GUILDify v2.0: A tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets

Joaquim Aguirre-Plans[1], Janet Piñero[2], Ferran Sanz[2], Laura I. Furlong[2], Narcis Fernandez-Fuentes[3,4], Baldo Oliva[1,*] and Emre Guney[2,5,*]

1- Structural Bioinformatics Group, Research Programme on Biomedical Informatics, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia 08003, Spain
2- Integrative Biomedical Informatics Group, Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute, Department of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Catalonia 08003, Spain
3- Department of Biosciences, U Science Tech, Universitat de Vic-Universitat Central de Catalunya, Vic, Catalonia 08500, Spain
4- Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, SY23 3EB Aberystwyth, United Kingdom
5- Department of Pharmacology and Personalised Medicine, CARIM, FHML, Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands

***Correspondence to Emre Guney**: emre.guney@upf.edu
***Correspondence may also be addressed to Baldo Oliva**: baldo.oliva@upf.edu

Abstract

The genetic basis of complex diseases involves alterations on multiple genes. Unravelling the interplay between these genetic factors is key to the discovery of new biomarkers and treatments. In 2014, we introduced GUILDify, a web server that searches for genes associated to diseases, finds novel disease-genes applying various network-based prioritisation algorithms and proposes candidate drugs. Here, we present GUILDify v2.0, a major update and improvement of the original method, where we have included protein interaction data for seven species and 22 human tissues and incorporated the disease-gene associations from DisGeNET. To infer potential disease relationships associated with multi-morbidities, we introduced a novel feature for estimating the genetic and functional overlap of two diseases using the top-ranking genes and the associated enrichment of biological functions and pathways (as defined by GO and Reactome). The analysis of this overlap helps to identify the mechanistic role of genes and protein-protein interactions in comorbidities. Finally, we provided an R package, guildifyR, to facilitate programmatic access to GUILDify v2.0 (http://sbi.upf.edu/guildify2)

Introduction

Complex diseases such as cancer, diabetes, neurodegenerative disorders or cardiovascular diseases are rarely caused by a single genetic perturbation and usually involve polygenic modifications on the underlying interconnected cellular network. Understanding the genetic basis of diseases and the interactions of disease-associated proteins in the protein interaction network (PIN) is essential for the development of new rational therapeutic strategies. Despite recent large-scale genotyping efforts, information on disease-gene associations is still limited, often explaining a small percentage of the phenotypic variance observed among individuals [1]. To address this limitation and infer novel disease-gene associations, various disease-gene prioritisation methods have been suggested, exploiting the “guilt-by-association” principle over certain features of disease-genes such as similarity in sequence and functional annotations, clustering in the linkage interval, or proximity in the PIN [2]. Indeed, albeit the PINs being incomplete [3], the proximity to disease-genes in the PIN has proven extremely useful in prioritising disease-associated genes [4]. Consequently, a number of tools and web servers has been developed to expand the number of disease-associated genes using the interactome [5–9].

Previously, we presented GUILDify, a web server that applies the prioritisation algorithms developed in GUILD software to find novel disease-gene associations based on the connectedness of genes in the PIN [10,11]. GUILDify searches for genes starting from user-provided keywords such as the names of diseases or gene symbols in the BIANA knowledge database. It uses the genes associated to the keywords as seeds and the PIN for the selected organism to apply graph theory algorithms to prioritise new disease genes. Recently, GUILDify has been applied to: (i) find comorbidities across genetic diseases [12]; (ii) construct PINs specific to breast cancer metastasis to lung and brain [13]; (iii) identify candidate genes for body size in sheep [14] and (iv) prioritise preeclampsia pathogenesis [15].

Here, we present a comprehensive upgrade, GUILDify v2.0, where we updated the underlying biological databases in BIANA knowledge database (protein and drug-target interactions, functional and disease annotations) and: (i) facilitated the use of seven species-specific PINs and 22 human tissue-specific PINs; (ii) increased the quality and number of disease-gene associations by incorporating DisGeNET to our datasets; (iii) incorporated the option to search by drug name, allowing the prioritisation of genes based on known drug targets to uncover the neighbourhood of the PIN affected by the drug; (iv) improved the visualisation of the results using cytoscape.js; (v) refined the definition of top-ranking genes based on whether they had similar functional annotations as the seeds, thus providing the biologically most coherent subnetwork relevant to a given disease; (vi) introduced a feature to measure the genetic and functional overlap of the top-ranking genes of two different diseases, supporting the investigation of disease comorbidities; (vii) implemented a new drug repurposing functionality to propose novel indications for a given drug based on the genetic and functional overlap; and (viii) developed an R package to facilitate the programmatic access to the methods implemented in the web server.

Results and Discussion

Advances

1. Identifying genetic and functional similarities across diseases

In recent works, we have shown that the genetic and functional similarities of diseases in the PIN can be used to characterise co- and multi-morbidities across diseases [12] and also to repurpose existing drugs targeting these diseases [16]. Motivated by these findings and to provide systematic insights on disease-disease relationships, GUILDify v2.0 now allows users to identify the overlap between two previously submitted results, i.e. sets of genes linked to two different diseases. Accordingly, given two job IDs corresponding to the prioritisation results of two different diseases, GUILDify v2.0 provides: (i) the overlap between the top-ranking genes of the two diseases; (ii) the overlap between the enriched functions among the top-ranking genes of the two diseases; (iii) the enriched functions among the common top-ranking genes; and (iv) a network visualisation of the interactions between common top-ranking genes. Moreover, GUILDify v2.0 also calculates the Fisher's exact test to quantify the significance of the overlap between genes and functions and report one-sided P-value (see details in Supplementary Material). GUILDify v2.0 is the first server that permits the use of gene prioritisation results to explore disease-disease relationships with such simplicity and flexibility.

2. Prioritisation of drug targets

GUILDify v2.0 now allows to search by a drug in addition to a phenotype and returns a list of drug-target associations integrated from DrugBank [17], DGIdb [18], DrugCentral [19] and ChEMBL [20] (see details in Supplementary Material). This new functionality allows the characterisation of the neighbourhood of the drug in the PIN, i.e. neighbouring proteins to those targeted by the drug, and thus providing insights on the potential mechanism of action of the drug. Moreover, the novel feature of assessing the overlap between two network expansion runs (i.e. two job IDs) can also be applied in multiple scenarios to: (i) identify the similarity between the neighbourhood of two drugs in the PIN, which can be useful to identify drug interactions; (ii) compare the neighbourhood of a disease with the neighbourhood of a drug in the PIN, which can be applied to drug repurposing. Such novel features make GUILDify v2.0 one of the most easy-to-use and flexible web servers to inspect the effect of drugs in the PIN.

3. Screening diseases to identify potential new indications of known drugs

Building upon new technical developments mentioned above, GUILDify v2.0 now offers a novel drug repurposing functionality. Given a job ID associated with a drug (or a list of drug targets), this feature automatically calculates the overlap of genes (or functions) between the given drug and a set of pre-calculated diseases. Details on the method and validation of drug repurposing are described in detail at Supplementary Material.

4. Tissue and species-specific PINs

The analysis of the protein interactions in a tissue-specific context is becoming increasingly relevant to understand genetic diseases and find improved treatments [21]. We have included tissue-specific networks derived from 22 different human tissues (see Supplementary Table S1). To create these networks, we filtered the interactions in the global PIN using RNAseq data from GTEx [22], keeping only the interactions between proteins encoded by genes that are expressed in a given tissue (i.e. considering only transcripts with TPM (transcripts per kilobase million) expression values of 1 or higher (see details in Supplementary Material). We have also included 7 species-specific PINs derived from experimentally determined protein-protein interactions. Although the coverage of interactomic data for some species is low (e.g., 11,943 interactions in rat vs 320,337 interactions in human), these PINs provide a reliable backbone for interactome-based analyses (e.g., in preclinical research) as opposed to PINs generated by predicted interactions based on homology information.

5. Disease-gene information from DisGeNET

We incorporated DisGeNET, one of the largest repositories of genes and variants associated to human diseases [23]. DisGeNET relies on data from UniProt [24], CTD [25], CLINVAR [26], ORPHANET [27], GWAS Catalog [28], PsyGeNET [29] and HPO [30] and is integrated in BIANA [31]. To investigate the increase in the number of disease-gene associations between versions 1 and 2 of GUILDify, we checked the number of associations for the lowest-level non-obsolete diseases from Disease Ontology [32] that were available in our repositories (2,190 terms). GUILDify v1 contains gene associations for 1,505 diseases and 4,171 genes (2.8 genes per disease), while updated GUILDify v2.0 has gene associations for 2,064 diseases and 11,615 genes (5.6 genes per disease on average).

6. Functional-coherency based selection of top-ranking genes

One of the main issues when working with disease-gene prioritisation is to select the most relevant (top ranked) genes associated with a given disease. The user can select top 1% or 2% highest scoring genes among all the proteins in the PIN as top ranked genes. In GUILDify v2.0, we also introduced a cutoff based on the functional validation approach described in Ghiassian *et al.* [5] and provided a new panel visualising the significance of the functional enrichment (P-value) as a function of the number of top-ranking genes included in the validation (implemented in Plotly). In brief, the highest-scoring non-seed proteins are iteratively included in the top-ranking set, provided that they maintain the functional coherency of the existing top-ranking set (see details in Supplementary Material). Note that this approach might be too restrictive for some complex diseases in which the information on known disease-gene associations is limited, failing to represent the functional diversity involved in the disease.

7. Visualisation of the top-ranking subnetwork

GUILDify v2.0 uses the JavaScript-based network visualisation library, Cytoscape.js [33], to show the subnetwork of the top-ranking proteins and the drugs targeting these proteins. The user can decide the cutoff to define the top ranked proteins to be visualised (top 1%, top 2% or functionally-coherent as mentioned above). In addition to seeds (green hexagons), top-ranking proteins (yellow circles) and drugs (blue diamonds), the subnetwork includes the proteins that connect the seeds to the largest connected component induced by seeds (named “linkers” and shown as grey circles, see details in Supplementary Material).

8. R package

We have included an R package in order to provide programmatic access to GUILDify v2.0 through R statistical computing environment (https://www.r-project.org/). The package implements methods to query and retrieve results from the web server as an R data frame, allowing users to run multiple queries for more high-throughput and/or systematic analyses. The package and documentation are available online at: http://sbi.upf.edu/guildify2.

