Prioritization of new candidate genes for rare genetic diseases by a disease-aware evaluation of heterogeneous molecular networks

Lorena de la Fuente[1,2,3], Marta Del Pozo-Valero[1,2], Irene Perea-Romero[1,2], Fiona Blanco-Kelly[1,2], Lidia Fernández-Caballero[1,2], Marta Cortón[1,2], Carmen Ayuso[1,2], Pablo Mínguez[1,2,3#]

[1]Department of Genetics, Health Research Institute–Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), 28049 Madrid, Spain

[2]Center for Biomedical Network Research on Rare Diseases (CIBERER), Instituto de Salud Carlos III (ISCIII), 28040 Madrid, Spain

[3]Bioinformatics Unit, Health Research Institute–Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), 28049 Madrid, Spain

[#]Corresponding author: Pablo Minguez (pablo.minguez@quironsalud.es)

ABSTRACT

Screening for pathogenic variants in the diagnosis of rare genetic diseases can now be performed on all genes thanks to the application of whole exome and genome sequencing (WES, WGS). Yet the repertoire of gene-disease associations is not complete. Several computer-based algorithms and databases integrate distinct gene-gene functional networks to accelerate the discovery of gene-disease associations. We hypothesize that the ability of every type of information to extract relevant insights is disease-dependent. We compiled 33 functional networks classified into 13 knowledge categories (KCs) and observed large variability in their ability to recover genes associated with 91 genetic diseases, as measured using efficiency and exclusivity. We developed GLOWgenes, a network-based algorithm that applies random walk with restart to evaluate KCs ability to recover genes from a given list associated with a phenotype, and modulates the prediction of new candidates accordingly. Comparison with other integration strategies and tools shows that our disease-aware approach can boost the discovery of new gene-disease associations, especially for the less obvious ones. KC contribution also varies if obtained using recently discovered genes. Applied to 15 unsolved WES, GLOWgenes proposed three new genes to be involved in the phenotypes of patients with syndromic inherited retinal dystrophies.

INTRODUCTION

A usual first step in Biomedicine is now to use omics to provide a first bunch of hypotheses on gene or protein associations with phenotypes that are afterward prioritized for further exploration. In this sense, rare diseases (RDs) are probably the paradigm in the application of bioinformatics tools for the prioritization of gene-disease associations. They are mainly Mendelian genetic diseases, and in their diagnosis it is now possible to screen the whole coding region or the entire genome using Whole Exome and Genome sequencing (WES and WGS) in search of causative variants. This extracts, for every patient, a large number of candidate variants beyond the genes already associated with the disease (1), including not only pathogenic mutations but also variants of uncertain significance and variants with conflicting interpretations. Their diagnosis needs to be boosted in order to increase the low ratio of solved cases (2). The challenge is to overcome the small cohort sizes and the low percentage of cases expected to be explained by unknown disease-causing genes.

Under this scenario, several computational disease-gene association prediction methods and databases have been developed to help in the prioritization of candidate genes. They can be divided into two general classes: 1) seed-based methods, which do not have *a priori* information about the interrogated trait but the functional landscape of a set of genes (seeds) provided by the user (3–7); and 2) predefined disease methods (8–10), which initiate the hunt from terms defining the disease. While seed-based algorithms are more flexible in the phenotype definition, predefined-disease methods take advantage of the knowledge accumulated for the pathologies. For both, the strategy for extracting new associations can also vary (11) from using text-mining techniques (DisGeNET (8), DISEASES (9), ENSEMBLE (12)), network-based methods (ToppNet (4), GUILDify (5), SNOW (13)), machine learning methods (ProDiGe (14), Phenolyzer (10)), or algorithms based on functional similarity (Endeavour (3), ToppFun (4)). They have in common the ability to screen large datasets to extract hidden associations. These sources can be of different nature, ranging from literature to different flavors of omics data, including their integration (5, 7) which has been reported to increase accuracy (15, 16). On top of this, resources such as PanelApp (17) provide manually curated candidate genes for genetic diseases based on community feedback.

A major challenge in the application of this type of resources to the diagnosis of RDs involves strengthening the discovery in both, diseases poorly studied, so lacking a solid background of knowledge, as well as in diseases where the missing genes have less obvious relationships to the known repertoire. With this goal in mind, we compiled several types of datasets with gene

and protein annotation and tested their ability to retrieve relevant gene-disease associations on many RDs using network biology. With the conclusions we developed a seed-based algorithm, called GLOWgenes, which is able to adapt its performance to every queried phenotype as well as to potentiate the extraction of less obvious associations. GLOWgenes has been benchmarked against current available tools and it has been implemented to work integrated into a variant calling pipeline in the diagnosis of RDs.

RESULTS

A compilation of heterogeneous gene-gene functional association networks

We wanted to build a diverse and rich gene functional information framework to be used for the prediction of new gene-disease associations. Thus, we compiled 33 publicly available sources with distinct functional information about human genes (Table S1). Of them, 22 are in a gene-gene network format (18–31) and 11 are gene annotations (32–39) that were transformed into co-annotation networks, with genes as nodes and pairwise relationships as edges (see Methods). When available, we coded the strength of the gene-gene associations using weighted edges (Table S1). We classified the 33 networks into 13 knowledge categories (KCs) covering different aspects of cell regulation and knowledge generation (Figure 1A, Table S1). The KCs include: i) gene cocitation in literature, ii) coessentiality as genetic interactions, iii) coexpression, iv) colocalization in cellular organelles, v) protein complexes, vi) targets drug sharing, vii) shared gene functional annotation, viii) features from genomic localization throughout evolution, ix) shared phenotypic annotation from mouse models, x) participation in molecular pathways, xi) shared human gene phenotypes, xii) protein-protein physical interactions (PPIs), and xiii) regulation of gene expression. Regarding the number of sources in each KC, coessentiality with five datasets is the KC with more contributing sources, and STRING (18), the database that contributes with more datasets spread over five KCs, in two of them alone (cocitations and genomic features). The networks have different sizes and shapes as shown by the number of nodes, edges and clustering coefficient (Figure 1B). The regulatory database RegNetwork (30) has the highest number of nodes, while data provided by CRISPR screenings the largest number of associations. Cocitation network from STRING stands out with the second largest number of nodes and edges, having a low clustering coefficient which suggests a hub-based connectivity. A grouping based on the overlap coefficients recovers the functional associations of the networks fitting quite well our classification into KCs (Figure S1).

KCs show different capabilities for the recovery of gene-phenotype associations

In a guilt-by-association approach, functional networks are a useful tool to associate genes with new functions or phenotypes based on their neighborhood (40). We aimed to assess the capability of the different KCs to recover the information from gene-disease associations. To perform this assessment, we selected 91 gene sets (or panels) used in the diagnosis of genetic diseases, classified into 20 disease families, from the PanelApp resource (17) (see Methods, Table S2) and applied a random walk with restart (RWWR) model (41) to every network using a training subset with 70% of the genes. The remaining genes of the gene sets (30%) were used to validate the prediction capability of each network by calculating two parameters: 1) efficiency (recall) reflecting the true associations caught between genes and phenotype, and 2) exclusivity, or capacity to recover genes others cannot, measured as the mean gene specificity (42). Both parameters are represented as the average of a 20-fold performance. For every KC we select a single individual network as the best representative based on its area under the precision-recall-gain (AUPRG) (43) (see Methods). This selection varies for every disease, particularly in the regulation networks (Figure S2).

To compare KCs, we calculated the efficiency and exclusivity of their best-performing networks at the n-top, taking n as the number of genes in the input disease panel. In a general view, KC efficiency varies substantially even within diseases of the same family (Figure 2A). In Figure 2A KCs are sorted by their overall performance. Generally, phenotype and cocitation KCs achieve the best efficiency, although all KCs show alternatively a high relevance in particular diseases (Figure S3). Some KCs work well in specific disease families: complexes in ciliopathies (Wilcoxon signed-rank test, WST, p-value = 6.08e−03), drug sharing in tumour syndromes (WST, p-value=6.01e-04) or coessentiality in metabolic disorders (WST, p-value = 2.31e−03) (Figure 2A, Figure S4). Remarkably, there is a high intrafamily disease variation with KCs outperforming the general trend for particular diseases (Figure S5). A focus on the efficiency for four distinct diseases (pilot diseases) accentuates the differential importance of specific KCs (Figure 2B).

We also calculated the exclusivity of KCs recovering gene-disease associations at n-top to measure their capacity to detect genes uniquely. The scenario here is much more diverse than the observed with the efficiency, there are no KCs that concentrate the exclusivity, but all are important in almost every disease family (Figure 2C, Figure S6). Interestingly, some KCs with low and medium overall recall, such as regulation or mouse phenotypes (Figure 2A), increase their importance in exclusivity having third and fourth best general performance respectively

(Figure 2C). On the other hand, in terms of exclusivity other KCs are globally downgraded compared to efficiency performance, as is the case of PPIs, falling from 4th to 11th place (Figure 2C).

Figure 2D displays a close look at the efficiency and exclusivity of KCs in the recovery of genes associated with our four pilot diseases. We observe no clear correlation between efficiency and exclusivity, but instead we found several KCs with very low efficiency but high exclusivity.

The history of acquired knowledge on a disease influences KCs contribution to the recovery of disease-associated genes

Since the discovery rate of gene-disease associations can vary over the years depending on several factors (disease prevalence, genetic heterogeneity, or scientific/monetary efforts), we hypothesize that the accumulated knowledge on each disease at the time of analysis may be also an aspect to be taken into account in the gene discovery process. Thus, we tested whether the performance of KCs in recovering genes for a particular disease changes if time is also considered. Thus, using 246 gene sets describing diseases and phenotypes from DisGeNET (44) (see Methods), we compared KCs performance in recovering genes using: 1) 30% of the genes chosen randomly in a 20-fold cross-validation, and 2) genes recently discovered (time-aware validation) (Figure 3A). Comparisons were made using an integration score that considers efficiency and exclusivity (see Methods).

Using the time-aware approach we observe a strong decrease in the performance of the KC cocitation (WST, p-value = 3.21 e-25) (Figure 3B). The same trend is reported for the KCs functional annotation and drug sharing (WST, p-values = 5.22 e-05 and 2.71 e-09 respectively). On the other hand, other KCs such as regulation and coessentiality increase their capacity to detect recent gene-disease associations (WST, p-values = 4.42 e-23 and 1.53 e-14 respectively). To illustrate these results, we selected the four DisGeNET gene sets with the most dramatic changes in KCs performance (Figure 3C and Methods). As shown above, cocitation displays a severe decrease in its performance for all four diseases using the time-aware validation. For the rest of the KCs, the patterns are highly variable. Thus, genes recovered for attention deficit hyperactivity disorder seem to be caught mainly by KCs coexpression, coessentiality, colocalization and drug sharing. In coronary heart disease, with a validation cut-off at year 2012, cocitation continues to play a predominant role followed by protein complexes and colocalization. In prostate carcinoma, coessentiality, regulation and canonical pathways information overtake cocitation whose influence using validation on random

genes was 50% of the total. The case of toxic hepatitis is especially striking, with the new pattern conserving the same trend but with a significant increase of regulation as KC with more potential to detect late insights (Figure S7).

A disease-aware algorithm to integrate several sources of knowledge

As the contribution of KCs in the recovery of causing genes is disease-specific, we developed a novel algorithm, named GLOWgenes, which adapts its performance to the special capabilities of every KC to recover genes associated with specific phenotypes. Thus, given a gene set associated to a particular disease (GAD), or any trait of interest, GLOWgenes applies two steps in parallel (Figure 4): 1) evaluates performance of each network (N=33) in recovering genes from the gene set (step repeated 20 times), where the algorithm performs a RWWR for all networks using 70% of the genes as seeds, calculates KC weights based on their efficiency and exclusivity using the rest of the genes from the gene set (test genes), and chooses the network with best peformance of each KC. If provided, seeds and test genes can also be defined using a cut-off defined by the year of publication of the gene-disease associations; and 2) for every KC (using the best network from each selected in step 1), apply RWWR on the whole set of GADs to rank all genes. Finally, GLOWgenes uses KC weights (step 1) to modulate the gene rankings (step 2) and produces a single gene ranking in which genes in the GAD have a value of zero and the rest are sorted in descending order with value 1 as the most associated.

We compared our disease-aware integration strategy with two widely used alternative approaches for disease candidate gene prediction (Figure 5A): 1) integration of datasets into a composite network prior to gene prioritization (15, 45, 46); and 2) combination of multiple rank lists using order statistics (3, 47, 48), represented by two methods, data fusion (47) and robust rank aggregation (3, 48). We applied these three methods, and ours, to disease gene sets from PanelApp. To evaluate the predictive capability in a gene-discovery scenario, we used their collection of candidate genes, called red genes (RGs), as validation sets. A total of 63 PanelApp gene sets, with more than 10 RGs, were used. Recall was calculated at various n-top genes. Using the same input information for all four algorithms (see Methods), the results show that our disease-aware integration approach achieves the best average performance across all disease gene sets, with a mean AUPRG=0.92 (Figure 5B) outperforming the rest data integration approaches together in 52% of the diseases looking only at the first 16 predicted genes (n-top=16), and in 80% of the diseases considering the first 256 genes (n-top=256) of the rankings (Figure 5C). The composite network performed second in the global ranking with 10 and 32

diseases with the best and second-best recalls respectively, considering the same number of genes as the validated set (Figure S8).

