bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genome-wide discovery of hidden genes mediating known drug-disease association

Hua Yu [1, 4#*], Lu Lu [2,#], Ming Chen [4], Chen Li [3,*], Jin Zhang [1,*]

[1] Center for Stem Cell and Regenerative Medicine, Department of Basic Medical Sciences & The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China. Institute of Hematology, Zhejiang University, Hangzhou, Zhejiang, China.

[2] Department of Human Genetics, Zhejiang University School of Medicine, Hangzhou, China.

[3] Department of Human Genetics, and Women's Hospital, Zhejiang University School of Medicine, Hangzhou, China

[4] College of Life Sciences, Zhejiang University, Hangzhou, China.

[5] Lead contact

[#] These authors contributed equally.

[*] Correspondence: huayu@zju.edu.cn, chenli2012@zju.edu.cn, zhgene@zju.edu.cn

Abstract

Identifying of hidden genes mediating Known Drug-Disease Association (KDDA) is of great significance for understanding disease pathogenesis and guiding drug repurposing. Here, we present a novel computational tool, called KDDANet, for systematic and accurate uncovering hidden genes mediating KDDA from the perspective of genome-wide gene functional interaction network. By implementing minimum cost flow optimization, combined with depth first searching and graph clustering on a unified flow network model, KDDANet outperforms existing methods in both sensitivity and specificity of identifying genes in mediating KDDA. Case studies on Alzheimer's disease (AD) and obesity uncover the mechanistic relevance of KDDANet predictions. Furthermore, when applied with multiple types of cancer-omics datasets, KDDANet not only recapitulates known genes mediating KDDAs related to cancer, but also uncovers novel candidates that offer new biological insights. Importantly, KDDANet can be used to discover the shared genes mediating multiple KDDAs. KDDANet can be accessed at http://www.kddanet.cn and the code can be freely downloaded at https://github.com/huayu1111/KDDANet/.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

The conventional development of novel promising drugs for treating specific diseases is a time-consuming and efforts-costing process, including discovery of new chemical entities, target detection and verification, preclinical and clinical trials and so on (Paul, et al., 2010). Furthermore, the success rate for a new drug to be taken to market is very low, usually only 10% per year of drugs approved by FDA and thus prevent their use in actual therapy (Weng et al., 2013). This results in that pharmaceutical research faces a decreasing productivity in drug development and a sustaining gap between therapeutic needs and available treatments (Alaimo and Pulvirenti, 2019). Compared with traditional drug development, drug repositioning, i.e., finding the novel indications of existing drugs, offers the possibility for safer and faster drug development because of several steps of traditional drug development pipeline can be avoided during repurposing efforts (Ashburn and Thor, 2004). Many successful cases of repositioned drugs have been shown: from *Minoxidil*, designed for treatment of hypertension and now indicated for hair loss (Bradley, 2005), to *Sildenafil*, developed for patients with heart problems and repurposed for erectile dysfunction (Ghofrani et al., 2006). Yet, these examples are based on clinical observations of secondary effects. Thanks to the advance in next generation omics sequencing and qualification technologies, a large volume of biomedical datasets has been rapidly accumulated. Based on these datasets, various computational tools, including machine learning, similarity computation and network-based models, have been proposed to systematically predict novel drug targets and drug-disease associations for facilitating drug repurposing (Ashburn and Thor, 2004; Lotfi Shahreza et al., 2018; Lu and Yu, 2018; Moridi et al., 2019; Xuan et al.,2019; Alaimo and Pulvirenti, 2019). These methods can exploit and integrate multi-level omics data sources for discovery of novel drug indications and thus accelerating drug repositioning process. However, no existing method is designed for identifying the hidden genes mediating Known Drug-Disease Association (KDDA), which we believe can provide additional contributions to the development of drug repurposing and

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

new therapies.