GUILDify v2.0 workflow

1. Input

The interface of GUILDify v2.0 is designed to be simple and intuitive. The input varies slightly depending on the desired task: (i) a new search; (ii) retrieving results from a previous run; and (iii) calculating genetic and functional overlap between two previous runs. For a new search, we require two steps: first the selection of seeds (genes associated with a phenotype or drug) and second the selection of parameters to run the prioritisation algorithms. For the selection of seeds the user has to provide: (i) either keyword(s) describing the phenotype/drug of interest or a set of specific gene names separated by a semicolon; (ii) the species of interest (default value: *Homo sapiens*); (iii) the tissue of interest (default value: *All*); and (iv) the PIN source (default value: BIANA). If the user provides a keyword (or set of keywords) describing a phenotype or drug, the server searches genes containing the keyword in BIANA knowledge database (i.e. integrating information from many resources), otherwise it uses the list of provided gene names. The server shows the selected seeds, which can still be filtered and selected by the user. Then, for the prioritisation parameters the user can select to run the "disease module detection algorithm" (DIAMOnD, downloaded from https://github.com/dinaghiassian/DIAMOnD) [5] or to use one of the several prioritisation algorithms from the GUILD package (default value: NetScore with default parameters). Finally, to retrieve results, the required input is the job ID of a previous run, while for calculating genetic and functional overlap the inputs are two job IDs of previous runs.

2. Output

GUILDify v2.0 outputs the ranking of the nodes in the PIN and the visualisation of the subnetwork involving the top-ranking genes in a cytoscape.js panel. In addition, the output page has: (i) a panel showing the P-values of functional enrichment of the ranked nodes; (ii) two panels with functions enriched among the top-ranking nodes and seeds, respectively; and (iii) one panel with the drugs that target the top-ranking proteins.

For the "*Overlap between two results*" option, the server provides: (i) the list of the common top-ranking genes and the significance of the overlap assessed by a Fisher's exact test (see details in Supplementary Material); (ii) the network visualisation of the common top-ranking genes including the "linkers" (see above); (iii) the list of enriched functions of the common genes; iv) the list of common enriched functions of both results and the significance of the overlap; and v) the drugs targeting the proteins of the common PIN. Using this functionality, the users can identify the overlap between any two queries such as between two diseases, two drugs or a disease and a drug. Although we do not provide the overlap between interactions of top-ranking proteins in a separate table, these interactions can be investigated in the network visualisation panel.

Case studies

1. Exploring the mechanistic links between rheumatoid arthritis and asthma

In multiple studies, rheumatoid arthritis and asthma are linked as a potential comorbidity, although the mechanisms underlying this association remain unclear [34]. Using the new functionality of GUILDify v2.0, we can assess the overlap between diseases and thus propose a potential mechanism to explain the association between them. Querying for “rheumatoid arthritis” and “asthma” returns 156 and 96 seeds, respectively coming from DisGeNET, OMIM, and UniProt. There are already 12 seeds in common (Fisher’s exact test, one-sided P-value = $1.4{\cdot}10^{-9}$) and 18 common functions out of the total enriched functions of the seeds (P-value = $9.3{\cdot}10^{-23}$). After running GUILDify v2.0, we select 290 and 181 top ranked genes using functional-coherency based cutoff for rheumatoid arthritis and asthma, respectively. We find that the number of common genes increases to 55 (yielding a P-value = $5.9{\cdot}10^{-48}$), while the number of common functions (biological processes) increases to 31 (P-value = $8.1{\cdot}10^{-46}$). The link between these diseases is significant even when the seeds are removed from the top-ranking genes (see Supplementary Material). Among the shared top-ranking genes, we find Tumor Necrosis Factor (TNF), which has been proposed as a potential drug target for asthma and rheumatoid arthritis, and highlighted as a potential precursor of the comorbidity [12]. We also find HLA-DRB1 and several interleukins (IL18, IL1B, IL3), taking part of the immune response potentially involved in both diseases. Furthermore, the most common enriched functions relate to inflammatory processes such as “inflammatory response”, “positive regulation of interferon-gamma production” and “positive regulation of T-helper 1 cell cytokine production”. These functions appear again if we check the functions enriched by the common genes, along with other functions such as “T-helper 1 type immune response” or “negative regulation of type 2 immune response”, highlighting the involvement of type 1 immune response in both diseases. As negative controls, we repeated the analysis using other disease pairs that are not likely to be comorbid such as “rheumatoid arthritis” - “breast cancer” and “asthma” - “breast cancer”, finding drastically reduced number of genes in the overlap between these disease pairs (see Supplementary Material). The results can be further explored in Figure 1 and in the pre-calculated examples section of the web. Additionally, we compared the functional relevance of the top-ranking genes identified by NetScore with DIAMOnD, based on the analysis in *Sharma et al.* [35] (see Supplementary Material). We checked the enrichment of top-ranking genes among the pathways containing the seed genes of asthma and rheumatoid arthritis, showing that both methods significantly recover the pathways in each disease. Furthermore, NetScore identified more genes that belonged to the pathways shared between asthma and rheumatoid arthritis compared to DIAMOnD.

2. Study of the mechanism of non-small cell lung carcinoma drugs

Non-small cell lung cancer (NSCLC) is the most common type of lung cancer. Typically induced by exposure to toxic substances, the NSCLC pathology has been specially associated with a mutation in the Epidermal Growth Factor Receptor (EGFR) [36]. In a recent study, 9 drugs were proposed to treat this disease [37], 6 of them having drug-target interactions reported: Afatinib, Ceritinib, Crizotinib, Erlotinib, Gefitinib and Palbociclib. Given that we can now identify potentially new relationships between drugs and diseases using drugs as queries, we investigate whether the neighbourhood of the targets of these drugs in the PIN significantly overlaps with the neighbourhood of the genes associated with NSCLC. We used GUILDify v2.0 to define this neighbourhood. We observe that the genetic overlap is always significant, except for one of the drugs (Palbociclib, see Table 1).

We confirm the significance by applying the same approach to breast cancer, showing that Ceritinib, Crizotinib and Palbociclib produce a significant genetic overlap, although the number of common genes in each case is substantially lower than it is in NSCLC (see Table 1). These results are consistent with the fact that Palbociclib is primarily indicated for breast cancer and it has been recently repurposed for NSCLC [38]. The small but significant overlap of Ceritinib and Crizotinib suggests that these two drugs might also be considered as potential repurposing candidates. We note that using the top-ranking nodes increases the significance of the genetic overlap (with lower P-values) compared to the overlap using only seeds (genes associated with a pathophenotype and direct targets of drugs). The significant overlap between the top ranked genes identified using these drugs and the top ranked genes for NSCLC (but not for the top ranked genes for breast cancer) suggests that GUILDify v2.0 can help understanding how drugs exert their action on certain diseases. Indeed, the characterisation of the neighbourhood in the PIN that is affected by drugs opens a wide range of possibilities for drug repurposing research.

Methods

Datasets

GUILDify v2.0 uses BIANA [31] for the integration of biological interaction databases with information on drugs, genes, proteins, functions, pathways and diseases. To create the tissue-specific PINs, we use the RNAseq data from GTEx V7 [22]. Phenotype-gene associations are extracted from DisGeNET, OMIM, Uniprot, and Gene Ontology. Drug-target associations are taken from DrugBank [17], DGIdb [18], DrugCentral [19] and ChEMBL [20]. See Supplementary Material for details on the datasets.

Prioritisation algorithms

GUILDify v2.0 uses four different network-based prioritisation algorithms: NetShort, NetZcore, NetScore and DIAMOnD. For details on these algorithms see references [5,10,11] and the Supplementary Material.

Overlap and functional enrichment analysis

We use one-sided Fisher’s exact test to calculate the overlap between two sets of genes or functions and use Benjamini-Hochberg multiple hypothesis testing procedure (where applicable). The functions enriched among seeds and top-ranking nodes as well as common functions between two diseases are calculated as explained in a previous work [12] (see details in Supplementary Material).

Conflicts of Interest Statement

None declared.

Acknowledgements

J.A.P. and E.G. would like to acknowledge the technical support from GRIB IT team, in particular A. Gonzalez Pauner and M. A. Sánchez Gómez.

The authors received support from: ISCIII-FEDER (PI13/00082, CP10/00524, CPII16/00026); IMI-JU under grants agreements no. 116030 (TransQST) and no. 777365 (eTRANSAFE), resources of which are composed of financial contribution from the EU-FP7 (FP7/2007- 2013) and EFPIA companies in kind contribution; the EU H2020 Programme 2014-2020 under grant agreements no. 634143 (MedBioinformatics) and no. 676559 (Elixir-Excelerate); the Spanish Ministry of Economy (MINECO) [BIO2017-85329-R] [RYC-2015-17519]; “Unidad de Excelencia María de Maeztu”, funded by the Spanish Ministry of Economy [ref: MDM-2014-0370]. The Research Programme on Biomedical Informatics (GRIB) is a member of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant PT13/0001/0023, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER.

References

[1] M.F. Wangler, S. Yamamoto, H.-T. Chao, J.E. Posey, M. Westerfield, J. Postlethwait, P. Hieter, K.M. Boycott, P.M. Campeau, H.J. Bellen, Model Organisms Facilitate Rare Disease Diagnosis and Therapeutic Research, Genetics. 207 (2017) 9–27. doi:10.1534/genetics.117.203067.

[2] Y. Bromberg, Chapter 15: Disease Gene Prioritization, PLoS Comput. Biol. 9 (2013) e1002902. doi:10.1371/journal.pcbi.1002902.

[3] J. Menche, A. Sharma, M. Kitsak, S.D. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabási, Uncovering disease-disease relationships through the incomplete interactome, Science (80-.). 347 (2014) 1257601-1-1257601–8. doi:10.1126/science.1257601.

[4] X. Wang, N. Gulbahce, H. Yu, Network-based methods for human disease gene prediction, Brief. Funct. Genomics. 10 (2011) 280–293. doi:10.1093/bfgp/elr024.

[5] S.D. Ghiassian, J. Menche, A.L. Barabási, A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome, PLoS Comput. Biol. 11 (2015) e1004120. doi:10.1371/journal.pcbi.1004120.

[6] D. Nitsch, L.C. Tranchevent, J.P. Gonalves, J.K. Vogt, S.C. Madeira, Y. Moreau, PINTA: A web server for network-based gene prioritization from expression data, Nucleic Acids Res. 39 (2011) 334–338. doi:10.1093/nar/gkr289.