Comparative assessment of disease candidate gene prioritization methods

To extend the evaluation of GLOWgenes beyond the integration method, we compared its performance against other tools that predict disease-gene associations using a list of genes as seeds and do not need a preselection of candidates, called here seed-association (SA) methods. In addition, we introduced as external references, methods that predict associations using disease terms instead of gene lists , called here predefined disease (PD) methods (9, 44, 49). In total, we selected six SA methods, including: DIAMOnD (6), Guildifyv2.0 (5), ToppGenet-net (4, 50), ToppGenet-fun (4, 51), Endeavour (3) and GeneMANIA (7), and three PD methods, including: DISEASES database (9), BEFREE (44, 49) (as extracted from DisGeNET) and predictions from DisGeNET (see Methods, Table S3-4). First, we tested the ability to recover genes with well-supported evidence in their association to the disease, represented by the PanelApp Green-Ambar validation genes (GAG) in our four pilot diseases. Figure 6A shows that PD methods overall outperform SA methods in all four diseases with better recall levels at different number of top genes selected. Only GLOWgenes has similar or better recall (e.g. in cardiomyopathies) than DisGeNET, BEFREE and DISEASE predictions, and performs widely better than its counterpart methods with only ToppGene (4) reaching similar recall levels in severe microcephaly (Figure 6A).

Next, we carried out a similar benchmark using PanelApp RGs of the same diseases as a proxy to test the ability to catch new gene-disease associations beyond the current knowledge. We observe a general decrease in the recall obtained by all methods, with SA and PD methods performing similarly (Figure 6B). In this scenario, GLOWgenes outperforms all other methods in three out of four evaluated diseases and gets similar recall levels for hearing loss. After GLOWgenes, BEFREE, DisGeNET, GeneMANIA, ToppGene and Endeavour ranked in alternate positions. GLOWgenes is also the method that captures more unique genes (Figure S9). Among all SA tools considered, only three of them provide programmatic access options for automatic task execution (NetCombo from GUILDIFY2.0, DIAMOnD and GeneMANIA). We compared their performance using RG validation sets for 63 PanelApp disease gene sets with GLOWgenes obtaining the best recall across all diseases globally (Figure S10).

Gene discovery in undiagnosed cases with syndromic inherited retinal dystrophy using GLOWgenes

A major application of gene-phenotype discovery methods is to propose candidate genes to be implicated in RDs. GLOWgenes is adapted to annotate a list of variants provided in a table format as it is coupled to our reanalysis pipeline (52). Here we have applied GLOWgenes for the discovery of candidate genes involved in syndromic forms of inherited retinal dystrophies (sIRD). As a proof of concept, 15 unsolved cases with sIRD and WES data available from the cohort of the Fundación Jiménez Díaz University Hospital (FJD-UH, Madrid, Spain) (2) were analyzed. These samples were previously analyzed at the FJD-UH with no variants found fitting the phenotype in known sIRD-associated genes (sIRD-virtual-panel with 198 genes used in the diagnostic protocol for sIRD, Table S5). After variant calling (52) and filtering by: i) quality (Q=100, DP=10), ii) predicted pathogenicity (CADD>15), and iii) allele frequency (gnomADg_AF_POPMAX<0.05/NA), we obtained an average of 1562 variants per sample in around 1000 genes without previous association to sIRDs. GLOWgenes was run using the sIRD-virtual-panel as seeds to provide a gene ranking that was used to prioritize the filtered variants of each case.

Thus, by interrogating the top 300 predicted gene list (Table 1), we yielded a total of three candidate genes with pathogenic variants that fitted the phenotype in two patients with sIRD. We found two heterozygous variants in Case 1 (Table 1). First, a pathogenic missense variant in the *SHH* gene previously reported to cause holoprosencephaly type 3 (MIM:142945), microphthalmia with coloboma type 5 (MIM:611638), schizencephaly (MIM:269160) or single median maxillary central incisor (MIM:147250) and with no clear mode of inheritance. And secondly, a heterozygous pathogenic stop-gain variant in the *DNAH5* gene, a dynein involved in primary ciliary dyskinesia type 3, with or without situs inversus (MIM:608644) and reported as autosomal recessive.

In a second case (Case 2, Table 1), we found a heterozygous likely pathogenic variant in the *GLI1* gene that has been associated with two types of polydactyly (phenotype present in our patient) and reported as autosomal recessive (source: OMIM database).

Regarding the contribution of KCs in the prioritization of the three genes proposed, reporting only those reaching >50% of the total score (Figure S11), *SHH* was found mainly by the signal provided by co-citation (46%) and mouse models (22%) networks, *DNAH1* by the co-citation (33%), co-expression (29%) and mouse models (19%) networks, and *GLI1* by the co-citation (26%), PPIs (19%) and regulation (10%) networks.

DISCUSSION

Around 50% of patients with RDs do not have an accurate diagnosis (53) and there is an urgent need to reduce this gap (54). One limitation is the lack of certainty about the clinical significance of the genomic variation data (1), with 40% of the variants annotated as having uncertain significance in the ClinVar database (54). If all genes were suspicious, a regular WES analysis may detect around 20,000 single nucleotide variants (55), 200 of them very rare (MAF<0.1%) and an average of 30 not detected in any other individual (56). All these numbers describe a scenario in which the use of WES and WGS to diagnose RD patients with causative variants outside the current knowledge needs the application of prioritization algorithms. Current major national and international initiatives focus on the use of these two sequencing techniques (WES and WGS) to assume discovery (56).

Methods for the prediction of new gene-disease associations evolved from those developed for predicting gene and protein function (57, 58) with social network (59) and data integration at the foundations (60). The needs are now focused on the development of specific and accurate resources that overcome the limitations of every area of application. Here, RDs represent a rather particular challenge. They are over 7000 diseases with different levels of genetic and phenotype heterogeneity. They are mainly Mendelian, so if the variant is in the coding DNA, we search for a single gene. Moreover, they generally have small cohorts available so a clear genotype-phenotype correlation can be difficult to find. Thus, RD diagnosis can become a “needle in a haystack” task where the solution is still far connected to current knowledge. In consequence, to increase their efficiency in difficult cases, gene-disease prediction methods are encouraged to use new and various types of functional annotations and combine them accurately (15, 16, 61, 62). In this study, we first confirmed that different types of gene-gene functional association networks had different capabilities in recovering known genes associated with a large collection of RDs. In terms of efficiency, the co-annotation network using phenotypes, as well as the cocitation network using STRING text-mining data, are the best performing sources in general, with particular diseases having alternative important sources. Differences are especially evident when considering genes detected uniquely by one source, which we called exclusivity (Figure 2). This fact might be of essential importance considering its application to Mendelian diseases.

With needs, limitations, and preliminary results in hand, we developed GLOWgenes (www.glowgenes.org), a seeds-based algorithm to prioritize candidate genes with main

application in the diagnosis of RDs using WES and WGS. GLOWgenes applies the guilt-by-association principle using the RWWR method on multiple networks as different layers of functional information (62–64), preserving not only the nature of the associations but also the topology of the networks. GLOWgenes is unique in the sense of evaluating the performance of different sources in every case to assign weights that are afterward integrated. We tested our integration method using well-known representatives of this kind of approaches with good results (Figure 3).

A second benchmark was performed comparing GLOWgenes with state-of-the-art tools to predict gene-disease associations based on seeds. Here, popular PD tools were used as external references. Using well-known gene-disease associations as testers (GAGs), PD methods achieved more prediction power than SA algorithms, with GLOWgenes behaving better than SA tools and similarly to PD tools (Figure 6A). When considering a more realistic discovery scenario, that is, testing the ability to recover candidate genes (RGs), the difference in the performance of SA and PD methods is less evident, with GLOWgenes also at the top performers (Figure 6B). The main difference between PD and SA tools is that PD methods cannot be easily adapted to new, or a combination of, phenotypes as required in genetic syndromes for instance (65). In fact, most PD algorithms are based on text-mining approaches which reward with a good recall but limit their ability to go beyond current knowledge. In this work, we provide two results that suggest that text-mining may mask interesting RD candidate genes. First, we observed a dramatic reduction in the importance of text-mining when trying to highlight recently discovered genes, using our time-aware partitioning strategy compared to recovering randomly selected genes (Figure 3). Second, if efficiency is measured in RGs, the methods based on text-mining, mainly PD methods, reduce their performance (Figure 6B).

Another interesting finding is that, in most cases, SA methods aggregating several functional sources (GLOWgenes, GeneMANIA, ToppGene, and Endeveour) performed better than network-based approaches using only PPI information (Guildifyv2.0, DIAMonD, ToppNet) (Table S4). In general, we observe that GLOWgenes behaves well under any scenario tested, being also the method catching more genes uniquely. Caveats specific to these type of approaches include the low overall recall of the algorithms and the accuracy, variety and level of coverage of the available functional gene and protein associations. In their application to the diagnosis of RDs it must be taken into account that the causal variants may not be in the coding DNA.

The initial implementation of GLOWgenes that we present here, is designed to be coupled to a pipeline of variant calling and annotation of DNASeq data. GLOWgenes is now serving

candidate genes in the research of RDs in the FJD-UH annotating variants using our reanalysis FJD-pipeline (52). Although its first motivation is to contribute to the genetic diagnosis of RDs, GLOWgenes is a predictor of new gene-disease associations that can help in other discovery scenarios. For instance, it has also been used in the search for genomic variants that provide susceptibility to COVID-19 (66). As a proof of concept, we provide here the results of the analysis of 15 WES of patients with sIRD. Three genes are reported here to provide new insights in two sIRD cases. In the first case, we found two pathogenic variants in heterozygosis involving two genes: *SHH* and *DNAH5*. Regarding the *SHH* gene, mutations produce a related phenotype, and it has a role in the neurogenesis of the zebrafish retina (52). Although the proband's father carries the same variant, *SHH* mutations are described to have a high variability of expression (67). On the other hand, *DNAH5* has been proposed as a candidate gene for retinal dystrophies using a different method based on the aggregation of variants in unsolved cases compared to controls (1). In a second case, we found a monoallelic variant in the recessive gene *GLI1*, which explains partially the patient's phenotype (68). Further analyses are needed to reach a conclusive diagnostic in both cases.

In conclusion, this work describes the special needs and limitations in the prioritization of candidate genes in the diagnosis of RDs using WES and WGS and presents a method that is hopefully useful to reduce the diagnostic gap of this type of patients.

MATERIALS AND METHODS

Compilation of gene-gene functional associations from public sources

We gathered functional gene or protein information from various sources (18–27, 29–39, 69) to build 33 networks of functional gene-gene interactions (Table S1). The sources with protein information had no isoform resolution, so a one-to-one correspondence with the genes can be assumed. Networks were classified into 13 classes, named here as knowledge categories (KCs) according to the nature of the functional information that contains (Table S1). Gene identifiers from sources were all mapped into HGNC gene symbols, excluding interactions involving genes without HGNC mapping. Weighted edges were kept and normalized. If unavailable, edge weights were all set up to 1 (Table S1).

Gene annotations not represented as gene or protein relationships were transformed into gene-gene co-annotation networks, describing genes (nodes) and their functional relationship (edges). Thus, for networks describing human gene-phenotype similarity using Human

Phenotype Ontology (HPO) terms (32) and phenotype similarity using mouse orthologs from the Mouse Genome Informatics (MGI) (16) we calculated Jaccard similarity for each pair of genes sharing at least one HPO term and constructed a null distribution of Jaccard values to compute z-scores. Significant gene interactions (z-score>1.96; p-values<0.05) were selected to generate the network of phenotypes. Z-score values were used as edge weights. Direct linkage between genes sharing annotations was used for network construction for the drug-gene interaction database (DGIdb) (34) and complexes from CORUM (36). The co-expression network from COXPRESdb (19) was created considering gene pairs with Mutual Rank (MR) co-expression correlations under 2000. The inverse of the MR was used as edge weights. The ProteomeHD (24) co-expression network was filtered to contain only the top-scoring 0.5% pairwise gene co-regulations, as recommended. Coessentiality networks were derived from: i) inferred genetic interactions from CRISPR screens by Rauscher *et al*. (27), ii) weighted coessentiality gene interactions obtained by Kim *et al*. (70) using CRISPR essentiality screens from the Avana Project (31); iii) gene fitness rank correlation coefficients calculated using two different RNAi- and CRISPR-Cas9-based screening datasets and following Pan *et al*. (35) processing pipeline (rank threshold = 1024; inverse rank threshold as edge weights); and iv) genetic interactions from the BIOGRID database (25). Networks derived from GO annotation were constructed using semantic similarity from GOGO (37) for gene pairs sharing at least one GO term. Semantic similarity scores were used as edge weights. DoRoThea evidence levels were transformed into numeric values for regulatory network weighting normalized to 0-1 range.