Theoretically, drug repurposing can be proposed based on two molecular aspects. I) On one hand, complex diseases often involve multiple genetic and environmental determinants, including multi-factor driven alterations and dysregulation of a series of genes (Goh, et al., 2007), which will propagate and perturb certain biological processes by the interactions among molecules, leading to the onset of diseases. II) On the other hand, one drug can exert impacts on many targets and perturb multiple biological processes (Yildirim, et al., 2007). As a result, the shared biological pathways manifested in certain disease state and induced by a known drug suggests potential drug repurposing. Therefore, developing the mechanism-oriented computational tools to unveil the hidden components of shared cellular pathways is of great significance for understanding disease pathogenesis and guiding drug repurposing. However, not much effort is spent to address this need from a perspective of genome-wide gene functional interaction network. To comprehensive understanding the molecular mechanism of individual KDDA and the shared molecular processes among multiple KDDAs, it is essential to find hidden genes, or intermediary genes that bridge drug and its related disease and obtain a global picture of disease state cellular responses with drug administration. To this end, we design a novel computational tool, called KDDANet, which uses known gene functional interaction network to identify hidden genes of cellular pathways mediating KDDA in a genome-wide scale. Our computational pipeline can be applied to two contexts: 1) uncovering hidden genes mediating the association between a query drug and its related disease (SDrTDi); 2) unveiling hidden genes mediating the association between a query disease and its related drug (SDiTDr). Fig 1 demonstrates the computational procedure of KDDANet in SDrTDi context (see Method for details). KDDANet first builds a unified flow model by integrating drugs, genes, and diseases into a heterogeneous network (Fig 1a). Then, the minimum cost flow optimization is designed and implemented to identify gene subnetwork

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mediating the association between query drug (disease) to all its related diseases (drugs) (Fig 1b). Finally, depth first searching, and Markov clustering algorithm are adopted to further uncover gene modules mediating the association between query drug (diseases) and each its related disease (drug) (Fig 1c-1e). To apply KDDANet in SDiTDr context, what the user need is just simply reconstruct the unified flow network model.

The key novelty of the KDDANet method lies in that it uncovers the hidden genes mediating KDDA through effective implementing minimum cost flow optimization, combined with depth first searching and graph clustering on a unified flow network model. We demonstrate that KDDANet outperforms existing methods in both sensitivity and specificity of identifying known and novel genes mediating KDDA. Case studies on Alzheimer's disease (AD) and obesity further show the mechanistic relevance of KDDAnet predictions. Validated with multiple types of cancer-omics' datasets, KDDANet does not only reveal known genes mediating KDDAs associating drug with cancer, but also uncovers new candidates that offer novel biological insights. Particularly, our results demonstrate that KDDANet can reveal the shared genes mediating multiple KDDAs. These show that importance of incorporating hidden genes in drug discovery pipelines. In summary, we develop an effective and universal computational tool for accurate and systematic discovering hidden genes mediating KDDA and thus providing novel insights into mechanism basis of drug repurposing and disease treatments. For facilitating biomedical researchers to explore the molecular mechanism of KDDA and guiding drug repurposing, an online web server, http://www.kddanet.cn, is provided for user to access the subnetwork of genes mediating KDDA, and the source codes of KDDANet are freely available at https://github.com/huayu1111/KDDANet.

Results

Evaluation of the performance and general applicability of KDDANet method

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

When analyzing the overlap ratio between Known Drug Target Genes (KDTGs) and Known Disease Related Genes (KDRGs) of 53124 KDDAs obtained from Comparative Toxicogenomics Database (CTD) (Davis, et al., 2009), we observe that the most gene sets demonstrate extremely low overlap ratio, with the increasing of gene number, the overlap ratio is sharply decreased (Fig S1a). We then check the overlap percentage in different types of diseases and find the overlaps between KDTGs and KDRGs are small for all 19 disease types in our dataset (Fig S1b). The discrepancy between KDTGs and KDRGs indicates that each gene set alone provides only a limited and biased view of KDDA, most of the genes in the cellular pathways mediating KDDA are not effectively identified from the experiments. To address this, we design a novel computational tool, KDDANet, which effectively integrates minimum cost flow optimization, depth first searching and graph clustering to systematically discover the hidden genes of cellular pathways mediating KDDA (see Methods for details).

To examine whether KDDANet can capture true genes mediating KDDA, we introduce two concepts: “known true KDDA genes (KTKGs)” and “novel true KDDA genes (NTKGs)”. For a given KDDA, “KTKGs” is defined as the shared genes between KDTGs and KDRGs feeded for constructing KDDANet flow network model (see Dataset for details). To obtain “NTKGs”, we collected a set of drug’s non-target genes (A) from SMPDB 2.0 database (Jewison et al., 2014) that are included in the drug’s ADME pathways and are responsible for mediating KDDA. Meanwhile, we collected a recently updated set of disease-related genes (B) from DisGeNet v6.0 database (Piñero et al., 2016), which are not included in KDDANet flow network model. Based on these, we define, for each KDDA, the “NTKGs” as the shared genes between A and B. Using gene set enrichment analysis, we observe that KDDANet can effectively capture true genes mediating KDDA (see Supplementary Text S1-S2 and Fig S1c-S1l for details).