[7] K. Zuberi, M. Franz, H. Rodriguez, J. Montojo, C.T. Lopes, G.D. Bader, Q. Morris, GeneMANIA prediction server 2013 update., Nucleic Acids Res. 41 (2013) 115–122. doi:10.1093/nar/gkt533.

[8] A. Gottlieb, O. Magger, I. Berman, E. Ruppin, R. Sharan, Principle: A tool for associating genes with diseases via network propagation, Bioinformatics. 27 (2011) 3325–3326. doi:10.1093/bioinformatics/btr584.

[9] T. Kacprowski, N.T. Doncheva, M. Albrecht, NetworkPrioritizer: A versatile tool for network-based prioritization of candidate disease genes or other molecules, Bioinformatics. 29 (2013) 1471–1473. doi:10.1093/bioinformatics/btt164.

[10] E. Guney, J. García-garcía, B. Oliva, GUILDify : A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms, Bioinformatics. 30 (2014) 1789–1790. doi:10.1093/bioinformatics/btu092.

[11] E. Guney, B. Oliva, Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization, PLoS One. 7 (2012) e43557. doi:10.1371/journal.pone.0043557.

[12] C. Rubio-Perez, E. Guney, D. Aguilar, J. Piñero, J. Garcia-Garcia, B. Iadarola, F. Sanz, N. Fernandez-Fuentes, L.I. Furlong, B. Oliva, Genetic and functional characterization of disease associations explains comorbidity, Sci. Rep. 7 (2017) 6207. doi:10.1038/s41598-017-04939-4.

[13] F. Halakou, E. Sen Kilic, E. Cukuroglu, O. Keskin, A. Gursoy, Enriching Traditional Protein-protein Interaction Networks with Alternative Conformations of Proteins, Sci. Rep. 7 (2017) 7180. doi:10.1038/s41598-017-07351-0.

[14] A. Kominakis, A.L. Hager-Theodorides, E. Zoidis, A. Saridaki, G. Antonakos, G. Tsiamis, Combined GWAS and 'guilt by association'-based prioritization analysis identifies functional candidate genes for body size in sheep, Genet. Sel. Evol. 49 (2017) 41. doi:10.1186/s12711-017-0316-3.

[15] E. Tejera, M. Cruz-Monteagudo, G. Burgos, M.E. Sánchez, A. Sánchez-Rodríguez, Y. Pérez-Castillo, F. Borges, M.N.D.S. Cordeiro, C. Paz-Y-Miño, I. Rebelo, Consensus strategy in genes prioritization and combined bioinformatics analysis for preeclampsia pathogenesis, BMC Med. Genomics. 10 (2017) 50. doi:10.1186/s12920-017-0286-x.

[16] J. Aguirre-Plans, J. Piñero, J. Menche, F. Sanz, L.I. Furlong, H.H.H.W. Schmidt, B. Oliva, E. Guney, Proximal pathway enrichment analysis for targeting comorbid diseases via network endopharmacology, Pharmaceuticals. 11 (2018) 61. doi:10.3390/ph11030061.

[17] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maclejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, Di. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: A major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082. doi:10.1093/nar/gkx1037.

[18] K.C. Cotto, A.H. Wagner, Y. Feng, S. Kiwala, C. Coffman, G. Spies, A. Wollam, N.C. Spies, O.L. Griffith, M. Griffith, DGIdb 3.0 : a redesign and expansion of the drug-gene interaction database, Nucleic Acids Res. 46 (2018) 1068–1073. doi:10.1093/nar/gkx1143.

[19] O. Ursu, J. Holmes, J. Knockel, C.G. Bologa, J.J. Yang, S.L. Mathias, S.J. Nelson, T.I. Oprea, DrugCentral : online drug compendium, Nucleic Acids Res. 45 (2017) 932–939. doi:10.1093/nar/gkw993.

[20] A. Gaulton, A. Hersey, A. Patr, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L.J. Bellis, E. Cibri, M. Davies, N. Dedman, A. Karlsson, P. Magari, J.P. Overington, G. Papadatos, I. Smit, The ChEMBL database in 2017, Nucleic Acids Res. 45 (2017) 945–954. doi:10.1093/nar/gkw1074.

[21] M. Kitsak, A. Sharma, J. Menche, E. Guney, S.D. Ghiassian, J. Loscalzo, A.L. Barabási, Tissue Specificity of Human Disease Module, Sci. Rep. 6 (2016) 35241. doi:10.1038/srep35241.

[22] G. Consortium, Genetic effects on gene expression across human tissues, Nature. 550 (2017) 204–213. doi:10.1038/nature24277.

[23] J. Piñero, Á. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, L.I. Furlong, DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res. 45 (2017) D833–D839. doi:10.1093/nar/gkw943.

[24] The UniProt Consortium, UniProt: The universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169. doi:10.1093/nar/gkw1099.

[25] A.P. Davis, C.J. Grondin, R.J. Johnson, D. Sciaky, B.L. King, R. McMorran, J. Wiegers, T.C. Wiegers, C.J. Mattingly, The Comparative Toxicogenomics Database: Update 2017, Nucleic Acids Res. 45 (2017) D972–D978. doi:10.1093/nar/gkw838.

[26] M.J. Landrum, J.M. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, J. Hoover, W. Jang, K. Katz, M. Ovetsky, G. Riley, A. Sethi, R. Tully, R. Villamarin-Salomon, W. Rubinstein, D.R. Maglott, ClinVar: Public archive of interpretations of clinically relevant variants, Nucleic Acids Res. 44 (2016) D862–D868. doi:10.1093/nar/gkv1222.

[27] A. Rath, A. Olry, F. Dhombres, M.M. Brandt, B. Urbero, S. Ayme, Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users, Hum. Mutat. 33 (2012) 803–808. doi:10.1002/humu.22078.

[28] J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. MayPendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, H. Parkinson, The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog), Nucleic Acids Res. 45 (2017) D896–D901. doi:10.1093/nar/gkw1133.

[29] A. Gutiérrez-Sacristán, À. Bravo, M. Portero, O. Valverde, A. Armario, M.C. Blanco-Gandía, A. Farré, L. Fernández-

Ibarrondo, F. Fonseca, J. Giraldo, A. Leis, A. Mané, M.A. Mayer, S. Montagud-Romero, R. Nadal, J. Ortiz, F.J. Pavón, E. Perez, M. Rodríguez-Arias, A. Serrano, M. Torrens, V. Warnault, F. Sanz, L.I. Furlong, Text mining and expert curation to develop a database on psychiatric diseases and their genes, Database. 1650 (2017) 48–55. doi:10.1093/database/bax043.

[30] S. Köhler, N.A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S.M. Bello, C.F. Boerkoel, K.M. Boycott, M. Brudno, O.J. Buske, P.F. Chinnery, V. Cipriani, L.E. Connell, H.J.S. Dawkins, L.E. DeMare, A.D. Devereau, B.B.A. De Vries, H. V. Firth, K. Freson, D. Greene, A. Hamosh, I. Helbig, C. Hum, J.A. Jähn, R. James, R. Krause, S.J.F. Laulederkind, H. Lochmüller, G.J. Lyon, S. Ogishima, A. Olry, W.H. Ouwehand, N. Pontikos, A. Rath, F. Schaefer, R.H. Scott, M. Segal, P.I. Sergouniotis, R. Sever, C.L. Smith, V. Straub, R. Thompson, C. Turner, E. Turro, M.W.M. Veltman, T. Vulliamy, J. Yu, J. Von Ziegenweidt, A. Zankl, S. Züchner, T. Zemojtel, J.O.B. Jacobsen, T. Groza, D. Smedley, C.J. Mungall, M. Haendel, P.N. Robinson, The human phenotype ontology in 2017, Nucleic Acids Res. 45 (2017) D865–D876. doi:10.1093/nar/gkw1039.

[31] J. Garcia-Garcia, E. Guney, R. Aragues, J. Planas-Iglesias, B. Oliva, Biana: a software framework for compiling biological interactions and analyzing networks, BMC Bioinformatics. 11 (2010) 56. doi:10.1186/1471-2105-11-56.

[32] W.A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C.J. Mungall, J.X. Binder, J. Malone, D. Vasant, H. Parkinson, L.M. Schriml, Disease Ontology 2015 update : an expanded and updated database of human diseases for linking biomedical knowledge through disease data, Nucleic Acids Res. 43 (2015) 1071–1078. doi:10.1093/nar/gku1011.

[33] M. Franz, C.T. Lopes, G. Huck, Y. Dong, O. Sumer, G.D. Bader, Cytoscape.js: A graph theory library for visualisation and analysis, Bioinformatics. 32 (2015) 309–311. doi:10.1093/bioinformatics/btv557.

[34] M.C. Rolfes, Y.J. Juhn, S.I. Wi, Y.H. Sheen, Asthma and the risk of rheumatoid arthritis: An insight into the heterogeneity and phenotypes of asthma, Tuberc. Respir. Dis. (Seoul). 80 (2017) 113–135. doi:10.4046/trd.2017.80.2.113.

[35] A. Sharma, J. Menche, C. Chris Huang, T. Ort, X. Zhou, M. Kitsak, N. Sahni, D. Thibault, L. Voung, F. Guo, S.D. Ghiassian, N. Gulbahce, F. Baribaud, J. Tocker, R. Dobrin, E. Barnathan, H. Liu, R.A. Panettieri, K.G. Tantisira, W. Qiu, B.A. Raby, E.K. Silverman, M. Vidal, S.T. Weiss, A.L. Barabási, A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma, Hum. Mol. Genet. 24 (2014) 3005–3020. doi:10.1093/hmg/ddv001.

[36] G. Bethune, D. Bethune, N. Ridgway, Z. Xu, Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update, J. Thorac. Dis. 2 (2010) 48–51. doi:10.3978/j.issn.2072-1439.2010.02.01.017.

