistically significant results and validation of these hypotheses would require extensive experimental work. We believe this is a major limitation of our work: we are acutely aware of the many challenges and low success rates of drug discovery programmes and recognise that a considerable proportion of our hits could be false positives. Diseases with several associated genes and drugs eliciting large transcriptional responses are more likely to result in significant results simply because of the size of these gene sets and the methodology used to compute significance.

Another issue is the lack of directionality in the genetics data we use to represent the disease space. While other Connectivity Map-inspired methods exploit up- or down-regulated genes in the transcriptomic data to identify compound profiles reversing a disease signature [7, 10, 11], our method does not take directionality into account. This could result in false positives, including predictions which could actually worsen the disease state.

In conclusion, our work represents a proof of concept that combining disease genetic and drug transcriptomic data is a valuable approach for GWAS-based drug repositioning. However, we recognise that much work remains to be done to improve its real-world applicability and would like to encourage further research in this area.

Additional files

Additional file 1: Table S1. List of diseases with a significant overlap between genes differentially expressed and genes genetically associated with disease. Results are filtered for adjusted *p*-value < 0.05. (CSV 9 kb)

Additional file 2: Table S2. List of current drug indications with a significant overlap between genes significantly up- or down-regulated after drug treatment and GWAS hits. Results are filtered for adjusted *p*-value < 0.05. (CSV 119 kb)

Additional file 3: Table S3. Drug repositioning hypotheses based on the significance of the overlap between genes genetically associated with the disease and genes differentially expressed after drug treatment. Results are filtered for adjusted *p*-value <1e-10. (CSV 1089 kb)

Abbreviations

AUC: Area under the curve; DEG: Differentially expressed gene; EFO: Experimental Factor Ontology; FDR: False discovery rate; GWAS: Genome-wide association study; LINCS: Library of Integrated Network-based Cellular Signatures; PheWAS: Phenome-wide association study; ROC: Receiver operating characteristic

Acknowledgements

We would like to thank Gautier Koscielny and Andrew Rouillard for enabling and facilitating access to data.

Funding

The authors received no specific funding for this work.

Availability of data and materials

The datasets analysed during the current study are available at the following repositories: STOPGAP (https://github.com/StatGenPRD/STOPGAP/blob/master/STOPGAP_data/stopgap.gene.mesh.RData); Harmonizome (http://amp.pharm.mssm.edu/static/hdfs/harmonizome/data/lincscmapchemical/gene_attribute_edges.txt.gz) and Open Targets (http://api.opentargets.io/v3/platform/docs).

The datasets generated during the current study are included in this published article and its additional files.

Authors' contributions

EF conceived the initial idea. EF and PA refined the proposed workflow. EF performed the data analysis. EF and PA worked on the interpretation of the results. EF wrote the first draft of the manuscript. EF and PA made changes and additions to the manuscript. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Competing interests

EF and PA are full-time employees of GSK (GlaxoSmithKline).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Computational Biology, Target Sciences, GSK, Gunnels Wood Road, Stevenage SG1 2NY, UK. [2]Computational Biology, Target Sciences, GSK, 1250 S. Collegeville Road, UP12-100, Collegeville, PA 19426-0989, USA. [3]Present Address: Autoimmunity, Transplantation and Inflammation, Novartis Institutes for Biomedical Research, Fabrikstrasse 2, Basel 4056, Switzerland.

Received: 26 October 2017 Accepted: 14 May 2018
Published online: 31 May 2018

References

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ. 2016;47:20–33.
2. Arrowsmith J, Miller P. Trial watch: phase II and phase III attrition rates 2011–2012. Nat Rev Drug Discov. 2013;12:569.
3. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, Pairaudeau G, Pennie WD, Pickett SD, Wang J, Wallace O, Weir A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. Nat Rev Drug Discov. 2015;14:475–86.
4. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3:673–83.
5. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther. 2013;93:335–41.
6. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. Brief Bioinform. 2011;12:303–11.
7. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J-P, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313:1929–35.
8. Qu XA, Rajpal DK. Applications of connectivity map in drug discovery and development. Drug Discov Today. 2012;17(23–24):1289–98.
9. Vidović D, Koleti A, Schürer SC. Large-scale integration of small molecule-induced genome-wide transcriptional responses, Kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action. Front Genet. 2014;5(SEP):1–14.
10. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Brief Bioinform. 2016;17:2–12.
11. Musa A, Ghoraie LS, Zhang S-D, Galzko G, Yli-Harja O, Dehmer M, Haibe-Kains B, Emmert-Streib F. A review of connectivity map and computational approaches in pharmacogenomics. Brief Bioinform. 2017;32:bbw112.
12. Brown AS, Patel CJ: A review of validation strategies for computational drug repositioning. Brief Bioinform 2016:bbw110.