Comparison of network topology and similarity measurement

Attributes from each compiled network (node size, number of edges, average clustering coefficient) were calculated using the python package NetworkX.

To compare the different networks, we computed edge-wise network similarity using the overlap coefficient (OC), also known as the Szymkiewicz–Simpson coefficient. OC is defined as the edge size of the union of two graphs (G1 and G2) over the size of the smaller set between G1 and G2.

Disease-associated gene sets used in the evaluation of KCs and benchmarks

Assessment of KCs importance in gene recovery was performed using disease-associated gene sets with high and moderate evidence (Green and Ambar genes, GAGs) from the Genomics

England PanelApp (17). Disease-associated genomic regions not fitted to genes such as copy number variations (CNVs) and short tandem repeats were filtered out from the panels. We grouped the panels into disease families using the level 2 provided in the disease classification by PanelApp. Panels lacking classification were manually curated by clinicians at FJD-UH. Although we observed that disease families might partially overlap in their definition, for instance with ciliopathies and ophthalmological disorders, we kept the PanelApp definition.

Alternative integration strategies were benchmarked using PanelApp red genes (RGs), defined as disease-promising genes that need further evidence for clinical disease diagnosis. Gene-disease prediction tools were benchmarked using both GAGs and RGs.

When PanelApp gene sets were used, super panels (involving a mix of phenotypes) were excluded. In the case of using GAGs, only panels with at least 40 GAGs were selected (91 panels). When RGs were used, only panels with at least 10 RGs were considered (70 panels).

Network signal propagation

Network propagation of signal (trait or disease annotated) provided by gene sets was performed using the random walk with restart (RWWR) model proposed by Kohler *et al*. (71) as implemented in python software NetworkX and modified to generate a degree weighted adjacency matrix for subsequent propagation in order to minimize node degree bias (avoid bias towards highly connected genes). The modification was done following the edge weight transformation proposed by Vanunu *et al*. (72) and used by others (73, 74). Restarting walk probability (rwp) was set up to 0.75. Convergence is decided when the probability difference between two consecutive time steps is less than 10e-6. Upon convergence, the frequency with which each node (gene) in the network has been visited is returned as a propagation score vector (S), which represents the probability of genes being associated with the input gene set. All networks were treated as weighted, if information was available, so adjacency matrices were defined providing edge weight attributes.

Systematic evaluation of evidence networks representing heterogeneous knowledge categories

To assess the performance of a network to prioritize phenotype-associated genes we performed a 20-fold random sub-sampling validation. Each phenotypic gene set was randomly split into 70%/30% training/test subsets. Training genes (seeds) were then propagated in each individual

network using the RWWR model and the resulting gene ranking was evaluated on test genes. The validation results were averaged over the 20 cross-validation rounds. Several metrics were calculated to assess the performance of a given network in a particular disease. Given the high imbalance nature of our dataset (high negative/positive instance ratio), the overall ranking was evaluated using the precision-recall analysis. We calculated the Precision-Recall-Gain (43) curve and its associated area (AUPRG). The network with the highest AUPRG in each KC is selected as the representative KC network and used for further analysis. To emphasize the top-ranked genes, instead of testing true/false separation in the total gene ranking, we also computed recall and gene specificity at different top-k thresholds. Recall evaluates the efficiency of networks to prioritize disease genes while gene specificity is informative of their degree of detection by particular KCs. Gene specificity was calculated for every true positive gene captured as in Martinez *et al*. 2008 (42). High specificity genes represent genes captured by a single KC while low specificity genes refer to genes detected by all KCs. The overall KC specificity level was calculated by averaging gene specificity scores and summarizes the ability of a KC to detect disease-specific genes, named here as exclusivity.

Time-aware network evaluation

The time-aware network evaluation analyses whether the contribution of KCs in gene recovery for a particular phenotype changes if time is also considered. As input, the publication year for each gene associated with the phenotype under study must be provided. GLOWgenes takes that time-sorted list of genes and evaluates each individual source of evidence (network) using a time-aware validation, which relies on gene publication year to split gene sets into 70/30 train/test partitions. The oldest genes are used as the training set to determine the ability of each network to recover the most recently discovered genes (test set). Partition publication time, metrics, and overall gene prioritization rank using time-aware mode are provided. Moreover, random vs time-aware network evaluation statistics are also computed in time-aware mode.

Gene-disease associations from DisGeNET (44) were used in the evaluation of the time-aware approach. The publication year was recovered from DisGeNET SQLite database and time-aware disease gene sets were generated based on the UMLS Concept ID (CUI) disease classification, comprising Phenotype and Disease categories. We only considered gene-disease associations from curated sources (CGI, CLINGEN, GENOMICS_ENGLAND, CTD_human, PSYGENET, ORPHANET, UNIPROT, GWASCAT, GWASDB, CLINVAR) and limited the

analysis to gene sets with at least 70 genes, remaining a total of 246 disease gene sets for time-aware disease candidate gene discovery.

GLOWgenes algorithm

GLOWgenes is based on the integration of gene prioritization results from *k* heterogeneous knowledge categories (KCs), called here disease-aware prioritization. Gene prioritization results from each KC representative network *i* are integrated into a unified ranking by incorporating *ad-hoc* information about their individual performance on the particular phenotype/disease under study.

The strategy involves three main steps: Step 1. RWWR propagation of *n* phenotype/disease-related genes on each individual network resulting in a gene ranking representing their association strength to the input gene set. Step 2. Networks evaluation and selection follows the steps listed in the previous section:

- For each network out of the 33 considered in this analysis, we performed a 20x random cross-validation of phenotype-associated input genes (70/30 training/test). For each partition:
    - The training subset is propagated using a RWWR model.
    - The area under the precision/recall gain curve is calculated.
- Networks with the highest mean AUPRG at each KC are selected as KC representative networks and used for further analysis.
- KC exclusivity and efficiency (recall) are calculated at different top-k for each representative KC network.
- Integration scores are calculated as the product of efficiency and exclusivity for each KC as a measure of the ability to recover input disease-associated genes.

Step 3. RWWR Scores (step 1) of representative KC networks (step 2) are integrated using the integration scores (step 2). In order to allow for unbiased integration of multiple networks, gene scores from network *i* are normalized by subtracting the mean of the scores of all nodes from the score of node n and then dividing by the standard deviation of the distribution. The normalized z-scores for representative KC networks are then merged in a k-n matrix (k = KCs; n=genes) and score imputation is applied for missing values. Disease-aware integration of the k

KC representative networks is performed by weighing the gene z-scores in each network using the performance factor obtained from the evaluation of *i* network and subsequent combination of ranks by averaging these weighted scores. A final overall gene ranking is generated. GLOWgenes is available at www.glowgenes.org.

Evaluation of alternative approaches for data integration in gene discovery

We compared the integration approach used by GLOWgenes (disease-aware integration) to two common strategies for network-based data integration in gene discovery: the fusion of rankings generated by individual modeling of datasets by order statistics (3, 47, 48) and the construction of a composite network for subsequent modeling (15, 45, 46). Regarding the first, we evaluated two rank aggregation algorithms based on order statistics: a) the optimization of the Stuart algorithm (75) by Aerts (47, 63) and implemented in Endeavour (3), and b) the Robust Rank Aggregation (RRA) algorithm (48). We ran the implementations of both algorithms using the R package RobustRankAggreg. For the network integration strategy, we constructed a composite network using our 33 defined networks by adapting the most optimal configuration of a composite network described by Huang *et al*. (15). Selection of edges supported by at least two sources from different KCs generated a composite network containing 4509215 edges and 19302 nodes. Alternative strategies were run across 63 PanelApp GAG gene sets and validated in RG genes. Recall-at-k was measured to quantify what fraction of all the disease genes are ranked within the first k predicted genes (k = 8, 16, 32, 64, 128, 256, 512). The best integration methods are selected at recall-at-n (validation set size). The mean area under the Precision-Recall curve (AUPRG) across gene sets was also calculated for each approach.

Benchmark of tools for the prioritization of new disease-related genes

For benchmarking, we selected tools that met four criteria: a) accessible, b) updated since 2016, c) predicting gene-disease associations from a user-input set of genes previously linked to a phenotype (seeds), d) no need for an implicit set of user-input candidate genes to operate. A total of six seed-association tools were selected and run on four pilot disease gene panels (described before) using their default parameters: Endeavour, which uses a whole-genome approach (3), ToppNet algorithm, which uses ToppGenet for selection of the first neighboring genes in PPI as candidates (ToppGenet - network based, ToppGenet-net) (4, 50), ToppGene algorithm, which uses also ToppGenet for candidate selection (ToppGenet - functional similarity, ToppGenet-fun) (4, 50), NetComb-GUILDIfy (5), the DIAMOnD (6) version implemented in

GUILDIFY2.0 web server, and GeneMANIA using its own API (76). In all cases, a ranked list of candidate genes was derived to evaluate their performance on GAGs and RGs validation sets.

We included as external references in the benchmark, pre-defined disease methods integrating gene-disease associations: DISEASES (9) and DisGeNET (44). In the case of DisGeNET, we separate predictions/inferences from curated sources in order to avoid knowledge bias during benchmarking. We performed a recalculation of gene-disease association scores from pre-computed DisGeNET data by removing the proportional score associated to curated evidence, generating our DisGeNET-noncurated dataset. Moreover, we evaluated just DisGeNET predictions by considering only associations derived from biomedical literature using BEFREE (DisGeNET-BEFREE). DISEASES, DisGeNET-noncurated and DisGeNET-BEFREE pre-computed datasets were interrogated for each evaluated PanelApp disease by selecting the gene-disease associations involving disease terms enriched in the corresponding PanelApp gene set (FDR Fisher-test <2e-16). To make the seed-association and pre-defined disease comparable, we also filtered out the computed rank list by removing genes contained in the PanelApp gene set under study (used as seeds for seed-associated methods).

The main characteristics and sources of evidence for both seed-association and pre-defined disease methods are listed in Tables S2 and S3. All methods were evaluated and compared across four selected diseases using both GAG and RG genes as validation sets. The selected diseases were cardiomyopathies including childhood onset, hearing loss, retinal disorders, and severe microcephaly. It was not feasible to use the full set of diseases, since most tools do not provide programmatic access.

Three SA methods (GeneMANIA, DIAMOnD and NetCombo-GUILDIfy) provide programmatic access, so they were run on 63 PanelApp disease gene sets using RGs as validation sets. Mean recall and error at different top-k were calculated for each approach. The mean area under the Precision-Recall curve (AUPRG) was computed for GLOWgenes and GeneMANIA, but not for DIAMOnD and NetCombo-GUILDIfy since they just returned the 500-top rank list of candidate genes.

Detection of variants fitting patient phenotype in GLOWgenes candidates using WES

We considered a total of 15 WES tests from patients with syndromic inherited retinal dystrophy (sIRD) phenotypes for clinical application of GLOWgenes. All cases have an inconclusive genetic diagnosis after clinical analysis of pathogenic variants at known sIRD-associated genes.

We used an in-house pipeline for the detection and annotation of germline variants (52), available at https://github.com/TBLabFJD/VariantCallingFJD. Variants were filtered by quality (Q=100, DP=10), predicted pathogenicity (CADD > 15) and low allele frequency (gnomADg_AF_POPMAX < 0.05 or NA). Selected pathogenic WES variants were prioritized using GLOWgenes sIRD ranking of candidate genes. To generate this sIRD ranking of candidate genes we ran GLOWgenes using as seeds the sIRD virtual gene panel used for diagnosis at FJD-UH, which contained 198 genes. Clinical diagnosis was re-assessed by analyzing GLOWgenes detected sIRD candidate genes that contained pathogenic variants in WES cases and considering information on the familiar phenotypic pattern and variant segregation, when available. Genes detected by GLOWgenes that overlapped sIRD phenotype were classified as phenotype-matched candidates and selected for further studies.

Relative contribution of each KC to the association to known sIRD genes

The relative contribution of each KC to the association with known sIRD genes was calculated for each WES-supported novel sIRD candidate and represented in pie charts using the GLOWgenes-normalized RWWR integration score for each KC representative network before KC integration by GLOWgenes.

Ethics Approval and Consent to Participate

The project was reviewed and approved by the Research Ethics Committee of FJD-UH (PIC086-19_FJD) and fulfills the principles of the Declaration of Helsinki and subsequent reviews. All patients signed an informed consent before their participation. All samples included in this work were pseudonymized.

Funding

This work was supported by Instituto de Salud Carlos III (ISCIII) of the Spanish Ministry of Health (PI18/00579, IMP/00019), the Comunidad de Madrid (CAM, RAREGenomics Project, B2017/BMD-3721), Centro de Investigación Biomédica en Red Enfermedades Raras (CIBERER, 06/07/0036) and Fundación Conchita Rábago. LdlF was supported by the platform technician contract of ISCIII (CA18/00017). PM is supported by a Miguel Servet program contract from ISCIII (CP16/00116, CPII21/00015).