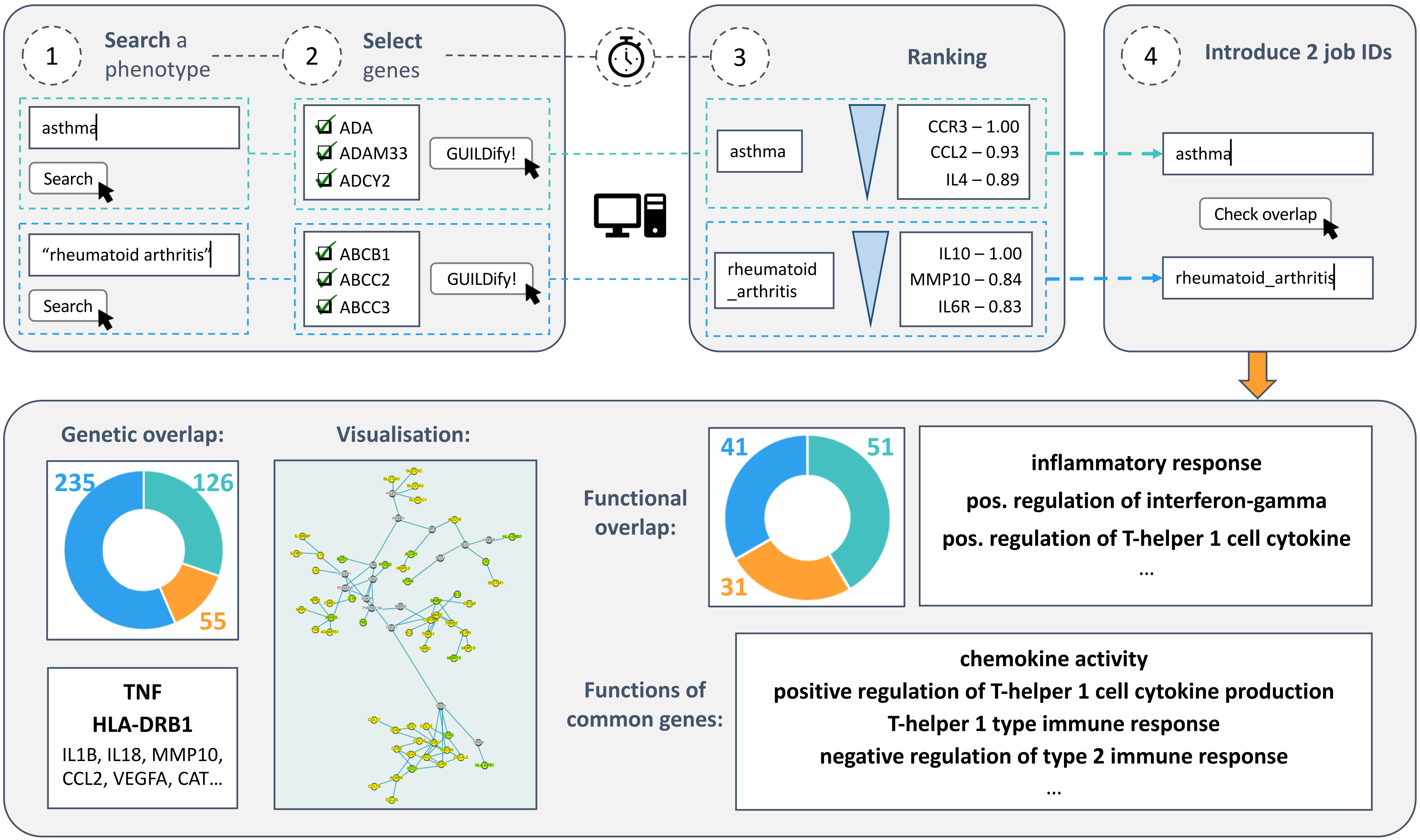
[37] C. Rubio-Perez, D. Tamborero, M.P. Schroeder, A.A. Antolín, J. Deu-Pons, C. Perez-Llamas, J. Mestres, A. Gonzalez-Perez, N. Lopez-Bigas, In Silico Prescription of Anticancer Drugs to Cohorts of 28 Tumor Types Reveals Targeting Opportunities, Cancer Cell. 27 (2015) 382–396. doi:10.1016/j.ccell.2015.02.007.

[38] J. Zhou, S. Zhang, X. Chen, X. Zheng, Y. Yao, G. Lu, J. Zhou, Palbociclib, a selective CDK4/6 inhibitor, enhances the effect of selumetinib in RAS-driven non-small cell lung cancer, Cancer Lett. 408 (2017) 130–137. doi:10.1016/j.canlet.2017.08.031.

TABLE AND FIGURES LEGENDS

Figure 1. GUILDify v2.0 example study on the comorbidity between asthma and rheumatoid arthritis. First, we run the prioritisations of the two diseases by searching (1) and selecting (2) the genes. After obtaining the ranking of proteins from the prioritisation (3), we use both job IDs to check their overlap (4) and inspect the genetic and functional relationships between them (see details at http://sbi.upf.edu/guildify2 in the pre-calculated examples section.

Table 1. Results of the genetic and functional overlap between the subnetwork of genes associated with "non small cell lung carcinoma" and "breast cancer" (top ranking genes and seeds) and the subnetwork of genes associated with the targets of drugs Afatinib, Ceritinib, Crizotinib, Erlotinib, Gefitinib and Palbociclib (drug targets and top-ranking genes obtained with GUILDify v2.0). P-values shown have been corrected using the Benjamini-Hochberg correction for multiple tests. Results with non-significant P-value are highlighted in red.


1
Search a phenotype
2
Select genes
3
Ranking
4
Introduce 2 job IDs
asthma
Search
"rheumatoid arthritis"
Search
☑ ADA
☑ ADAM33
☑ ADCY2
GUILDify!
☑ ABCB1
☑ ABCC2
☑ ABCC3
GUILDify!
asthma
CCR3 – 1.00
CCL2 – 0.93
IL4 – 0.89
rheumatoid_arthritis
IL10 – 1.00
MMP10 – 0.84
IL6R – 0.83
asthma
Check overlap
rheumatoid_arthritis
Genetic overlap:
235
126
55
TNF
HLA-DRB1
IL1B, IL18, MMP10, CCL2, VEGFA, CAT…
Visualisation:
Functional overlap:
41
51
31
inflammatory response
pos. regulation of interferon-gamma
pos. regulation of T-helper 1 cell cytokine
…
Functions of common genes:
chemokine activity
positive regulation of T-helper 1 cell cytokine production
T-helper 1 type immune response
negative regulation of type 2 immune response
…

	“non-small cell lung carcinoma”								“breast cancer”							
	Genetic overlap				Functional overlap				Genetic overlap				Functional overlap			
	Top		Seeds		Top		Seeds		Top		Seeds		Top		Seeds	
	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.	Nº	P-val.
Afatinib	9	7.80E-06	2	1.90E-03	0	1	6	1.70E-05	3	1.40E-01	0	1	0	1	1	7.50E-01
Ceritinib	18	4.20E-15	4	6.60E-05	5	1.70E-07	9	1.80E-06	6	1.20E-02	0	1	0	1	0	1
Crizotinib	13	1.20E-09	4	7.00E-05	3	1.10E-04	8	7.70E-06	5	2.00E-02	0	1	0	1	0	1
Erlotinib	16	6.30E-13	4	6.60E-05	5	1.20E-06	10	2.60E-07	3	1.40E-01	0	1	0	1	0	1
Gefitinib	10	1.10E-06	3	1.20E-04	1	3.60E-02	11	7.20E-10	3	1.40E-01	0	1	0	1	0	1
Palbociclib	3	1.40E-01	0	1	0	1	2	8.90E-02	5	2.00E-02	1	4.70E-01	0	1	1	7.50E-01

Supplementary material for:
GUILDify v2.0: A tool to identify molecular networks underlying human diseases, their comorbidities and their druggable targets

Joaquim Aguirre-Plans, Janet Piñero, Ferran Sanz, Laura I. Furlong, Narcis Fernandez-Fuentes, Baldo Oliva* and Emre Guney*

1. Datasets

1.1. Protein-protein interaction networks

GUILDify v2.0 relies on a knowledge database called BIANA [1], which integrates biological interaction databases together with information on genes and proteins and its associated functions, diseases and phenotypes. Currently, we have up-to-date information of protein-protein interactions from IntAct [2], BioGRID [3] and DIP [4]. As an additional option, we provide the user with 5 other PIN sources: HIPPIE (high confidence score threshold >= 0.7) [5], InBio_Map (score threshold >= 0.15) [6], ConsensusPathDB [7], I2D [8], and STRING (score >= 0.7) [9].

1.2. Tissue-specific protein-protein interaction networks

To create the tissue-specific PINs, we retrieved the RNA-sequencing gene TPMs from GTEx V7 [10]. We use the samples from subjects for which the reason for death was traumatic injury (point 1 in Hardy Scale) and we discard the tissues with less than 5 samples (a total of 675 samples from 40 tissues). For each gene, we calculate the median expression of all samples of a tissue. We unify tissues into a unique "main" tissue (i.e. "Adipose – Subcutaneous" and "Adipose – Visceral Omentum" belong to the main tissue "Adipose") by considering the highest median expression of all samples [11]. We note that using this approach, the final tissue profiles could be biased towards the subtypes of tissue that have a higher number of samples. This is a limitation, as there may be subtypes of tissue that are more represented than others in the final expression of the tissue (Supplementary Table S1). GUILDify v2.0 includes a total of 22 tissues (see details in Supplementary Material).

1.3. Phenotype-gene associations

Phenotype-gene associations are extracted from DisGeNET, OMIM, Uniprot and Gene Ontology extracting information from relevant sections. In the case of DisGeNET, we parse disease-gene associations from curated sources: UniProt [12], CTD [13], ORPHANET [14], PsyGeNET [15] and HPO [16]. In OMIM [17], we retrieve disease-gene associations from the OMIM's Synopsis of the Human Gene Map. For Uniprot [30], we collect the protein information of the categories "Description", "Function", "Keyword" and "Disease". Finally, in the case of the Gene Ontology (GO) [18,19], we parse the functional annotations of genes.

1.4. Drug-target integration

Drug-target interactions are retrieved from DrugBank [20], DGIdb [21], DrugCentral [22] and ChEMBL [23], and integrated following the procedure in *Piñero et al.* [24]. In DrugBank, we only select the therapeutic targets (excluding enzymes, transporters and carriers). In the case of data from DrugCentral, we retrieve the targets in the "Tclin" category. From DGIdb we select the targets from "Chembl", "GuideToPharmacology", "Tdg Clinical Trial", "FDA", "TEND" and "TTD". Finally in ChEMBL, we collect targets with a DrugBank identifier cross-reference.