Based on this, we compile a standard set of positive and negative KDDA genes for

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

each KDDA to unbiasedly evaluate the capability of KDDANet method on uncovering the true genes mediating KDDA using Receiver operating characteristic (ROC) and precision recall (PR) curves (see Methods for details). We observe that the performance of KDDANet is obviously better than random permutation across a widely setting of γ (Fig2a and Fig2b). We next select γ = 6 for SDrTDi and γ = 8 for SDiTDr for subsequent evaluation (see Supplement Text S2 for reasons). With this setting, we further compare KDDANet with other existing methods, including PPA (Kutalik, et al., 2008), SNPLS (Chen and Zhang, 2016), comCHIPER (Zhao and Li, 2012) and DGPsubNet (Wang, et al., 2014). Among all the methods tested, KDDANet obtains the highest sensitivities and the lowest false-positive rates (Fig 2c). These results suggest that KDDANet is a powerful tool for achieving the goal of uncovering true genes mediating KDDA genome wide. We next test whether KDDANet has the general application value across a variety of disease types and find that KDDANet can make effective capturing of true genes mediating KDDAs for all 19 types of diseases (Fig S2a-S2d). We further test KDDANet using different types of networks, including HINT+HI2012, iRefIndex, MultiNet and STRINGv10. We find that the performances of KDDANet are consistent well across all these networks (Fig 2d and Fig S2e-S2f). Collectively, our results fully demonstrate that KDDANet is a powerful and general computational tool for uncovering hidden genes mediating KDDA across broad types of diseases.

Mechanistic relevance of KDDANet predictions

We examine whether the enriched pathways of KDDANet resulting subnetworks have the mechanistic relevance with KDDA by carrying out a global enrichment of all predicted KDDA subnetwork genes against 53 classical KEGG pathways. The obtained enrichment results can be validated by existing knowledge (see Supplementary Text S4 and Fig S3a-S3b for details). We aim to provide two cases to intuitively describe the mechanistic relevance of KDDANet resulting subnetwork. DB01022 (phylloquinone)-104300 (Alzheimer's disease, AD) association has been reported in

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

previous study (Alisi, et al., 2019). A subnetwork including 46 genes and 44 links are uncovered mediating this association (Fig 3a). The AUROC and AUPRC values of KDDANet for this subnetwork are 0.864 and 0.795, respectively (Fig S3c). Two known targets of phylloquinone and 18 AD-related genes are identified in this subnetwork. Against with genome background, this subnetwork captures 5 NTKGs with ~108-fold enrichment. The top 10 enriched KEGG terms of this subnetwork, such as Phospholipase D signaling pathway and Neurotrophin signaling pathway (Chen et al., 2018; Oliveira and Di Paolo, 2010), are closely related with AD (Fig 3b). Interestingly, for this association, two separated gene modules (M1 and M2) are detected (Fig 3a). The module M1 mainly functions in Insulin signaling pathway, ErbB signaling pathway, FoxO signaling pathway and growth hormone synthesis, secretion and action, which play important roles in neural system development and the onset and development of AD (Buonanno and Fischbach, 2001; Gomez, 2008; Pardeshi et al., 2017) (Fig S3d). The enriched KEGG pathways of M2 genes, including Complement and coagulation cascades, Glycolysis and AGE-RAGE signaling pathway in diabetic complications, are dysfunction in AD (Yan et al., 2009; Vlassenko and Raichle, 2015; Krance et al., 2019) (Fig S3d). We further analyze the published RNA-seq data to detect the expressional change of these two modules in normal individuals and AD patients (Scheckel et al., 2016). We observe that the averaged expression level of M1 genes is significantly upregulated in AD (Fig S3e). Interestingly, a predicted novel gene, STAT3, is obviously activated in AD patients (Fig 3c). Two newest studies published in years 2019 and 2020 report that STAT3 is a potential therapeutic target for cognitive impairment in AD (Reichenbach et al., 2019; Choi et al., 2020). Another predicted novel gene, GNAI1, also obviously activated in AD patients. This gene is contained in the causal pathways associated with an imaging endophenotype characteristic of longitudinal structural change in the brains of patients with AD (Silver et al., 2012). For module M2 genes, the obviously expressional changes between normal individuals and AD patients are not observed (Fig S3e). However, we find that a predicted novel KDDA gene, GC, is activated in AD patients (Fig 3c). This gene encodes Vitamin D Binding

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Protein, which is recently evidenced as a potential therapeutic agent for the treatment of AD (Zhang et al., 2020).