13. Plenge RM. Disciplined approach to drug discovery and early development. Sci Transl Med. 2016;8:349ps15.
14. Cook D, Brown D, Alexander R, March R, Morgan P, Satterthwaite G, Pangalos MN. Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. Nat Rev Drug Discov. 2014;13:419–31.
15. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. Nat Rev Drug Discov. 2013;12:581–94.
16. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. The support of human genetic evidence for approved drug indications. Nat Genet. 2015; 47:856–60.
17. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 years of GWAS discovery: biology, function, and Translation. Am J Hum Genet. 2017;101:5–22.
18. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to Omnigenic. Cell. 2017;169: 1177–86.
19. Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, Mooser V. Use of genome-wide association studies for drug repositioning. Nat Biotechnol. 2012;30:317–20.
20. Rastegar-Mojarad M, Ye Z, Kolesar JM, Hebbring SJ, Lin SM. Opportunities for drug repositioning from phenome-wide association studies. Nat Biotechnol. 2015;33:342–5.
21. Dudley JT, Sirota M, Shenoy M, Pai RK, Roedder S, Chiang AP, Morgan AA, Sarwal MM, Pasricha PJ, Butte AJ. Computational repositioning of the anticonvulsant Topiramate for inflammatory bowel disease. Sci Transl Med. 2011;3:96ra76.
22. Cheng J, Yang L, Kumar V, Agarwal P. Systematic evaluation of connectivity map for disease indications. Genome Med. 2014;6:95.
23. Fortney K, Griesman J, Kotlyar M, Pastrello C, Angeli M, Sound-Tsao M, Jurisica I. Prioritizing therapeutics for lung Cancer: an integrative meta-analysis of Cancer gene signatures and Chemogenomic data. PLoS Comput Biol. 2015;11:e1004068.
24. Khaladkar M, Koscielny G, Hasan S, Agarwal P, Dunham I, Rajpal D, Sanseau P. Uncovering novel repositioning opportunities using the open targets platform. Drug Discov Today. 2017;22(12):1800–1807.
25. R Core Team: R: A Language and Environment for statistical Computing 2017:https://www.r-project.org/.
26. Shen J, Song K, Slater AJ, Ferrero E, Nelson MR. STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. Bioinformatics. 2017;33:2784–6.
27. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, Hasan S, Karamanis N, Maguire M, Papa E, Pierleoni A, Pignatelli M, Platt T, Rowland F, Wankar P, Bento AP, Burdett T, Fabregat A, Forbes S, Gaulton A, Gonzalez CY, Hermjakob H, Hersey A, Jupe S, Kafkas Ş, Keays M, Leroy C, Lopez F-J, Magarinos MP, Malone J, et al. Open targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2017;45:D985–94.
28. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016;2016:baw100.
29. Rainer J. EnsDb.Hsapiens.v75. 2016. https://doi.org/10.18129/B9.bioc.EnsDb.Hsapiens.v75.
30. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an experimental factor ontology. Bioinformatics. 2010;26:1112–8.
31. EMBL-EBI: Zooma. 2017:http://www.ebi.ac.uk/spot/zooma/.
32. Chambers J, Davies M, Gaulton A, Hersey A, Velankar S, Petryszak R, Hastings J, Bellis L, McGlinchey S, Overington JP. UniChem: a unified chemical structure cross-referencing and identifier tracking system. J Cheminform. 2013;5:3.
33. Fisher RA. On the interpretation of χ^2 from contingency tables, and the calculation of P. J R Stat Soc. 1922;85:87.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 1995;57:289–300.
35. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18:50–60.
36. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
37. Weiner J. riverplot. 2017. https://CRAN.R-project.org/package=riverplot.
38. Wickham H. ggplot2. New York, NY: Springer New York; 2009.