2. Prioritisation algorithms

GUILDify v2.0 uses four different network-based prioritisation algorithms: NetShort, NetZcore, NetScore and DIAMOnD. For details on these algorithms see references [25–27]. In brief, **NetShort** (10-20 minutes of computation time) incorporates "phenotypic-relevance" of the path between a node and the nodes of a given phenotype by considering the number of edges to phenotype-associated nodes (seeds). **NetZcore** (5-10 minutes of computation time) iteratively assesses the relevance of a node for a given phenotype by averaging the normalised scores of the neighbours. **NetScore** (5-10 minutes of computation time) is based on the propagation of information through the nodes of the network by considering multiple shortest paths from the source of information to the target. **NetCombo** (10-20 minutes of computation time) combines NetScore, NetShort and NetZcore by calculating the mean of the normalised score of each prioritisation method. **DIAMOnD** [27] (5 minutes of computation time) determines the "connectivity significance" of all the proteins of the network, iteratively ranking and selecting the nodes with highest scores.

3. Functional enrichment analysis

We calculate the enriched functions of seeds and top-ranking nodes and calculate the significance of common functions between two diseases (or phenotypes) as in a previous work on comorbidities [28]. Briefly, functions are defined by GO biological processes, GO molecular functions and Reactome pathways. In the case of GO, we only use high confident annotations (codes of evidence EXP, IDA, IMP, IGI, IEP, ISS, ISA, ISM or ISO). We calculate the significance of the enrichment using a one-sided Fisher's exact test (the alternative hypothesis is that the overlap would be greater than observed overlap). Then, we correct the P-value by either applying the Benjamini-Hochberg correction for multiple tests and keeping the functions for which the adjusted P-value < 0.05. We also offer the user the possibility to use Bonferroni correction at the results page.

4. Description of BIANA integration pipeline

We used BIANA [1] to compile different types of biological data in an integrated database and to create the protein-protein interaction networks (PIN). The information in BIANA is updated annually to keep the resources underlying the web server up-to-date. The details of data retrieval, integration, unification and network generation pipeline are as follows:

1. **Download the data**

We use five sources of protein-protein interaction data:

- **IntAct**: retrieved from https://www.proteinatlas.org/download/normal_tissue.tsv.zip (Release of 22-Mar-2018).
- **BioGRID**: downloaded from https://downloads.thebiogrid.org/BioGRID (Version 3.4.159).
- **DIP**: downloaded from http://dip.doe-mbi.ucla.edu/dip/Download.cgi (Release of 05-Feb-2017).
- **iRefIndex**: downloaded from http://irefindex.org/download/irefindex/data/archive/release_14.0/psi_mitab/MITAB2.6/ (Version 15.0).
- **HIPPIE**: downloaded from http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/download.php (Version 2.1).

And we also incorporate additional databases to complement interactomics data:

- **UniProt swissprot**: retrieved from ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/ (Release of 28-Mar-2018).
- **Taxonomy**: downloaded from ftp://ftp.ncbi.nih.gov/pub/taxonomy (Release of 19-Apr-2018).
- **Gene Ontology**: downloaded from http://www.geneontology.org/ontology/ (Release of 19-Apr-2018).
- **NCBI Gene**: downloaded from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2ensembl.gz (Release of 28-Nov-2017).
- **DisGeNET**: downloaded from http://www.disgenet.org/web/DisGeNET/menu/downloads (Version 5.0).
- **DrugBank**: downloaded from https://www.drugbank.ca/releases/latest (Version 5.1.0).
- **DrugCentral**: downloaded from http://drugcentral.org/download (Release of 29-Aug-2017).
- **DGIdb**: downloaded from http://dgidb.org/downloads (Version 3.0.2).
- **ChEMBL**: (Version ChEMBL_24) downloaded from:
  - https://www.ebi.ac.uk/chembl/drug/targets > Downloads > Download all txt
  - https://www.ebi.ac.uk/chembl/drugstore > Downloads > Download all txt
  - https://www.ebi.ac.uk/chembl/drug/indications > Downloads > Download all txt
  - https://www.ebi.ac.uk/chembl/target/browser > Select All > Fetch selected targets > Please select… > Download All (tab-delimited)

2. Parse and unify the data

We parse all the external databases according to the manual of BIANA (available at http://sbi.imim.es/web/BIANA.php). Once we have all the databases incorporated in BIANA knowledge database, we find the equivalent entries across databases to unify the data. This means that two entities coming from different databases (or in some cases from the same database) can be unified in a unique entity provided that they satisfy certain equivalence criteria.

If they do so, they will be given the same unique ID called **BIANA ID**. The rules to unify data (equivalence criteria) are the following:

- Same **Entrez Gene ID** (applied to all databases)
- Same **Taxonomy ID AND protein sequence** (applied to all databases)
- Same **UniProt entry** (only applied to ConsensusPathDB and Uniprot databases)
- Same **UniProt accession** (applied to InBio_Map, I2D, HitPredict and Uniprot)
- Same **UniProt accession** (applied to DrugBank, DrugCentral, ChEMBL and Uniprot databases, to unify the drug targets with the rest of proteins)
- Same **DrugBank ID** (applied to DrugBank, DCDB and DrugCentral to unify drugs)
- Same **PubChem Compound** (applied to DrugBank and DCDB to unify drugs)
- Same **ChEMBL ID** (applied to DrugBank, DGIdb and ChEMBL to unify drugs)

3. Generate protein-protein interaction networks (PIN)

To generate PINs, we retrieve the protein-protein interactions from BIANA knowledge database that have the same Taxonomy ID (reported in the same organism) for human, mouse, rat, yeast, worm, fly and plant.

Once we have the PIN, we filter the interactions depending on the detection method used to characterise each interaction. Recent studies highlighted that several protein interaction detection techniques tended to provide a higher number of interactions for more studied proteins in the interactome [29,30]. Thus, for human, we only include interactions coming from the following detection methods that we consider to be less biased:

- Two hybrid (ID: 18).
- Cross-linking study (ID: 30).
- Protein array (ID: 89).
- Two hybrid array (ID: 397).
- Two hybrid pooling approach (ID: 398).
- Biochemical (ID: 401).
- Enzymatic study (ID: 415).
- Two hybrid prey pooling approach (ID: 1112).
- Proximity labelling technology (ID: 1313).
- Validated two hybrid (ID: 1356).

For mouse, we included all the methods listed above and the following ones:

- Affinity chromatography technology (ID: 4)
- Anti tag coimmunoprecipitation (ID: 7)
- Coimmunoprecipitation (ID: 19)
- Cosedimentation in solution (ID: 28)
- Pull down (ID: 96)

- X-ray crystallography (ID: 375)
- Chromatin immunoprecipitation assay (ID: 810)
- Tandem affinity purification (ID: 676)

5. Description of tissue-specific PIN generation pipeline

1. Download the data

We used the RNA-Seq data from GTEx Portal version 7, downloaded from https://gtexportal.org/home/datasets.

2. Process the data

- Process the subjects file (GTEx_v7_Annotations_SubjectPhenotypesDS.txt) and get the subjects with cause of death by traumatic injury (DTHHRDY=1). We end up with 29 subjects.
- Process the samples file (GTEx_v7_Annotations_SampleAttributesDS.txt) and get the samples coming from the subjects filtered in the previous step. We end up with 699 samples.
- Count the tissues present in the samples and the number of samples for each tissue. Remove the tissues that have less than 5 samples. We end up having 40 tissues and 675 samples.
- Read the TPM file (GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_tpm.gct.gz) and get the TPM values of the 675 samples mentioned in the previous step.
- For each gene in the TPM file, calculate the median of TPM values across all the samples in each tissue.
- If several tissues belong to a more general tissue (i.e. “Adipose – Subcutaneous” and “Adipose – Visceral Omentum” belong to the main tissue “Adipose”), we unify them by considering the highest median expression value among these tissues. After unifying the tissues, we end up having 22 main tissues. The tissues unified are listed in Supplementary Table S1.

3. Filter the PIN

The last step is to filter the interactions of the PIN based on the tissue annotation of the proteins. For each pair of interacting proteins, first, we check if we have information of the proteins in the GTEx file that we have processed. If so, we get the TPM values for the tissue of interest and the two proteins, and if in both cases the TPM values are higher or equal than 1, we maintain the interaction in the network. If not, we remove the interaction.

Supplementary Table S1. Tissues considered for the creation of tissue-specific PIN. In the left we show the 40 initial tissues and in the right the 22 final tissues after the unification of some of the tissues into a broader one. We included the number of samples considered for each tissue.

GTEx tissues		Unified tissues	
Name	**Num. samples**	**Name**	**Num. samples**
Adipose – Subcutaneous	18	Adipose	28
Adipose – Visceral (Omentum)	10		
Adrenal Gland	5	Adrenal Gland	5
Artery – Aorta	16	Artery	52
Artery – Coronary	9		
Artery – Tibial	27		
Brain – Amygdala	10	Brain	144
Brain – Anterior cingulate cortex (BA24)	12		
Brain – Caudate (basal ganglia)	13		
Brain – Cerebellar Hemisphere	13		
Brain – Cerebellum	16		
Brain – Cortex	14		
Brain – Frontal Cortex (BA9)	8		
Brain – Hippocampus	13		
Brain – Hypothalamus	11		
Brain – Nucleus accumbens (basal ganglia)	12		
Brain – Putamen (basal ganglia)	12		
Brain – Spinal cord (cervical c-1)	10		
Breast – Mammary Tissue	16	Breast – Mammary Tissue	16
Cells – Transformed fibroblasts	20	Cells – Transformed fibroblasts	20
Colon – Sigmoid	8	Colon – Sigmoid	8
Esophagus – Gastroesophageal Junction	8	Esophagus	36
Esophagus – Mucosa	15		
Esophagus – Muscularis	13		
Heart – Atrial Appendage	14	Heart	39
Heart – Left Ventricle	25		
Liver	11	Liver	11
Lung	25	Lung	25
Muscle – Skeletal	31	Muscle – Skeletal	31
Nerve – Tibial	26	Nerve – Tibial	26

Ovary	5	Ovary	5
Pituitary	14	Pituitary	14
Prostate	10	Prostate	10
Skin – Not Sun Exposed (Suprapubic)	16	Skin	44
Skin – Sun Exposed (Lower leg)	28		
Testis	15	Testis	15
Thyroid	25	Thyroid	25
Uterus	5	Uterus	5
Vagina	6	Vagina	6
Whole Blood	110	Whole Blood	110