Author contribution

Study concept and design: PM and LdlF. Software implementation: LdlF. Data analysis and interpretation: LdlF., MDP, IPR, FBK, MC, LFC, CA and PM. Drafting of the manuscript: LdlF and PM. Manuscript reviewing and editing: LdlF, IPR, MDP, FBK, MC, LFC, CA and PM. Resources: CA and PM. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

GLOWgenes code and data is available at https://github.com/TBLabFJD/GLOWgenes.

Tables and Figures

Table 1: Syndromic inherited retinal dystrophy candidate genes captured by GLOWgenes that hold pathogenic/likely pathogenic variants and fulfill phenotype in cases analyzed by WES. Variant pathogenicity is coded according to ACMG guidelines as: 1 (benign), 2 (likely benign), 3 (uncertain significance), 4 (likely pathogenic) and 5 (pathogenic).

Family	Gene	Variant	Consequence	CADD	Max POP Freq	Pathogenicity	Genotype	Inheritance
Case 1	SHH	NM_000193.2:c.676G>A	Missense variant	27,3	8,8E-05	5	0/1	AD
Case 1	DNAH5	NM_001369.2:c.13486C>T	Stop gained	59	1.34E-04	5	0/1	AR
Case 2	GLI1	NM_005269.2:c.762C>T	Splice region variant	15,05	6.67E-05	4	0/1	NA

Figures

Figure 1. Composition of knowledge categories (KCs) and their network properties. A) 13 KCs grouping 33 datasets. The pie chart indicates the number of source contributors per KC, ranging from 1 (drug sharing, co-citation and mouse models) to 5 (coessentiallity). **B)** Network attributes of the 33

networks generated from the compiled datasets. Here we present number of nodes, number of edges, the mean of their clustering coefficient and whether edges are weighted or not. Color code is assigned according to their KC.

Figure 2. Disease-dependent performance of heterogeneous knowledge categories (KC) to recover gene-disease associations. A) Heatmap representing the efficiency of every KC (x axis) in recovering genes in 91 disease gene sets from PanelApp (y axis). Efficiency is measured as the recall at top n (gene set size). Only genes with high and moderate evidence status at PanelApp (Green and Ambar) were considered. Diseases are classified into families. Disease families with only one gene set were discarded for plotting purposes (7 gene sets). Diseases within families are sorted according to their similarity in the efficiency pattern using hierarchical clustering. KCs were also sorted according to median recall levels across disease gene sets. **B)** Ranking of KC efficiency for four selected diseases (pilot diseases). **C)** Heatmap representing the exclusivity of every KC (x axis) in gene recovery in 91 disease gene sets from PanelApp (y axis). Exclusivity is calculated as the normalized value of the mean gene specificity of genes at top n (gene set size). For representation purposes, disease filtering and plot layout were done as in A. **D)** Scatter plot representing KC efficiency *versus* KC exclusivity for four selected diseases.

Figure 3. Time-aware evaluation of the contribution of knowledge categories (KC) in recovering genes associated with genetic diseases. A) A time-aware evaluation approach is compared to an evaluation based on a random partition of the initial gene set in assessing the impact of KC contribution in the gene recovery associated with genetic diseases. Time-aware evaluation consists of dividing the gene set into training and testing subsets based on the year in which genes were associated with disease. The training set is composed by the older genes, and the testing subset by the newer. The KC importance is measured using the integration score, calculated as the product of efficiency and exclusivity of KC. **B)** Comparison of integration scores when considering random and time-aware approaches for 246 curated disease/phenotype gene sets extracted from DisGeNET. Mean recall levels and error bars are represented. KCs are sorted based on median integration scores at random evaluation. **D.** Four disease gene sets ranked in the top 10 when sorted by additive KC variation.

Figure 4. General schema of the GLOWgenes algorithm. GLOWgenes takes as input a set of genes associated with a disease (GAD) and performs two steps in parallel: 1) Step 1, the disease-aware network evaluation (red arrows), where an integration score is calculated for each knowledge category (KC). The integration scores merge efficiency and exclusivity obtained for the best performing KC network; 2) Step 2, gene ranking for each KC (purple arrows), where GADs are propagated in every best performing network (Step 1) using random walk with restart (RWWR) algorithm, producing a ranking of all genes for every KC. KCs' integration scores and KCs' gene rankings are integrated into a GLOWgenes

gen ranking as the KC performance-weighted average of normalized gene scores.

Figure 5. Benchmarking of three data integration approaches for disease candidate gene prediction. A) General schema of three approaches for the integration of various datasets applied to the prioritization of phenotype-associated genes. In red, the composite network approach first merged the available networks into a composite network (1) that is the subject of the modeling and scoring step (2) to generate the gene ranking. In blue, the data fusion approach applies first modeling and scoring in every dataset (1) to merge them into a unique ranking using order statistics (2). In green, GLOWgenes approach takes two steps, the evaluation of the performance of every network (1) and the modeling and scoring of the networks producing a ranking for each one (2). In a third step, (1) and (2) are merged to produce a single gene ranking (3). **B)** Recall at various n-top of the performance of data integration methods. Precision-Recall-Gain curve associated area (AUPRG) is shown for every method. Methods are colored based on the types of approaches described in (A). C) Percentage of diseases (PanelApp disease gene sets) for which each method performed best, measure at various n-top. Methods are colored based on the types of approaches described in (A).

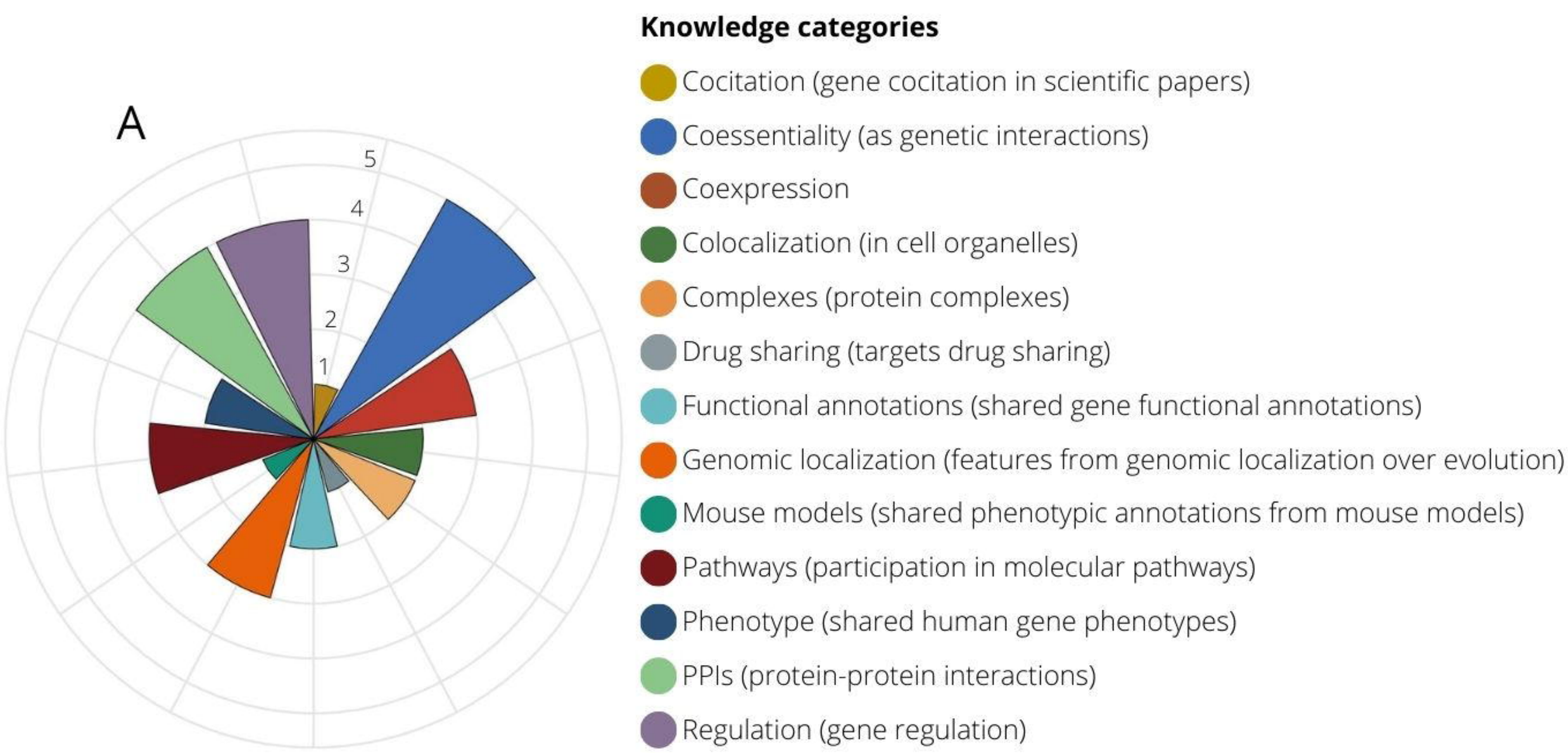
Figure 6. Benchmarking of tools for disease candidate genes prioritization. A) Comparative evaluation of 10 tools for the prioritization of gene disease candidates using Green-Ambar genes from PanelApp (GAGs, high evidence disease-associated genes) for validation in four selected diseases (pilot diseases). Recall levels at different n-top genes are represented. **B)** Same to A, but using red genes from PanelApp (RGs, low-evidence disease-associated genes) as validation sets**.**