The association between DB01109 (heparin) and 601665 (obesity) is inferred by multiple genes as described in CTD database. For this association, KDDANet predicts a subnetwork containing 168 edges connecting 169 genes (Fig 3d). We find that two known drug target genes and 67 disease-related genes are captured in this subnetwork. Particularly, 9 genes are the NTKGs with ~56-fold enrichment. Three genes used to infer this KDDA, including ARK1, PARP1 and TNF, are also effectively captured in the resulting subnetwork. The top 10 enriched functions of this subnetwork are shown in Fig 3e. As expected, Insulin resistance, Type II diabetes mellitus, Insulin signaling pathway and Adipocytokine signaling pathway are the frequently reported events and molecular processes associated with obesity (Kahn and Flier, 2000; Taylor and Macqueen, 2010). In addition, Proteoglycans, Lipolysis and AMPK signaling pathway are also highly related with Insulin resistance (Langin, et al., 2005; Olsson, et al., 2001; Viollet, et al., 2009). In consistent with this, the AUROC and AUPRC values of KDDANet for this subnetwork are 0.854 and 0.849, respectively (Fig S3e). We next delineated the subnetwork into gene modules. The enriched functions of top 3 gene modules are demonstrated in Fig S3f. The genes of module M1 mainly participate in Glycolysis and Carbon metabolism. Further analysis of public RNA-seq data of normal individuals and obesity patients (Väremo et al., 2017; Väremo et al., 2015) demonstrates the significantly repressed expression of M1 genes in patients with obesity, such as GADPH (Fig S3g and S3h). This is consistent with the fact that enhancing the level of glycolysis reduce obesity (Wu et al., 2005). The mainly enriched pathways of M2 genes are related to Insulin resistance, a frequently happened events in obesity patients. Interestingly, Cell adhesion molecules is a significantly enriched KEGG term of M2 genes (Fig S3f), which is elevated in patients with obesity (Isoppo de Souza et al., 2012). The M3 genes function in Complement and coagulation cascades and Platelet activation (Fig S3f).

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As reported, these two terms are closely associated with obesity (Santilli et al., 2012; Zheng Li., 2017). In support with these, the expression of genes in M2 and M3 are activated in patients with obesity (Fig S3g and S3h). Together, these results indicate that KDDANet can serve as a useful tool to unveil the molecular basis of KDDA.

KDDANet provides novel molecular insights on KDDAs related to cancer

Cancer is a frequently happened complex genetic disease caused by DNA abnormalities (Vogelstein et al., 2013). For this reason, substantial genetic, genomic and pharmacogenomics efforts, including The Cancer Genome Atlas (TCGA), The Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer 1000 human cancer cell lines (GSDC1000), have been undertaken to improve existing therapies or to guide early-phase clinical trials of compounds under development. With these efforts, an increasing amount of available high-throughput data sets at both levels of genomic data and pharmacogenomics data are produced at recent years. In addition, COSMIC, the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer, collects a catalogue of genes with mutations that are causally implicated in cancer. With these datasets, we observe that the genes of KDDANet resulting subnetworks mediating the associations between drugs and cancer are significantly enriched in COSMIC Cancer Gene Census, and these genes harbors more oncogenic alterations in tumor samples than randomly selected genes (Supplementary Text S5, Fig S4a and Fig S4b). Moreover, we find that oncogenic alterations of genes in KDDANet resulting subnetworks mediating the associations between drugs and cancer are more correlated with the responses of cancer cell lines under anti-cancer drug treatment than randomly selected genes (see Supplementary Text S5, Fig S4c-S4f). We provide two detailed examples to describe the potential values of KDDANet in revealing novel genes mediating the associations between drugs and cancer.

Sotalol (DB00489) is normally used to treat life threatening ventricular arrhytmias. It