6. Functional-based selection of top-ranking genes

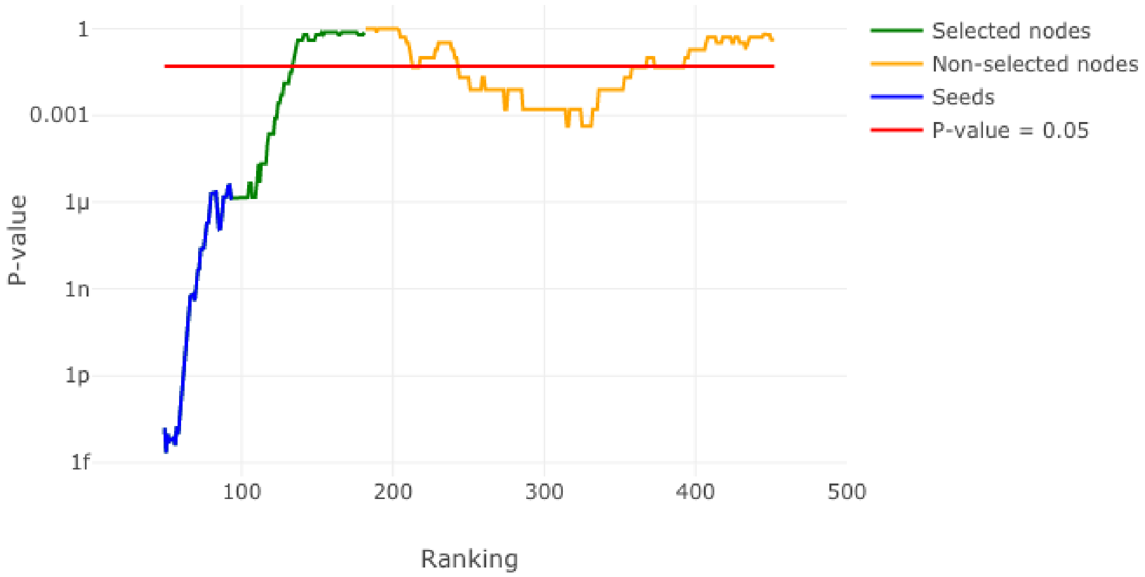
The functional-based selection of top-ranking genes is a procedure that we followed to identify if the set of ranking genes is functionally similar to the set of initial seed genes. The procedure was first implemented in *Ghiassian et al*. [27].

First, we find the enriched Gene Ontology (GO) terms in the seeds by calculating the functional enrichment following the procedure in Rubio-Perez, et al [28]. We only use high confidence annotations from Biological Processes or Molecular Functions associated with the evidence codes EXP, IDA, IMP, IGI, IEP, ISS, ISA, ISM or ISO. From the enriched GO terms, we identify the ones that are significantly enriched using a one-sided Fisher's exact test of significance where the alternative hypothesis is that the overlap would be greater than observed overlap. We correct the significance applying either a Bonferroni or a Benjamini-Hochberg correction for multiple tests and selecting a P-value < 0.05 (the case studies use Benjamini-Hochberg).

For each candidate gene in the top-ranking genes, we search if it is annotated within any of the significant GO terms of the seeds. The genes annotated are considered true positives.

For each candidate gene in the ranking, we define a sliding window with a size corresponding to the number of seeds. For instance, if there are 66 seeds, the interval for top ranking node i will be [i-66/2, i+66/2]. We calculate the number of true positives among the proteins in the sliding window. We calculate the statistical significance in the sliding window by using a Fisher’s exact test.

In the end, we obtain a plot that goes from the first position of the sliding window (ranking = # of seeds / 2 + 1) to the final position (500 - # of seeds/2). In each position, we show the result of the Fisher’s test calculation for the positions of the sliding window. We consider as *enriched positions* all the positions until the last sliding window giving a P-value < 0.05.



Supplementary Figure S1. Functional-based selection plot in the case of the example GUILDify v2.0 run using 96 seeds for *asthma* keyword and the NetScore algorithm with default parameters.

In the figure S1, we observe an example of the functional-based selection plot for *asthma* using 96 seeds. Therefore, the plot starts at position 49 and ends at position 452. The last sliding window with a P-value < 0.05 is at position 133, and ranges from position 85 to 181. We consider as enriched positions all the top-ranking genes until the 181st position, including the 96 initial seeds and 85 additional non-seed genes.

7. Significance of the overlap between genes/functions

To calculate the significance of the overlap between top-ranking genes and their functions, we use a one-sided (the alternative hypothesis is that the odds ratio based on the overlap is greater than the observed odds ratio) Fisher's exact test with the contingency table given in Supplementary Table S2.

Supplementary Table S2. Contingency table used to calculate the significance of the overlap between top-ranking genes and their functions.

	Top 2	Non-top 2
Top 1	Nº common	Nº top 1 – Nº common
Non-top 1	Nº top 2 – Nº common	Nº total – Nº top 1 – Nº top 2 – Nº common

Supplementary Table S3 is an example of the contingency table for the genetic overlap between asthma and rheumatoid arthritis, where we obtain an Odds Ratio of 23.542 and a P-value of $5.9*10^{-48}$.

Supplementary Table S3. Contingency table used in Fisher's exact text for the genetic overlap between asthma and rheumatoid arthritis.

	Top genes 2	Non-top genes 2
Top genes 1	55	290 – 55 = 235
Non-top genes 1	181 – 55 = 126	13,090 – 181 – 290 + 55 = 12,674

The contingency table for the functional overlap of the same example, where we obtain an Odds Ratio of 155.334 and a P-value of $1.3*10^{-58}$ is below (Supplementary Table S4).

Supplementary Table S4. Contingency table used in Fisher's exact text for the functional overlap between asthma and rheumatoid arthritis

	Top functions 2	Non-top functions 2
Top functions 1	38	84 – 38 = 46
Non-top functions 1	94 – 38 = 56	10,670 – 94 – 84 + 38 = 10,530

Supplementary Table S5. Job IDs to access to the case studies in the web server and parameters used.

Keyword	Parameters	Job ID
asthma	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**asthma**
“rheumatoid arthritis”	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**rheumatoid_arthritis**
“non small cell lung carcinoma”	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**non_small_cell_lung_carcinoma**
“breast cancer”	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**breast_cancer**

“breast cancer”	Seeds: DisGeNET and OMIM Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**breast_cancer_omim_disgenet**
afatinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**afatinib**
ceritinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**afatinib**
crizotinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**crizotinib**
erlotinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**erlotinib**
gefitinib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**gefitinib**
palbociclib	Seeds: all genes Organism: Homo sapiens Tissue: All Network: BIANA Method: NetScore (nRepetition=3, nIteration=2)	**palbociclib**

8. Results of the case study of asthma and arthritis rheumatoid and the negative controls with breast cancer

We query “asthma” in GUILDify v2.0, obtaining 96 seeds. After the running GUILD and selecting the functionally-coherent top genes, we obtain 181 genes conforming the neighbourhood of asthma. We do the same for “rheumatoid arthritis”, retrieving 158 seeds and creating a neighbourhood of 290 functionally-coherent top genes. Between the top ranking genes of the two phenotypes there are 55 common genes (Fisher’s exact test, one-sided P-value = $5.9 \cdot 10^{-48}$) which is more significant than the 12 common genes between the seeds (P-value = $1.4\cdot10^{-9}$). When removing the seeds from the top ranking genes of the two phenotypes, we find an overlap of 43 genes which is even more significant than before (P-value = $3.7\cdot10^{-65}$).

If we focus on the functional overlap, we find 38 common enriched functions from the top ranking genes (P-value = $1.3\cdot10^{-58}$), 18 common enriched functions from the seeds (P-value = $1.7\cdot10^{-24}$), and 24 common functions removing the seed-functions from the top-functions (P-value = $3.5\cdot10^{-47}$).

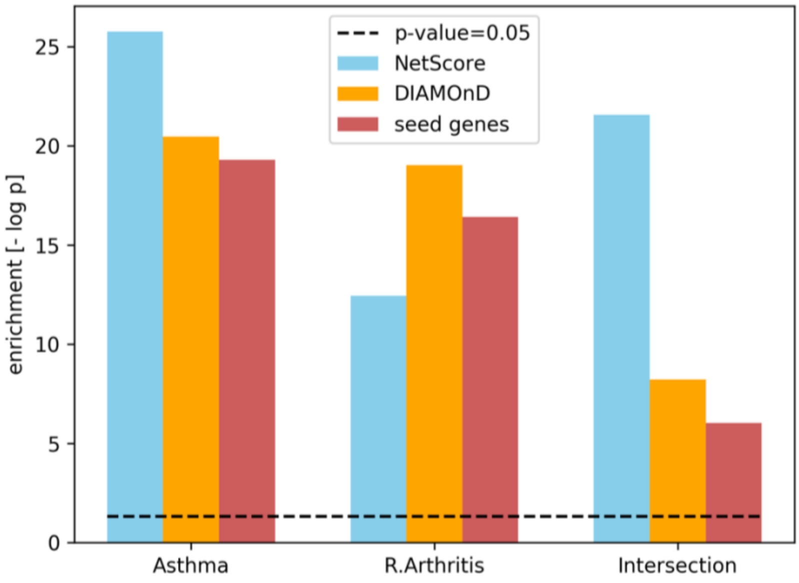
To have a negative control, we query “breast cancer” and retrieve 119 seeds. We select the 182 functionally-coherent top genes and compare the neighbourhood with the phenotypes of asthma and rheumatoid arthritis. In the case of asthma and breast cancer, we observe a significant overlap between 8 genes (P-value = $3.8\cdot10^{-3}$), but the functional overlap is not significant (only 1 common function). In the case of rheumatoid arthritis and breast cancer the overlap is not significant. It is important to remark that 101 of the 119 breast cancer seeds are from Uniprot and may not be as much reliable. We repeated the same analysis selecting only 28 seeds from OMIM and DisGeNET, and in general the results of the overlap are not significant neither for asthma nor rheumatoid arthritis. The results can be explored with more detail in Supplementary Table S6.

Supplementary Table S6. Results of the genetic and functional overlap between the subnetwork of genes associated to asthma and rheumatoid arthritis, asthma and breast cancer, and rheumatoid arthritis and breast cancer. Breast cancer has been calculated either using DisGeNET and OMIM seeds (D+O) or using all the seeds (all). P-values have been corrected using the Benjamini-Hochberg correction for multiple tests. Results with non-significant P-value are highlighted in red.