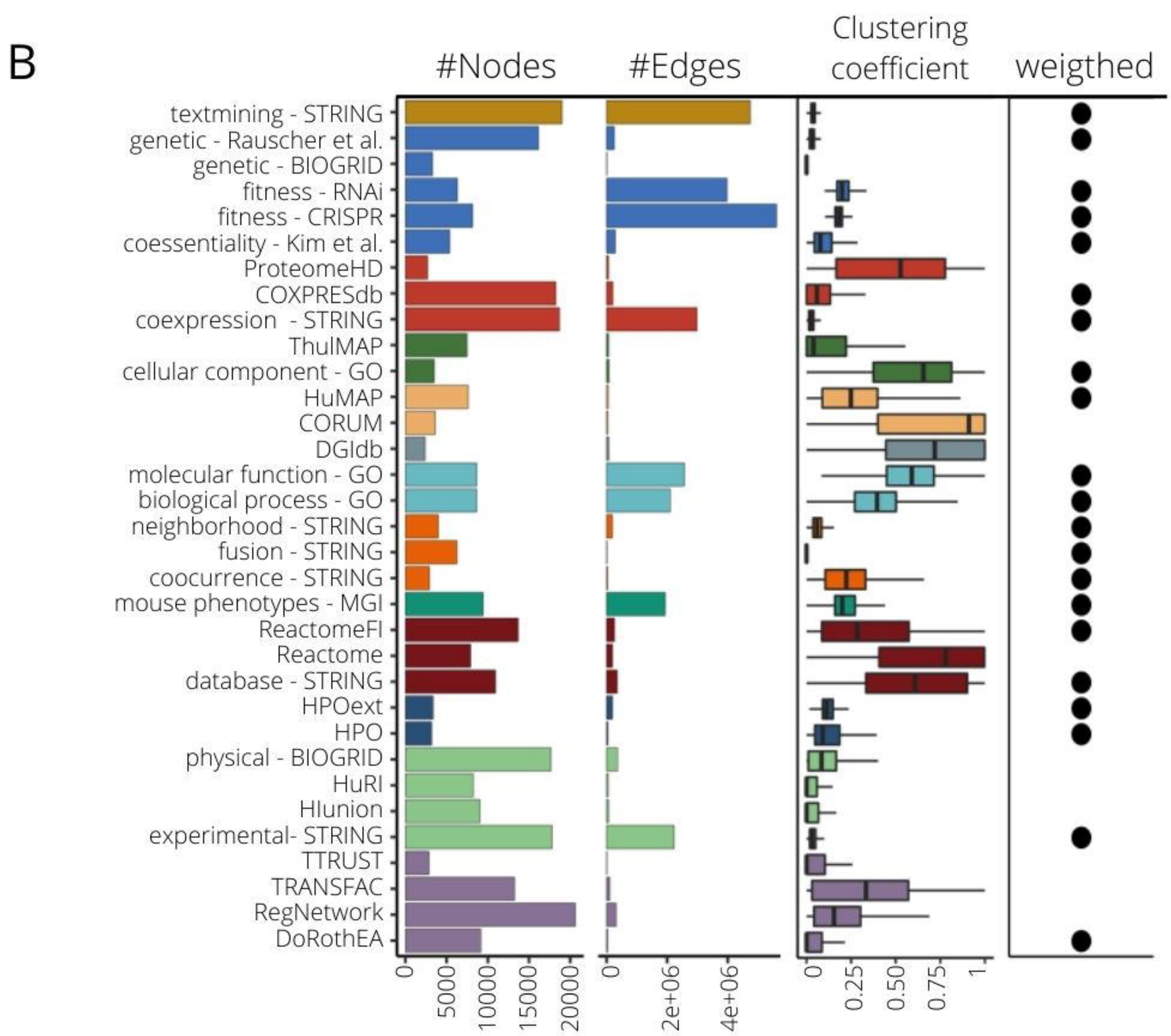
REFERENCES

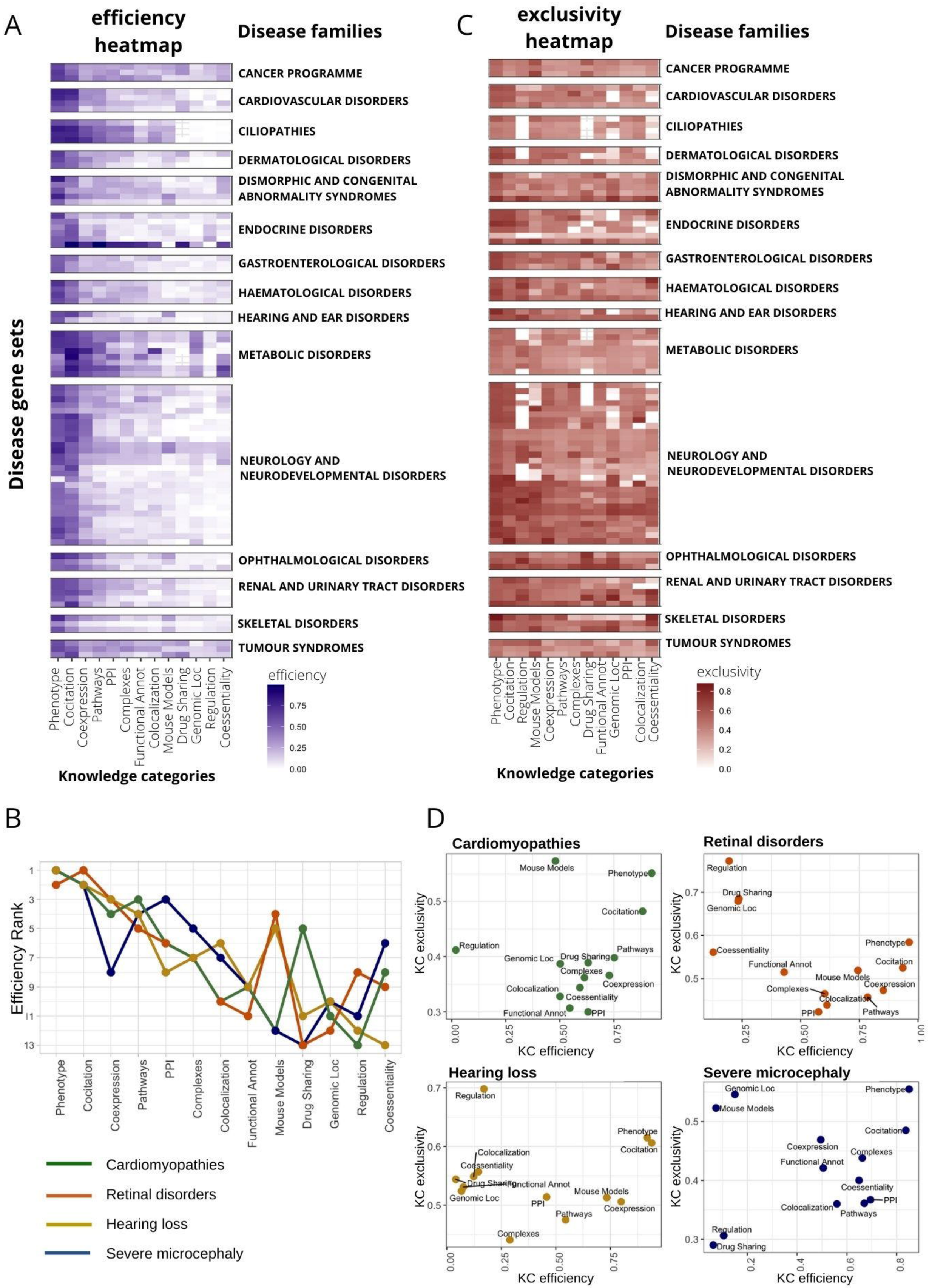
1. Iancu,I.-F., Perea-Romero,I., Núñez-Moreno,G., de la Fuente,L., Romero,R., Ávila-Fernandez,A., Trujillo-Tiebas,M.J., Riveiro-Álvarez,R., Almoguera,B., Martín-Mérida,I., *et al.* (2022) Aggregated Genomic Data as Cohort-Specific Allelic Frequencies can Boost Variants and Genes Prioritization in Non-Solved Cases of Inherited Retinal Dystrophies. *Int J Mol Sci*, **23**, 8431.
2. Perea-Romero,I., Gordo,G., Iancu,I.F.I.F.I.F., Del Pozo-Valero,M., Almoguera,B., Blanco-Kelly,F., Carreño,E., Jimenez-Rolando,B., Lopez-Rodriguez,R., Lorda-Sanchez,I., *et al.* (2021) Genetic landscape of 6089 inherited retinal dystrophies affected cases in Spain and their therapeutic and extended epidemiological implications. *Scientific Reports*, **11**.
3. LC,T., A,A., S,E., D,A.A., J,A., D,A.A., Y,M., Tranchevent,L.C., Ardeshirdavani,A., ElShal,S., *et al.* (2016) Candidate gene prioritization with Endeavour. *Nucleic acids research*, **44**, W117–W121.
4. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, **37**.
5. Aguirre-Plans,J., Piñero,J., Sanz,F., Furlong,L.I., Fernandez-Fuentes,N., Oliva,B. and Guney,E. (2019) GUILDify v2.0: A Tool to Identify Molecular Networks Underlying Human Diseases, Their Comorbidities and Their Druggable Targets. *Journal of molecular biology*, **431**, 2477–2484.
6. Ghiassian,S.D., Menche,J. and Barabási,A.L. (2015) A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLOS Computational Biology*, **11**, e1004120.
7. D,W.-F., SL,D., O,C., K,Z., R,B., P,C., M,F., C,G., F,K., CT,L., *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, **38**.
8. Pinero,J., Queralt-Rosinach,N., Bravo,A., Deu-Pons,J., Bauer-Mehren,A., Baron,M., Sanz,F. and Furlong,L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, **2015**, bav028–bav028.
9. S,P.-F., A,P., K,T., JX,B. and LJ,J. (2015) DISEASES: text mining and data integration of disease-gene associations. *Methods (San Diego, Calif.)*, **74**, 83–89.
10. Yang,H., Robinson,P.N. and Wang,K. (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature methods*, **12**, 841–3.
11. Luo,P., Chen,B., Liao,B. and Wu,F.-X. (2021) Predicting disease-associated genes: Computational methods, databases, and evaluations. *WIREs Data Mining and Knowledge Discovery*, **11**, e1383.
12. Bhasuran,B. and Natarajan,J. (2018) Automatic extraction of gene-disease associations from literature using joint ensemble learning. *PLOS ONE*, **13**, e0200699.
13. Minguez,P., Götz,S., Montaner,D., Al-Shahrour,F. and Dopazo,J. (2009) SNOW, a web-based tool for the statistical analysis of protein-protein

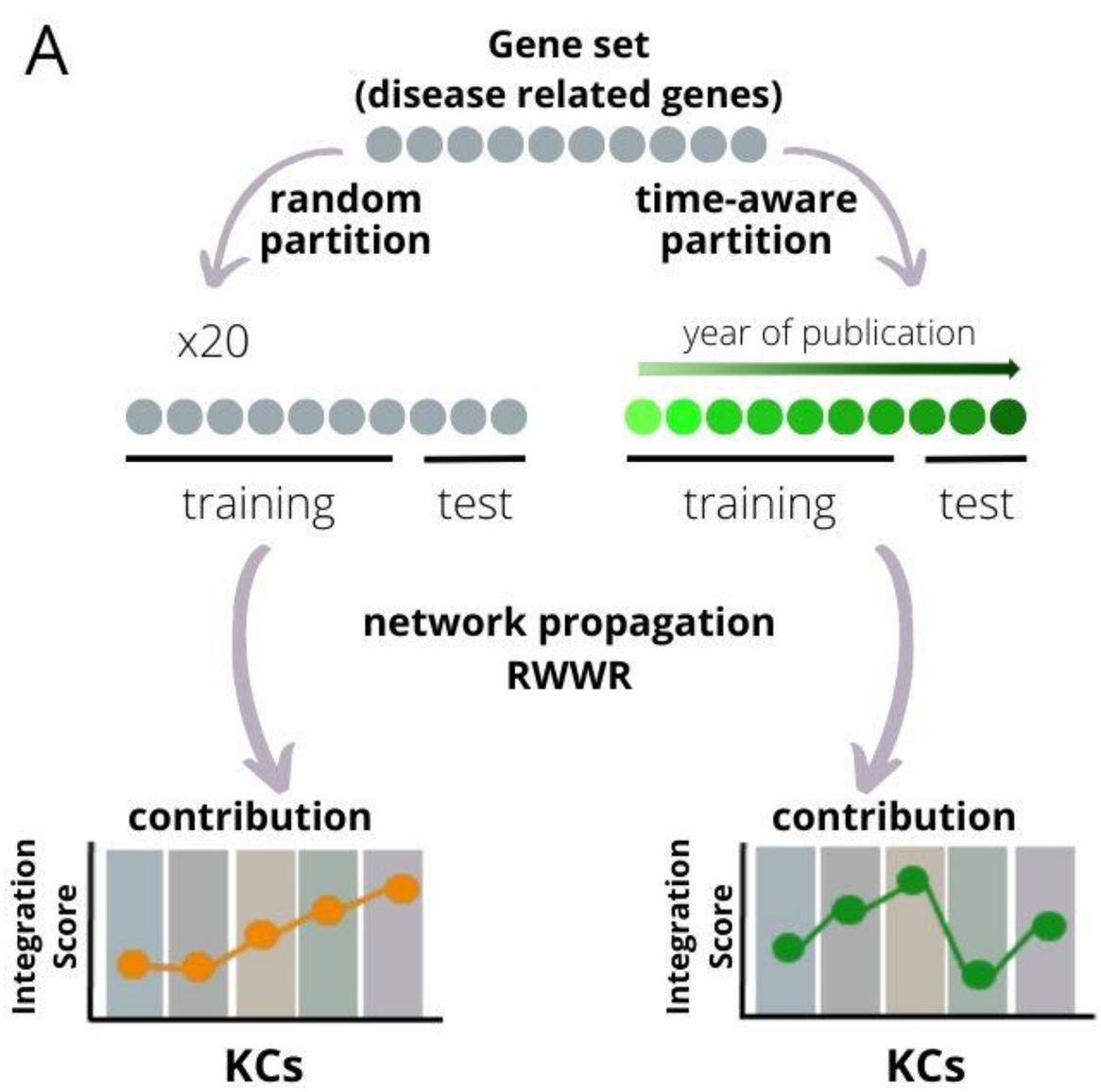
interaction networks. *Nucleic acids research*, **37**, W109-14.
14. Mordelet,F. and Vert,J.-P. (2011) ProDiGe: Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, **12**, 389.
15. JK,H., DE,C., MK,Y., W,Z., JF,K., P,T. and T,I. (2018) Systematic Evaluation of Molecular Networks for Discovery of Disease Genes. *Cell systems*, **6**, 484-495.e5.
16. Fernández-Torras,A., Duran-Frigola,M., Bertoni,M., Locatelli,M. and Aloy,P. (2022) Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. *Nat Commun*, **13**, 5304.
17. Martin,A.R., Williams,E., Foulger,R.E., Leigh,S., Daugherty,L.C., Niblock,O., Leong,I.U.S., Smith,K.R., Gerasimenko,O., Haraldsdottir,E., *et al.* (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature genetics*, **51**, 1560–1565.
18. D,S., AL,G., D,L., A,J., S,W., J,H.-C., M,S., NT,D., JH,M., P,B., *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, **47**, D607–D613.
19. T,O., Y,K., Y,A., S,T. and K,K. (2019) COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic acids research*, **47**, D55–D62.
20. H,H., JW,C., S,L., A,Y., H,K., D,B., S,Y., CY,K., M,L., E,K., *et al.* (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, **46**, D380–D386.
21. L,G.-A., CH,H., MM,I., D,T. and J,S.-R. (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research*, **29**, 1363–1375.
22. E,W. (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in bioinformatics*, **9**, 326–332.
23. PJ,T., L,Å., M,W., D,M., A,G., H,A.B., T,A., A,A., L,B., LM,B., *et al.* (2017) A subcellular map of the human proteome. *Science (New York, N.Y.)*, **356**.
24. G,K., P,G., TA,S., JB,P., M,S. and J,R. (2019) Co-regulation map of the human proteome enables identification of protein functions. *Nature biotechnology*, **37**, 1361–1371.
25. R,O., C,S., BJ,B., J,R., L,B., C,C., N,K., L,O., G,L., R,M., *et al.* (2019) The BioGRID interaction database: 2019 update. *Nucleic acids research*, **47**, D529–D541.
26. K,L., DK,K., L,L., K,S., BE,B., W,B., R,B., T,C., FJ,C.-L., B,C., *et al.* (2020) A reference map of the human binary protein interactome. *Nature*, **580**, 402–408.
27. B,R., F,H., L,H., T,H., O,V. and M,B. (2018) Toward an integrated map of genetic interactions in cancer cells. *Molecular systems biology*, **14**.
28. Wu,G., Haw,R., G,W., R,H., Wu,G. and Haw,R. (2017) Functional Interaction Network Construction and Analysis for Disease Discovery Methods Mol Biol.
29. E,K., M,D., WF,L., G,W., S,S., M,C. and T,H. (2019) A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life science alliance*, **2**.
30. ZP,L., C,W., H,M. and H,W. (2015) RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database: the journal of biological databases and curation*, **2015**, 1–12.
31. K,D., C,L., RL,H., F,T., B,B., CD,M., Y,M., JB,W. and EM,M. (2017) Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular systems biology*, **13**, 932.
32. S,K., NA,V., M,E., E,F., J,M., S,A., G,B., SM,B., CF,B., KM,B., *et al.* (2017) The human phenotype ontology in 2017. *Nucleic Acids Research*, **45**, D865–D876.
33. CJ,B., JA,B., CL,S., JA,K. and JE,R. (2019) Mouse Genome Database (MGD) 2019. *Nucleic acids research*, **47**, D801–D806.
34. KC,C., AH,W., YY,F., S,K., AC,C., G,S., A,W., NC,S., OL,G. and M,G. (2018) DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic acids research*, **46**, D1068–D1073.
35. J,P., RM,M., BC,M., N,M., AE,S., JN,W., SH,C., F,V., BA,W., WC,H., *et al.* (2018) Interrogation of Mammalian Protein Complex Structure, Function, and Membership Using Genome-Scale Fitness Screens. *Cell systems*, **6**, 555-568.e7.
36. M,G., J,R., B,B., I,D.-K., G,F., G,F., C,M. and A,R. (2019) CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic acids research*, **47**, D559–D563.
37. C,Z. and Z,W. (2018) GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Scientific reports*, **8**.
38. The Gene Ontology Resource: 20 years and still GOing strong (2019) *Nucleic acids research*, **47**, D330–D338.
39. B,J., L,M., G,V., C,G., P,L., A,F., K,S., J,C., M,G., R,H., *et al.* (2020) The reactome pathway knowledgebase. *Nucleic acids research*, **48**, D498–D503.
40. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P., *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–7.
41. K?hler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, **82**, 949–958.
42. Martínez,O. and Reyes-Valdés,M.H. (2008) Defining diversity, specialization, and gene specificity in transcriptomes through information theory. *Proceedings of the National Academy of Sciences*, **105**, 9709–9714.
43. Flach,P. and Kull,M. Precision-Recall-Gain Curves: PR Analysis Done Right. *NeurIPS Proceedings*.
44. J,P., JM,R.-A., J,S.-P., F,R., E,C., F,S. and LI,F. (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, **48**, D845–D855.
45. J,L., X,L., Y,T., S,Q., D,X., J,Z. and Y,K. (2016) A Comprehensive Evaluation of Disease Phenotype Networks for Gene Prioritization. *PloS one*, **11**.
46. G,V., A,P., H,C., AE,R. and M,R. (2014) An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artificial intelligence in medicine*, **61**, 63–78.
47. S,A., D,L., S,M., P,V.L., B,C., F,D.S., LC,T., B,D.M., P,M., B,H., *et al.* (2006) Gene prioritization through genomic data fusion. *Nature biotechnology*, **24**, 537–544.
48. R,K., S,L., P,A., J,V., Kolde,R., Laur,S., Adler,P. and Vilo,J. (2012) Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics (Oxford, England)*, **28**, 573–580.
49. Piñero,J., Bravo,A., Queralt-Rosinach,N., Gutierrez-Sacristan,A., Deu-Pons,J., Centeno,E., Garcia-Garcia,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, **45**, D833–D839.
50. Chen,J., Aronow,B.J. and Jegga,A.G. (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, **10**, 1–14.
51. Chen,J., Xu,H., Aronow,B.J. and Jegga,A.G. (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC bioinformatics*, **8**.
52. Romero,R. et al. et al. (2021) An evaluation of pipelines for DNA variant detection can guide a reanalysis protocol to increase the diagnostic ratio of genetic diseases. *npj genomic medicine*, **in press**.
53. Graessner,H., Zurek,B., Hoischen,A. and Beltran,S. (2021) Solving the unsolved rare diseases in Europe. *Eur J Hum Genet*, **29**, 1319–1320.
54. Austin,C.P., Cutillo,C.M., Lau,L.P.L., Jonker,A.H., Rath,A., Julkowska,D., Thomson,D., Terry,S.F., de Montleau,B., Ardigò,D., *et al.* (2018) Future of Rare Diseases Research 2017-2027: An IRDiRC Perspective. *Clin Transl Sci*, **11**, 21–27.
55. Dopazo,J., Amadoz,A., Bleda,M., Garcia-Alonso,L., Alemán,A., García-García,F., Rodriguez,J.A., Daub,J.T., Muntané,G., Rueda,A., *et al.* (2016) 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. *Molecular biology and evolution*, 10.1093/molbev/msw005.

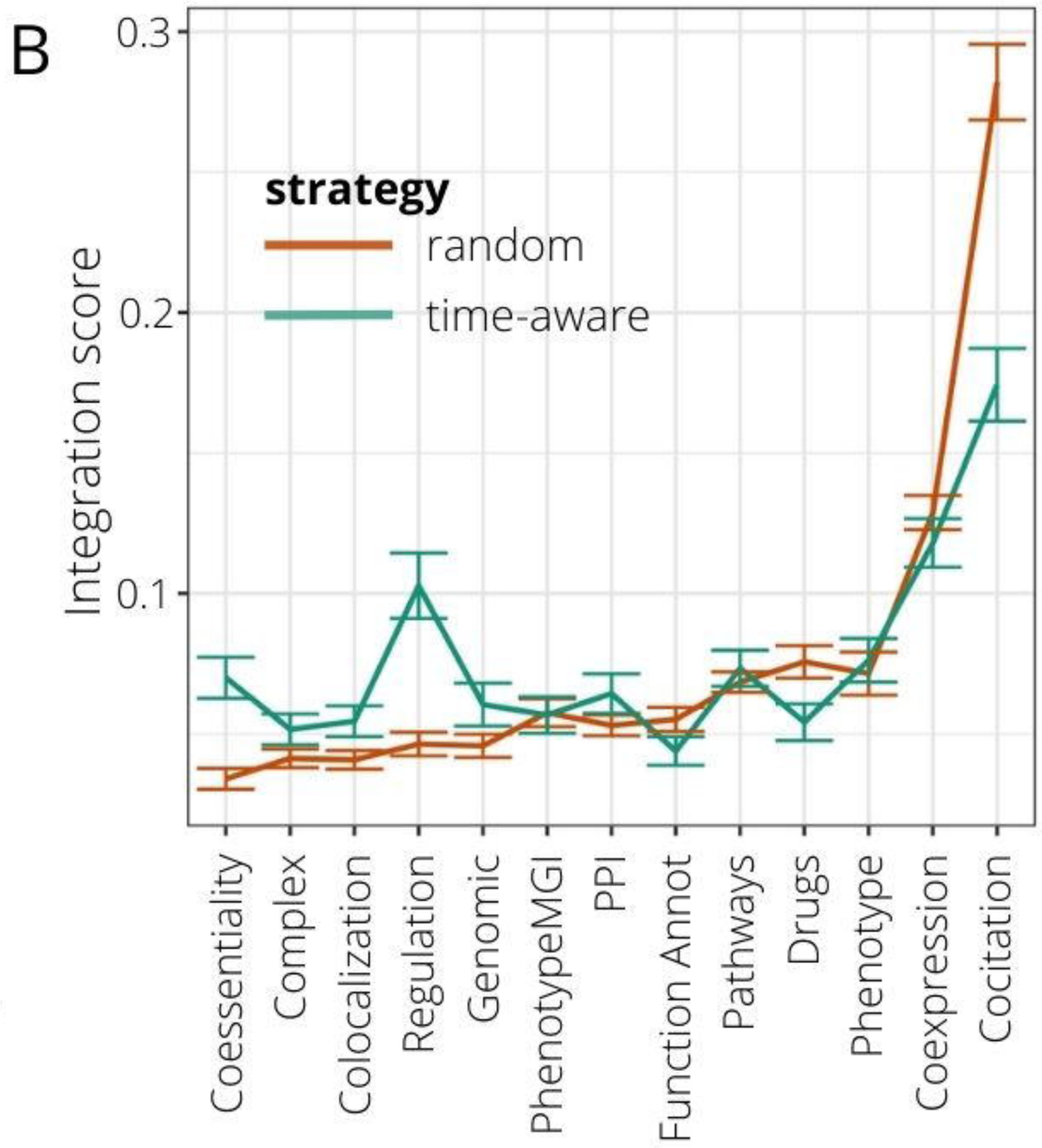
56. Gudmundsson,S., Singer-Berk,M., Watts,N.A., Phu,W., Goodrich,J.K., Solomonson,M., Genome Aggregation Database Consortium, Rehm,H.L., MacArthur,D.G. and O'Donnell-Luria,A. (2022) Variant interpretation using population databases: Lessons from gnomAD. *Hum Mutat*, **43**, 1012–1030.
57. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.
58. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic acids research*, **31**, 258–61.
59. Page,L., Brin,S., Motwani,R. and Winograd,T. (1999) The PageRank Citation Ranking: Bringing Order to the Web.
60. Vidal,M., Cusick,M.E. and Barabási,A.-L. (2011) Interactome Networks and Human Disease. *Cell*, **144**, 986–998.
61. Menche,J., Sharma,A., Kitsak,M., Ghiassian,S.D., Vidal,M., Loscalzo,J. and Barabási,A.-L. (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.
62. Valdeolivas,A., Tichit,L., Navarro,C., Perrin,S., Odelin,G., Levy,N., Cau,P., Remy,E. and Baudot,A. (2019) Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, **35**, 497–505.
63. Cantini,L., Medico,E., Fortunato,S. and Caselle,M. (2015) Detection of gene communities in multi-networks reveals cancer drivers. *Sci Rep*, **5**, 17386.
64. Núñez-Carpintero,I., Petrizzelli,M., Zinovyev,A., Cirillo,D. and Valencia,A. (2021) The multilayer community structure of medulloblastoma. *iScience*, **24**.
65. Reiter,J.F. and Leroux,M.R. (2017) Genes and molecular pathways underpinning ciliopathies. *Nat Rev Mol Cell Biol*, **18**, 533–547.
66. López-Rodríguez,R., Del Pozo-Valero,M., Corton,M., Minguez,P., Ruiz-Hornillos,J., Pérez-Tomás,M.E., Barreda-Sánchez,M., Mancebo,E., Villaverde,C., Núñez-Moreno,G., *et al.* (2022) Presence of rare potential pathogenic variants in subjects under 65 years old with very severe or fatal COVID-19. *Sci Rep*, **12**, 10369.
67. Heussler H.S., Suri M., Young I.D., Muenke M. (2002) Extreme variability of expression of a Sonic Hedgehog mutation: attention difficulties and holoprosencephalyArchives of Disease in Childhood;**86**:293-296.
68. Perea-Romero,I., Blanco-Kelly,F., Sanchez-Navarro,I., Lorda-Sanchez,I., Tahsin-Swafiri,S., Avila-Fernandez,A., Martin-Merida,I., Trujillo-Tiebas,M.J.J., Lopez-Rodriguez,R., Rodriguez de Alba,M., *et al.* (2021) NGS and phenotypic ontology-based approaches increase the diagnostic yield in syndromic retinal diseases. *Human genetics*, **140**, 1665–1678.
69. Wu,G. and Haw,R. (2017) Functional Interaction Network Construction and Analysis for Disease Discovery. *Methods Mol Biol*, **1558**, 235–253.
70. Kim,E., Dede,M., Lenoir,W.F., Wang,G., Srinivasan,S., Colic,M. and Hart,T. (2019) A network of human functional gene interactions from knockout fitness screens in cancer cells. *Life Sci Alliance*, **2**, e201800278.
71. Köhler,S., Bauer,S., Horn,D. and Robinson,P.N. (2008) Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics*, **82**, 949–958.
72. Vanunu,O., Magger,O., Ruppin,E., Shlomi,T. and Sharan,R. (2010) Associating Genes and Protein Complexes with Disease via Network Propagation. *PLOS Computational Biology*, **6**, e1000641.
73. Hofree,M., Shen,J.P., Carter,H., Gross,A. and Ideker,T. (2013) Network-based stratification of tumor mutations. *Nat Methods*, **10**, 1108–1115.
74. Leiserson,M.D.M., Vandin,F., Wu,H.-T., Dobson,J.R., Eldridge,J.V., Thomas,J.L., Papoutsaki,A., Kim,Y., Niu,B., McLellan,M., *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet*, **47**, 106–114.
75. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science (New York, N.Y.)*, **302**, 249–55.
76. Zuberi,K., Franz,M., Rodriguez,H., Montojo,J., Lopes,C.T., Bader,G.D. and Morris,Q. (2013) GeneMANIA Prediction Server 2013 Update. *Nucleic Acids Research*, **41**, W115–W122.