	Genetic overlap						**Functional overlap**					
	Top		**Top without seeds**		**Seeds**		**Top**		**Top without seeds**		**Seeds**	
	Nº	**P-val.**	**Nº**	**P-val.**	**Nº**	**P-val.**	**Nº**	**P-value**	**Nº**	**P-value**	**Nº**	**P-value**
Asthma – Rheumatoid arthritis	55	2.90E-47	43	1.80E-64	12	7.00E-09	31	4.00E-45	18	5.50E-34	18	4.60E-22
Asthma – Breast cancer (all)	8	9.50E-03	5	1.30E-04	3	9.50E-02	1	2.10E-01	1	1.40E-02	0	1
Rheumatoid arthritis – Breast cancer (all)	4	7.20E-01	2	2.30E-01	2	5.20E-01	0	1	0	1	0	1
Asthma – Breast cancer (D+O)	2	2.30E-01	0	1	2	4.50E-02	0	1	0	1	0	1
Rheumatoid arthritis – Breast cancer (D+O)	0	1	0	1	0	1	0	1	0	1	1	1

9. Comparison of the case study results of asthma and rheumatoid arthritis with DIAMOnD

We have compared the functional enrichment of top-ranking genes identified by NetScore and DIAMOnD in asthma and rheumatoid arthritis. We based the analysis in *Sharma et al*., where the authors present a comparison of DIAMOnD and several other prioritisation algorithms, showing that DIAMOnD outperforms existing algorithms in predicting asthma related genes [31]. Following the procedure described in the original article, for each seed-gene, we determine the set of MSIgDB pathways [32] associated with the gene. For each pathway, we analyse its enrichment among the set of seed-genes using Fisher's exact test. The p-values are corrected using the Benjamini-Hochberg correction procedure for multiple tests. We choose the pathways with a significance level of p-value<0.01 as being associated with the set of seed-genes (enriched pathways). We calculate a Fisher's exact test between the top-ranking genes (based on functional-coherency) of the algorithm under analysis and the genes of the enriched pathways. The p-value of the Fisher's exact test gives the enrichment of the top-ranking genes. The enrichment of the pathways using top-ranking genes from NetScore and DIAMOnD as well as using the original set of seed-genes are shown in Supplementary Figure S2. When measuring the overlap between the two diseases, NetScore outperforms DIAMOnD, finding more genes involved in the pathways enriched in both diseases.



Supplementary Figure S2. Bar plot showing the enrichment of the top-ranking genes in terms of -log p-value.

10. Screening diseases to identify potential new indications of known drugs

GUILDify v2.0 introduces a “Drug Repurposing” functionality that can be accessed from the home page of the web server. This functionality takes a job ID as input, i.e. results for a drug (or a disease) and screens across a set of pre-calculated diseases (or drugs) for the significance of the overlap of genes and functions between the given job ID and the set of pre-calculated diseases (or drugs). The generation of the pre-calculated sets and the validation of the drug repurposing approach using these sets are explained below.

1. Set of pre-calculated diseases

We created a list of diseases using the UMLS concept unique identifiers from DisGeNET. Specifically, we obtained all diseases with gene associations reported by curated sources (UniProt, ORPHANET, PsyGeNET and HPO). We did not include CTD because it provides several clinically ambiguous phenotypes such as “liver cirrhosis, experimental”. From this list, we selected those diseases associated to at least 10 gene from DisGeNET, obtaining a final list of 757 diseases. We ran the prioritisation using the guildifyR package and the default parameters (BIANA network, and NetScore algorithm).

2. Set of pre-calculated drugs

The list of drugs was obtained by retrieving all drugs with targets stored in BIANA knowledge database. Out of this list, we selected those drugs with at least 10 targets (retrieved from at least one of the following databases in BIANA: DrugBank, ChEMBL, DGIdb, DrugCentral), obtaining a final list of 362 drugs. We run the prioritisation using the R package and default parameters (BIANA network, NetScore algorithm).

3. Set of drug-disease indications

We tested the quality of the predictions of indications of drugs using as benchmark the indications of Hetionet [33]. Accordingly, we used 161 drugs and 64 diseases that appeared in both Hetionet and the lists of pre-calculated drugs and diseases, producing a final set of 329 drug-disease pairs with known indications.

4. Finding the indication among the top-ranked results

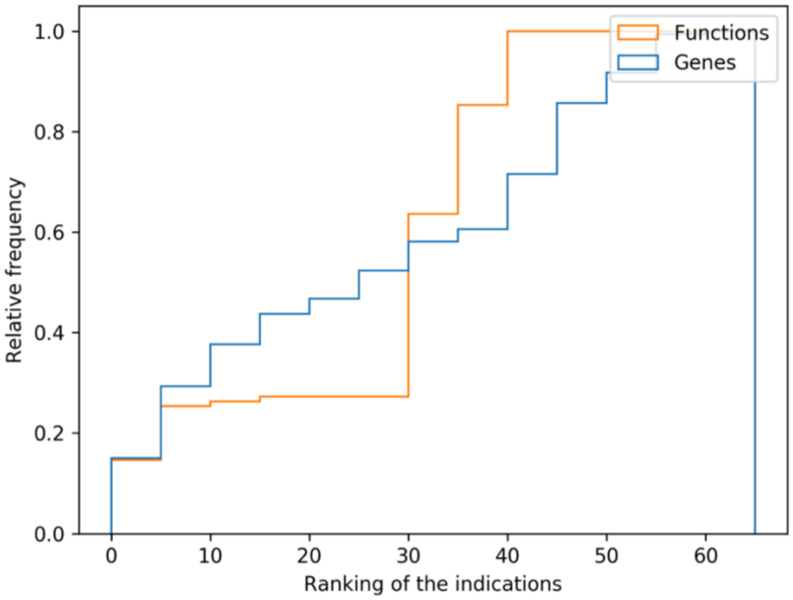
We plotted a histogram of the correct indications among the top ranked (see Supplementary Figure S3). This showed that 30% of the correct indications were already among the top 10 indications selected (and 50% of correct indications appeared among the top 25 predicted indications).

We calculated how many disease indications could be guessed depending on the number of top-ranked indications for a drug. We transformed this calculation in True Positive Rate (TPR) and False Positive Rate (FPR), calculated as:

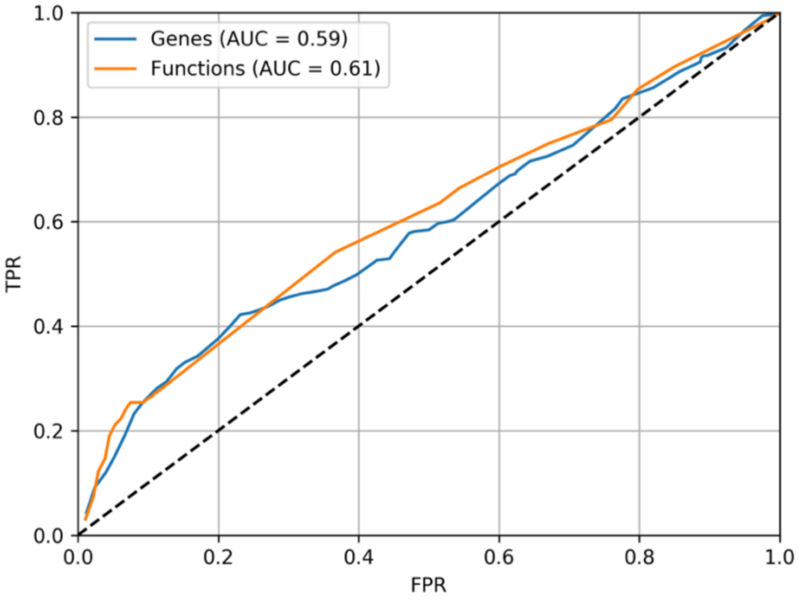
$$TPR = \frac{\#\ true\ positives}{\#\ positives} = \frac{\#\ guessed\ indications}{\#\ total\ indications}$$

$$FPR = \frac{\#\, false\ positives}{\#\, negatives} = \frac{\#\, non - indications\ in\ predictions}{\#\, non - indications}$$

Using the TPR and FPR we plotted the Receiver Operating Characteristic (ROC) curve (see Supplementary Figure S3). Using the top-scoring 1% of genes and functions, we obtained an Area Under the Curve (AUC) of 0.59 for genetic overlap and 0.61 for functional overlap.



Supplementary Figure S3. Cumulative distribution of the ranking positions achieved by the indications using the drug repurposing feature of GUILDify v2.0.



Supplementary Figure S4. Receiver Operating Characteristic (ROC) curve. The values of the Area Under the Curve (AUC) are indicated in the legend.

11. Visualisation of the top-ranking subnetwork

In the results page, we provide a visualisation panel to inspect in detail the interactions between the top-ranking nodes. Initially, we provide the user with an image of the subnetwork created with Matplotlib Python library [34]. The users have the option to click to “activate interactive visualization”, where the top-ranking nodes and interactions are displayed in a visualisation panel using the JavaScript-based network visualisation library, Cytoscape.js [35]. In addition to seeds (green hexagons), top-ranking proteins (yellow circles) and drugs (blue diamonds), the subnetwork includes the proteins that connect the seeds to the largest connected component induced by seeds (named “linkers” and shown as grey circles). The procedure to get the linkers is the following:

- Calculate the connected components of the top-ranking nodes.
- Check the size of the connected components and get the largest connected component (LCC). If there are more than one, we get the component with the highest mean score among all the nodes of the component.
- We order the rest of the components by their size and we find the shortest paths between the LCC and the remaining components (starting from the component with the highest size) to connect them to the LCC. If there is more than one shortest path, we get the shortest path that contains the node with the maximum score. If they have the same maximum score, we get the shortest path with highest mean score between all the nodes.

References

[1] J. Garcia-Garcia, E. Guney, R. Aragues, J. Planas-Iglesias, B. Oliva, Biana: a software framework for compiling biological interactions and analyzing networks, BMC Bioinformatics. 11 (2010) 56. doi:10.1186/1471-2105-11-56.

[2] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N.H. Campbell, G. Chavali, C. Chen, N. Del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R.C. Lovering, B. Meldal, A.N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. Van Roey, G. Cesareni, H. Hermjakob, The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases, Nucleic Acids Res. 42 (2014) 358–363. doi:10.1093/nar/gkt1115.

[3] A. Chatr-Aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N.K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam, C. Stark, B.J. Breitkreutz, K. Dolinski, M. Tyers, The BioGRID interaction database: 2017 update, Nucleic Acids Res. 45 (2017) D369–D379. doi:10.1093/nar/gkw1102.

[4] L. Salwinski, The Database of Interacting Proteins: 2004 update, Nucleic Acids Res. 32 (2004) 449D–451. doi:10.1093/nar/gkh086.