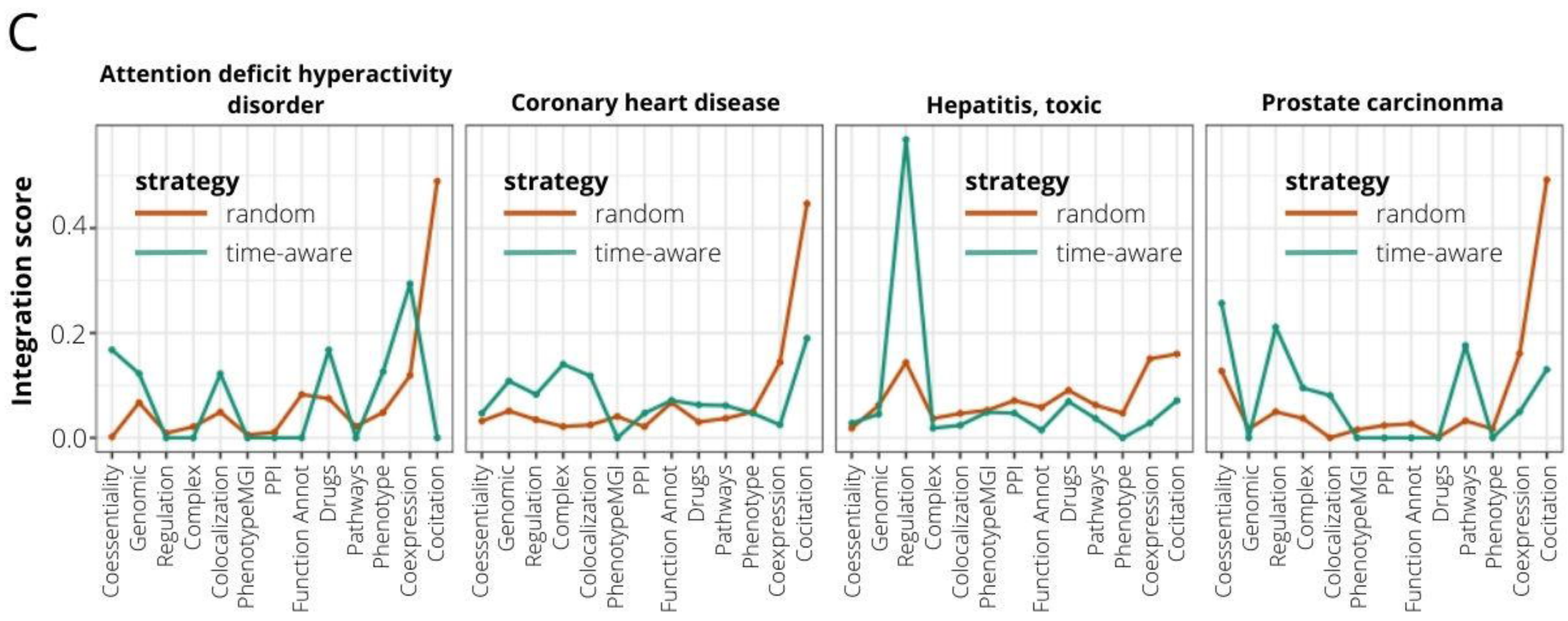

A
5
4
3
2
1
Knowledge categories
Cocitation (gene cocitation in scientific papers)
Coessentiality (as genetic interactions)
Coexpression
Colocalization (in cell organelles)
Complexes (protein complexes)
Drug sharing (targets drug sharing)
Functional annotations (shared gene functional annotations)
Genomic localization (features from genomic localization over evolution)
Mouse models (shared phenotypic annotations from mouse models)
Pathways (participation in molecular pathways)
Phenotype (shared human gene phenotypes)
PPIs (protein-protein interactions)
Regulation (gene regulation)

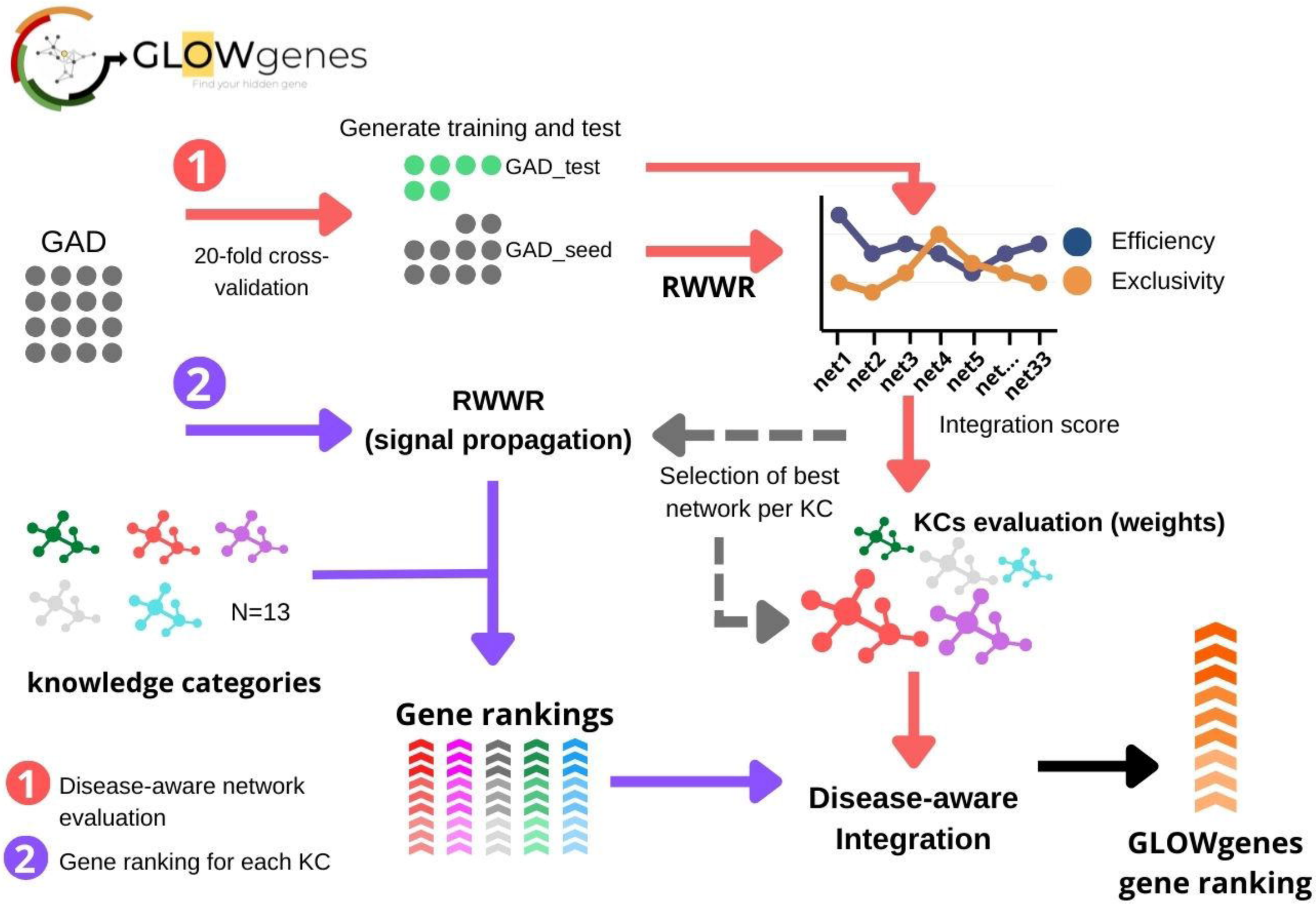

B
#Nodes
#Edges
Clustering coefficient
weigthed
textmining - STRING
genetic - Rauscher et al.
genetic - BIOGRID
fitness - RNAi
fitness - CRISPR
coessentiality - Kim et al.
ProteomeHD
COXPRESdb
coexpression - STRING
ThulMAP
cellular component - GO
HuMAP
CORUM
DGIdb
molecular function - GO
biological process - GO
neighborhood - STRING
fusion - STRING
coocurrence - STRING
mouse phenotypes - MGI
ReactomeFI
Reactome
database - STRING
HPOext
HPO
physical - BIOGRID
HuRI
HIunion
experimental- STRING
TTRUST
TRANSFAC
RegNetwork
DoRothEA
0
5000
10000
15000
20000
0
2e+06
4e+06
0
0.25
0.50
0.75
1

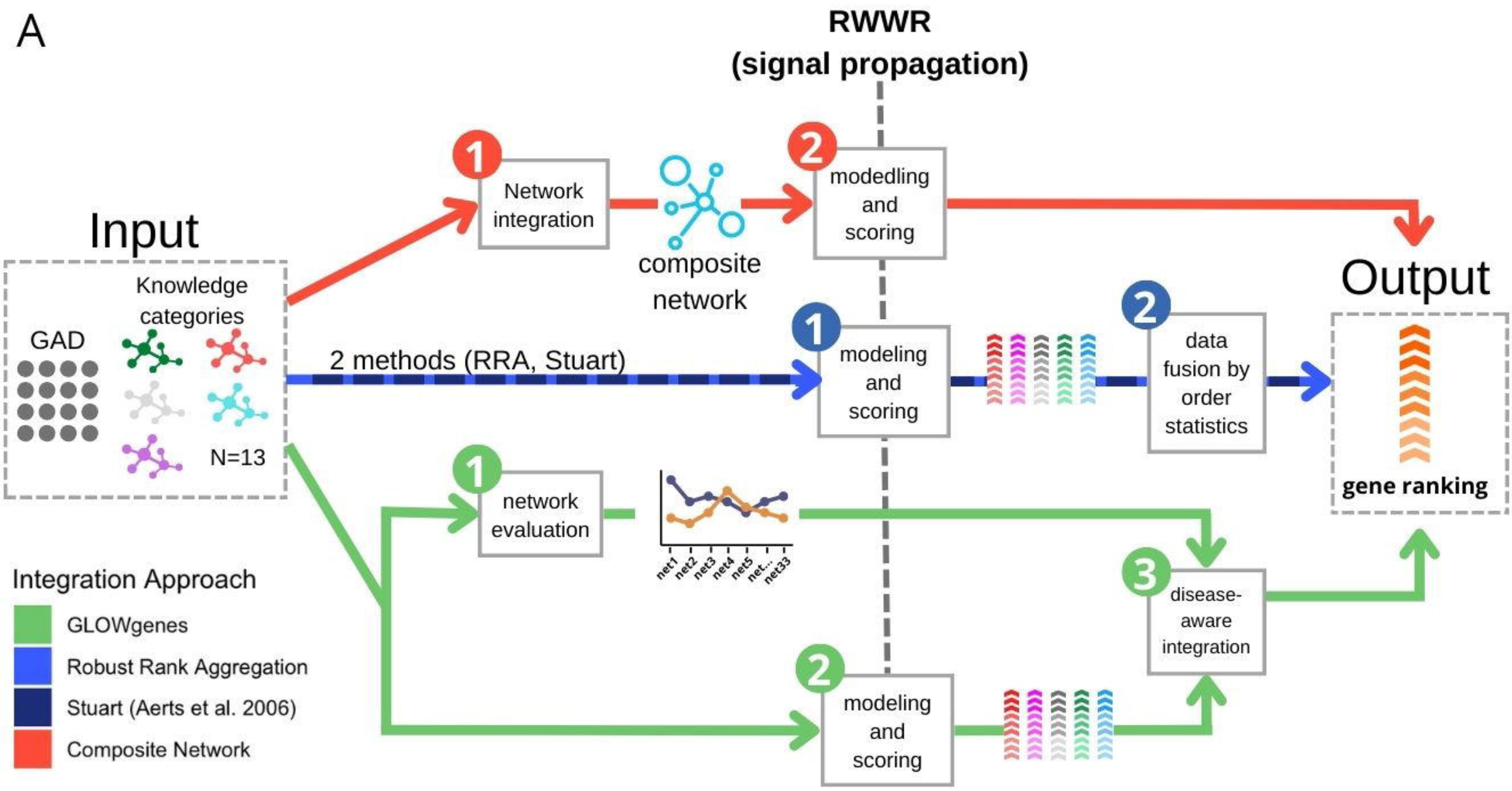

A
efficiency heatmap
Disease families
CANCER PROGRAMME
CARDIOVASCULAR DISORDERS
CILIOPATHIES
DERMATOLOGICAL DISORDERS
DISMORPHIC AND CONGENITAL ABNORMALITY SYNDROMES
ENDOCRINE DISORDERS
GASTROENTEROLOGICAL DISORDERS
HAEMATOLOGICAL DISORDERS
HEARING AND EAR DISORDERS
METABOLIC DISORDERS
NEUROLOGY AND NEURODEVELOPMENTAL DISORDERS
OPHTHALMOLOGICAL DISORDERS
RENAL AND URINARY TRACT DISORDERS
SKELETAL DISORDERS
TUMOUR SYNDROMES
Disease gene sets
Phenotype
Cocitation
Coexpression
Pathways
PPI
Complexes
Functional Annot
Colocalization
Mouse Models
Drug Sharing
Genomic Loc
Regulation
Coessentiality
Knowledge categories
efficiency
0.75
0.50
0.25
0.00
C
exclusivity heatmap
Disease families
Phenotype
Cocitation
Regulation
Mouse Models
Coexpression
Pathways
Complexes
Drug Sharing
Funtional Annot
Genomic Loc
PPI
Colocalization
Coessentiality
Knowledge categories
exclusivity
0.8
0.6
0.4
0.2
0.0
B
Efficiency Rank
Cardiomyopathies
Retinal disorders
Hearing loss
Severe microcephaly
D
Cardiomyopathies
Retinal disorders
Hearing loss
Severe microcephaly
KC exclusivity
KC efficiency

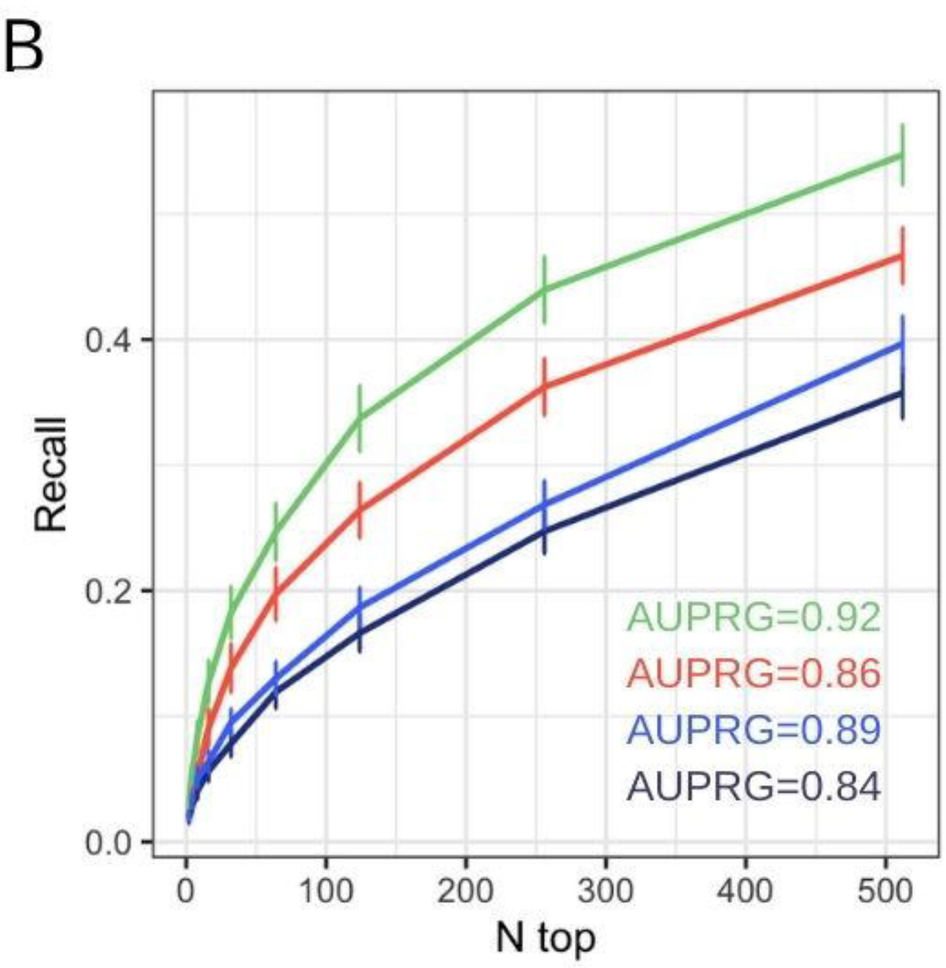

A
Gene set
(disease related genes)
random
partition
time-aware
partition
x20
year of publication
training
test
training
test
network propagation
RWWR
contribution
Integration
Score
KCs
contribution
Integration
Score
KCs

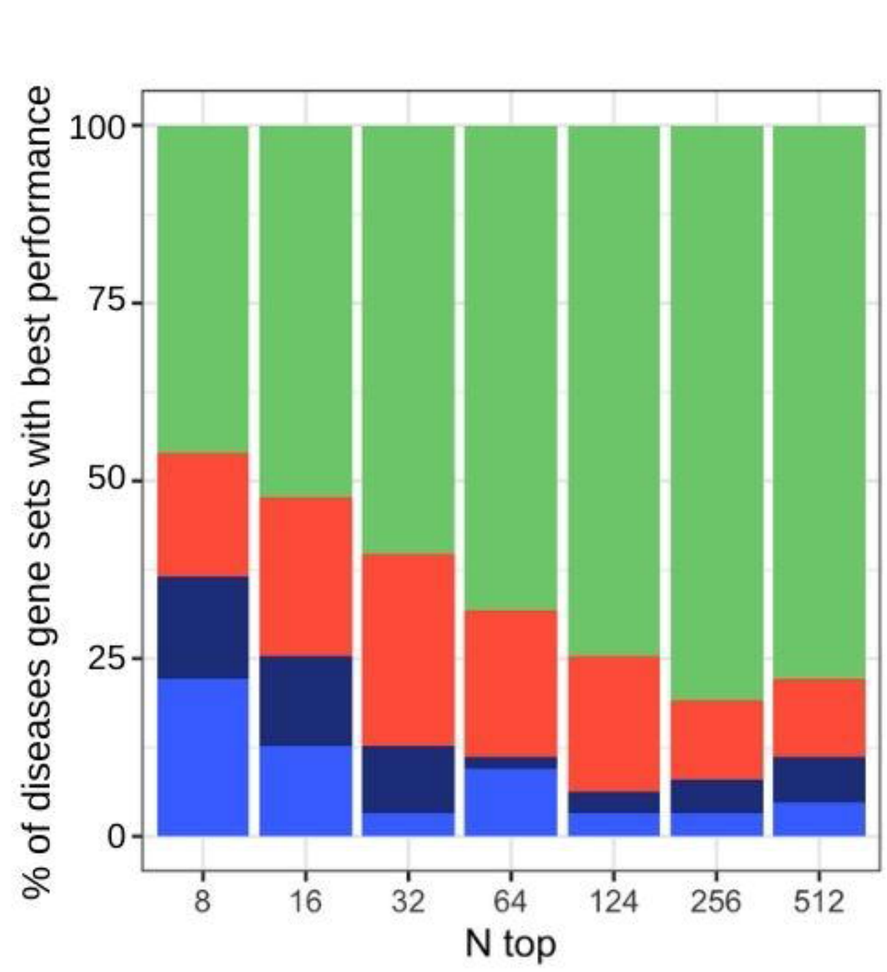

B
Integration score
0.3
0.2
0.1
strategy
random
time-aware
Coessentiality
Complex
Colocalization
Regulation
Genomic
PhenotypeMGI
PPI
Function Annot
Pathways
Drugs
Phenotype
Coexpression
Cocitation

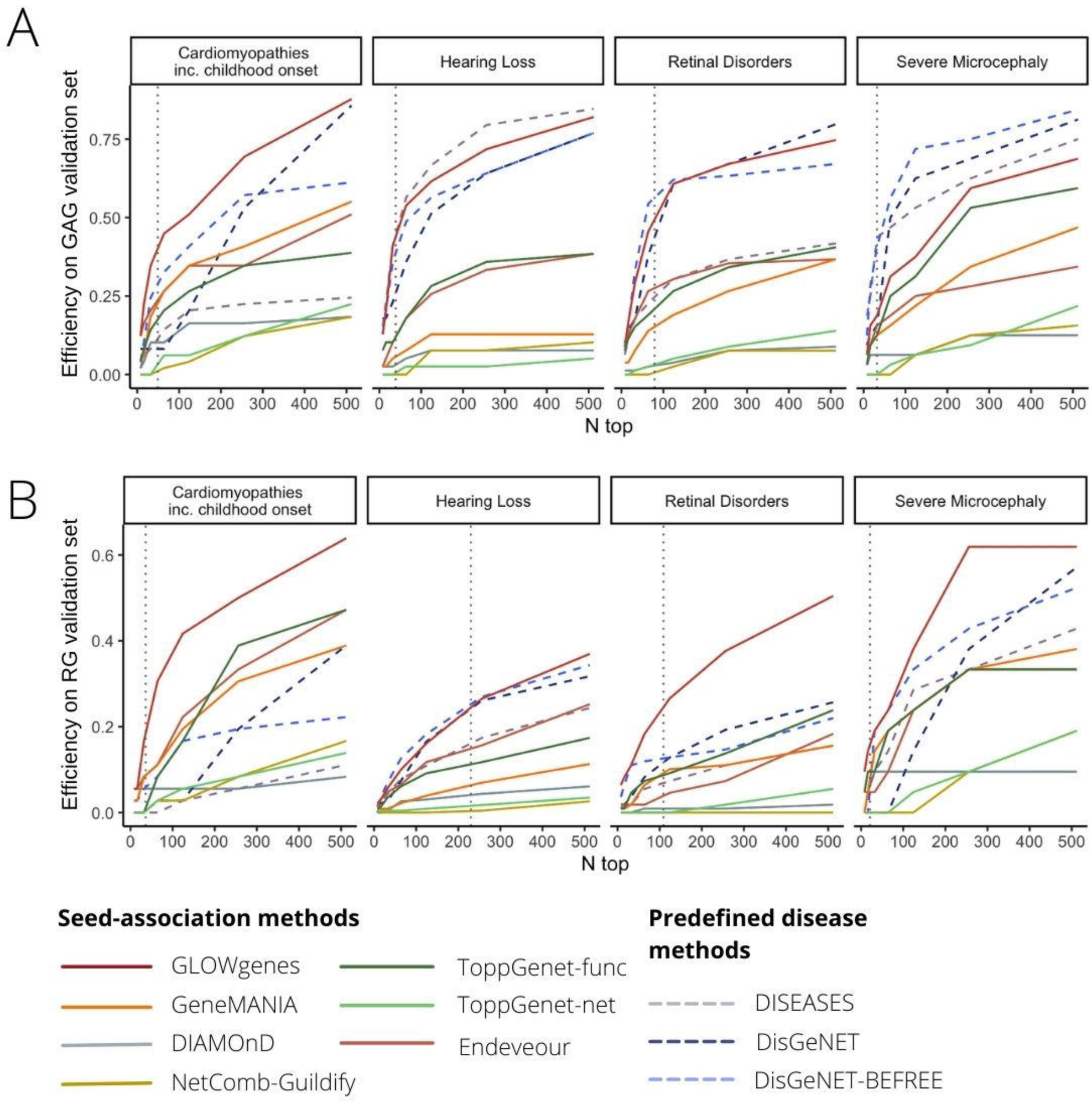

C
Attention deficit hyperactivity disorder
Coronary heart disease
Hepatitis, toxic
Prostate carcinonma
Integration score
0.4
0.2
0.0
strategy
random
time-aware
Coessentiality
Genomic
Regulation
Complex
Colocalization
PhenotypeMGI
PPI
Function Annot
Drugs
Pathways
Phenotype
Coexpression
Cocitation


GLOWgenes
Find your hidden gene
Generate training and test
1
GAD_test
GAD
GAD_seed
20-fold cross-validation
RWWR
Efficiency
Exclusivity
net1
net2
net3
net4
net5
net...
net33
2
RWWR
(signal propagation)
Integration score
Selection of best network per KC
KCs evaluation (weights)
N=13
knowledge categories
Gene rankings
1 Disease-aware network evaluation
2 Gene ranking for each KC
Disease-aware Integration
GLOWgenes gene ranking


A
RWWR
(signal propagation)
Input
Knowledge categories
GAD
N=13
Network integration
composite network
modedling and scoring
2 methods (RRA, Stuart)
modeling and scoring
data fusion by order statistics
Output
gene ranking
network evaluation
net1
net2
net3
net4
net5
net...
net33
disease-aware integration
modeling and scoring
Integration Approach
GLOWgenes
Robust Rank Aggregation
Stuart (Aerts et al. 2006)
Composite Network


B
Recall
0.0
0.2
0.4
0
100
200
300
400
500
N top
AUPRG=0.92
AUPRG=0.86
AUPRG=0.89
AUPRG=0.84


C
% of diseases gene sets with best performance
0
25
50
75
100
8
16
32
64
124
256
512
N top


A
Cardiomyopathies inc. childhood onset
Hearing Loss
Retinal Disorders
Severe Microcephaly
Efficiency on GAG validation set
0.00
0.25
0.50
0.75
0
100
200
300
400
500
N top
B
Efficiency on RG validation set
0.0
0.2
0.4
0.6
Seed-association methods
GLOWgenes
GeneMANIA
DIAMOnD
NetComb-Guildify
ToppGenet-func
ToppGenet-net
Endeveour
Predefined disease methods
DISEASES
DisGeNET
DisGeNET-BEFREE