[5] G. Alanis-Lobato, M.A. Andrade-Navarro, M.H. Schaefer, HIPPIE v2.0: Enhancing meaningfulness and reliability of protein-protein interaction networks, Nucleic Acids Res. 45 (2017) D408–D414.

doi:10.1093/nar/gkw985.

[6] T. Li, R. Wernersson, R.B. Hansen, H. Horn, J. Mercer, G. Slodkowicz, C.T. Workman, O. Rigina, K. Rapacki, H.H. Stærfeldt, S. Brunak, T.S. Jensen, K. Lage, A scored human protein-protein interaction network to catalyze genomic interpretation, Nat. Methods. 14 (2016) 61–64. doi:10.1038/nmeth.4083.

[7] R. Herwig, C. Hardt, M. Lienhard, A. Kamburov, Analyzing and interpreting genome data at the network level with ConsensusPathDB, Nat. Protoc. 11 (2016) 1889–1907. doi:10.1038/nprot.2016.117.

[8] K.R. Brown, I. Jurisica, Unequal evolutionary conservation of human protein interactions in interologous networks, Genome Biol. 8 (2007) R95. doi:10.1186/gb-2007-8-5-r95.

[9] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N.T. Doncheva, A. Roth, P. Bork, L.J. Jensen, C. Von Mering, The STRING database in 2017 : quality-controlled protein – protein association networks , made broadly accessible, Nucleic Acids Res. 45 (2017) 362–368. doi:10.1093/nar/gkw937.

[10] G. Consortium, Genetic effects on gene expression across human tissues, Nature. 550 (2017) 204–213. doi:10.1038/nature24277.

[11] O. Basha, R. Barshir, M. Sharon, E. Lerman, B.F. Kirson, I. Hekselman, E. Yeger-Lotem, The TissueNet v.2 database: A quantitative view of protein-protein interactions across human tissues, Nucleic Acids Res. 45 (2017) D427–D431. doi:10.1093/nar/gkw1088.

[12] The UniProt Consortium, UniProt: The universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169. doi:10.1093/nar/gkw1099.

[13] A.P. Davis, C.J. Grondin, R.J. Johnson, D. Sciaky, B.L. King, R. McMorran, J. Wiegers, T.C. Wiegers, C.J. Mattingly, The Comparative Toxicogenomics Database: Update 2017, Nucleic Acids Res. 45 (2017) D972–D978. doi:10.1093/nar/gkw838.

[14] A. Rath, A. Olry, F. Dhombres, M.M. Brandt, B. Urbero, S. Ayme, Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users, Hum. Mutat. 33 (2012) 803–808. doi:10.1002/humu.22078.

[15] A. Gutiérrez-Sacristán, À. Bravo, M. Portero, O. Valverde, A. Armario, M.C. Blanco-Gandía, A. Farré, L. Fernández-Ibarrondo, F. Fonseca, J. Giraldo, A. Leis, A. Mané, M.A. Mayer, S. Montagud-Romero, R. Nadal, J. Ortiz, F.J. Pavón, E. Perez, M. Rodríguez-Arias, A. Serrano, M. Torrens, V. Warnault, F. Sanz, L.I. Furlong, Text mining and expert curation to develop a database on psychiatric diseases and their genes, Database. 1650 (2017) 48–55. doi:10.1093/database/bax043.

[16] S. Köhler, N.A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S.M. Bello, C.F. Boerkoel, K.M. Boycott, M. Brudno, O.J. Buske, P.F. Chinnery, V. Cipriani, L.E. Connell, H.J.S. Dawkins, L.E. DeMare, A.D. Devereau, B.B.A. De Vries, H. V. Firth, K. Freson, D. Greene, A. Hamosh, I. Helbig, C. Hum, J.A. Jähn, R. James, R. Krause, S.J.F. Laulederkind, H. Lochmüller, G.J. Lyon, S. Ogishima, A. Olry, W.H. Ouwehand, N. Pontikos, A. Rath, F. Schaefer, R.H. Scott, M. Segal, P.I. Sergouniotis, R. Sever, C.L. Smith, V. Straub, R. Thompson, C. Turner, E. Turro, M.W.M. Veltman, T. Vulliamy, J. Yu, J. Von Ziegenweidt, A. Zankl, S. Züchner, T. Zemojtel, J.O.B. Jacobsen, T. Groza, D. Smedley, C.J. Mungall, M. Haendel, P.N. Robinson, The human phenotype ontology in 2017, Nucleic Acids Res. 45 (2017) D865–D876. doi:10.1093/nar/gkw1039.

[17] J.S. Amberger, C.A. Bocchini, F. Schiettecatte, A.F. Scott, A. Hamosh, OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders, Nucleic Acids Res. 43 (2015) D789–D798. doi:10.1093/nar/gku1205.

[18] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: Tool for the unification of biology, Nat. Genet. 25 (2000) 25–29. doi:10.1038/75556.

[19] The Gene Ontology Consortium, Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium, Nucleic Acids Res. 45 (2017) D331–D338. doi:10.1093/nar/gkw1108.

[20] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. MacIejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, Di. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: A major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (2018) D1074–D1082. doi:10.1093/nar/gkx1037.

[21] K.C. Cotto, A.H. Wagner, Y. Feng, S. Kiwala, C. Coffman, G. Spies, A. Wollam, N.C. Spies, O.L. Griffith, M. Griffith, DGIdb 3.0 : a redesign and expansion of the drug-gene interaction database, Nucleic Acids Res. 46 (2018) 1068–1073. doi:10.1093/nar/gkx1143.

[22] O. Ursu, J. Holmes, J. Knockel, C.G. Bologa, J.J. Yang, S.L. Mathias, S.J. Nelson, T.I. Oprea, DrugCentral : online drug compendium, Nucleic Acids Res. 45 (2017) 932–939. doi:10.1093/nar/gkw993.

[23] A. Gaulton, A. Hersey, A. Patr, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L.J. Bellis, E. Cibri, M. Davies, N. Dedman, A. Karlsson, P. Magari, J.P. Overington, G. Papadatos, I. Smit, The ChEMBL database in 2017, Nucleic Acids Res. 45 (2017) 945–954. doi:10.1093/nar/gkw1074.

[24] J. Piñero, A. Gonzalez-Perez, E. Guney, J. Aguirre-Plans, F. Sanz, B. Oliva, L.I. Furlong, Network, Transcriptomic and Genomic Features Differentiate Genes Relevant for Drug Response, Front. Genet. 9 (2018) 412. doi:10.3389/fgene.2018.00412.

[25] E. Guney, B. Oliva, Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization, PLoS One. 7 (2012) e43557. doi:10.1371/journal.pone.0043557.

[26] E. Guney, J. García-garcía, B. Oliva, GUILDify : A web server for phenotypic characterization of genes through biological data integration and network-based prioritization algorithms, Bioinformatics. 30 (2014) 1789–1790. doi:10.1093/bioinformatics/btu092.

[27] S.D. Ghiassian, J. Menche, A.L. Barabási, A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome, PLoS Comput. Biol. 11 (2015) e1004120. doi:10.1371/journal.pcbi.1004120.

[28] C. Rubio-Perez, E. Guney, D. Aguilar, J. Piñero, J. Garcia-Garcia, B. Iadarola, F. Sanz, N. Fernandez-Fuentes, L.I. Furlong, B. Oliva, Genetic and functional characterization of disease associations explains comorbidity, Sci. Rep. 7 (2017) 6207. doi:10.1038/s41598-017-04939-4.

[29] T. Rolland, M. Taşan, B. Charloteaux, S.J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, A. Kamburov, S.D. Ghiassian, X. Yang, L. Ghamsari, D. Balcha, B.E. Begg, P. Braun, M. Brehme, M.P. Broly, A.R. Carvunis, D. Convery-Zupan, R. Corominas, J. Coulombe-Huntington, E. Dann, M. Dreze, A. Dricot, C. Fan, E. Franzosa, F. Gebreab, B.J. Gutierrez, M.F. Hardy, M. Jin, S. Kang, R. Kiros, G.N. Lin, K. Luck, A. Macwilliams, J. Menche, R.R. Murray, A. Palagi, M.M. Poulin, X. Rambout, J. Rasla, P. Reichert, V. Romero, E. Ruyssinck, J.M. Sahalie, A. Scholz, A.A. Shah, A. Sharma, Y. Shen, K. Spirohn, S. Tam, A.O. Tejeda, S.A. Trigg, J.C. Twizere, K. Vega, J. Walsh, M.E. Cusick, Y. Xia, A.L. Barabási, L.M. Iakoucheva, P. Aloy, J. De Las Rivas, J. Tavernier, M.A. Calderwood, D.E. Hill, T. Hao, F.P. Roth, M. Vidal, A proteome-scale map of the human interactome network, Cell. 159 (2014) 1212–1226. doi:10.1016/j.cell.2014.10.050.

[30] K. Luck, G.M. Sheynkman, I. Zhang, M. Vidal, Proteome-Scale Human Interactomics., Trends Biochem. Sci.

42 (2017) 342–354. doi:10.1016/j.tibs.2017.02.006.

[31] A. Sharma, J. Menche, C. Chris Huang, T. Ort, X. Zhou, M. Kitsak, N. Sahni, D. Thibault, L. Voung, F. Guo, S.D. Ghiassian, N. Gulbahce, F. Baribaud, J. Tocker, R. Dobrin, E. Barnathan, H. Liu, R.A. Panettieri, K.G. Tantisira, W. Qiu, B.A. Raby, E.K. Silverman, M. Vidal, S.T. Weiss, A.L. Barabási, A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma, Hum. Mol. Genet. 24 (2014) 3005–3020. doi:10.1093/hmg/ddv001.

[32] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, J.P. Mesirov, Molecular signatures database (MSigDB) 3.0, Bioinformatics. 27 (2011) 1739–1740. doi:10.1093/bioinformatics/btr260.

[33] D.S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S.L. Chen, D. Hadley, A. Green, P. Khankhanian, S.E. Baranzini, Systematic integration of biomedical knowledge prioritizes drugs for repurposing, Elife. 6 (2017) e26726. doi:10.7554/eLife.26726.

[34] J.D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. 9 (2007) 90–95. doi:10.1109/MCSE.2007.55.

[35] M. Franz, C.T. Lopes, G. Huck, Y. Dong, O. Sumer, G.D. Bader, Cytoscape.js: A graph theory library for visualisation and analysis, Bioinformatics. 32 (2015) 309–311. doi:10.1093/bioinformatics/btv557.