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

has been reported that sotalol is associated with decreased prostate cancer (176807) risk (Kaapu, et al., 2015). For sotalol-prostate cancer association, KDDANet predicts a subnetwork consisting of 31 genes and 28 links (Fig 4a). All three known target genes of sotalol are included in this subnetwork. Meanwhile, this subnetwork also captures 12 prostate cancer-related genes. Among these genes, ADRB1, ADRB2, PRKACA and PRKAR1A are NTKGs with ~60-fold enrichment. The top 10 enriched KEGG terms are shown in Fig 4b. Among these, PI3K-Akt signaling pathway, MAPK signaling pathway and FoxO signaling pathway are related to cancer formation and development. The relationships between EGFR tyrosine kinase inhibitor resistance, Relaxin signaling pathway, AGE-RAGE signaling pathway and prostate cancer have been widely investigated and reported in the previous works (Bao, et al., 2015; Neschadim, et al., 2015; Ozvegy-Laczka, et al., 2005). In addition, Autophagy and Focal adhesion are two widely observed processes in cancer (Eke and Cordes, 2015; Farrow, et al., 2014). By applying MCL with default parameters, this subnetwork is further decomposed into three gene modules, M1, M2 and M3. As expected, the enriched KEGG signaling pathways of M1 are closely related to tumorigenesis (Fig S4g). The expression of M1 genes are activated in tumor samples, such as an oncogene YWHAE (Fig S4h and S4i). Interestingly, M1 captures that IGF1R, a gene encoding insulin-like growth factor receptor, harbors SNVs and CNVs in TCGA prostate cancer samples, and has been reported to be oncogenic genes of prostate cancer (Heidegger et al., 2014). The genes of M2 are not enrich to any KEGG term and but have lower expression in TCGA tumor samples (Fig S4h). Surprisingly, SPARC, a reported prostate cancer-related gene (Tai and Tang, 2008) with significantly lower expression in TCGA tumor samples is captured in this module (Fig S4i). Moreover, we find that both SPARC and COL1A1 harbor SNVs and CNVs in TCGA prostate cancer samples and their expressions are obviously correlated to the survival of patients (Fig 4c). The M3 genes mainly participate in ErbB signaling pathways, a biological process involved in prostate cancer progression (Brizzolara et al., 2017). The expression of M3 genes are not obviously changed in TCGA prostate cancer samples (Fig S4h). However, we find that

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

GRB2 is over-expressed in TCGA prostate cancer samples and obviously correlated to the survival of patients at the later stage of disease (Fig S4i and Fig 4c).

Another example is the association between nebularine (DB04440)-lung cancer (211980) that is inferred by ADA targeted by nebularine (Gannon et al., 2018). For this association, KDDANet predicts 237 genes connected by 238 links that constitutes a subnetwork without apparent modular structure (Fig 4d). The only known target gene ADA of nebularine is connected to 119 lung cancer-related genes and 117 novel predicted KDDA genes in this subnetwork. As expected, KEGG enrichment demonstrates that the genes in this subnetwork are involved in various cancers (Fig 4e). Interestingly, we find that two highly connected novel genes BYSL and BOP1 are significantly over-expressed in TCGA tumor samples and their expression levels are obviously correlated to the survival of patients (Fig 4f). These results indicate that KDDANet does not only capture known genes mediating KDDAs linking drug with cancer, but also uncovers novel candidates that offer novel biological insights

KDDANet uncovers the shared genes mediating multiple KDDAs

Analysis on above fully demonstrates that KDDANet can uncover true genes mediating individual KDDA. We further ask whether KDDANet can reveal the shared genes mediating multiple KDDAs. We answer this from two aspects as follow: I) Multiple Diseases associating with One Drug (MDiODr); II) Multiple Drugs associating with One Disease (MDrODi). Considering the practical merits for the first one analysis, we require multiple diseases belongs to the same type of diseases. To evaluate the capability of KDDAnet for revealing the shared genes mediating multiple KDDAs, we produce meta-subnetworks by integrating multiple KDDANet resulting subnetworks for 12386 MDiODr combinations and 773 MDrODi combinations produced in SDrTDi context, and 12189 MDiODr combinations and 773 MDrODi combinations in SDiTDr context. The weight of an edge in the meta-subnetwork is defined as the number of KDDA resulting subnetworks containing this edge divided by the total number of

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

KDDA resulting subnetworks. The higher weight value of a link in the meta-subnetwork indicates more conservation and commonality. Thus, we use the weight value to evaluate the capacity of KDDANet in unveiling the shared gene interactions mediating multiple KDDAs. We carry out a permutation test by producing random meta-subnetworks with the same number of edges for comparing with random background. As shown in Fig 5a and Fig S5a-S5b, the weights of KDDANet meta-subnetworks are significantly higher than random meta-subnetworks across different types of diseases. This indicate that the genes tend to be shared in KDDANet meta-subnetworks than random one. We also conduct the same analysis in SDiTDr context and obtain the similar results (Fig 5b and Fig S5c-S5d). Collectively, these results indicate that KDDANet can effectively uncover the shared genes mediating multiple KDDAs.

We present some examples for describing the capability of KDDANet in identifying shared genes mediating multiple KDDAs. For MDiODr, we select two cases: I) DB00392 (profenamine) and neurological disease associations and II) DB00370 (mirtazapine) and cancer associations. As reported in CTD database, profenamine is associated with three neurological diseases, including Parkinsonian Disorders, Multiple Sclerosis and Alzheimer Disease. The shared meta-subnetwork contained 75 edges linking 43 unknown genes with 3 profenamine's target genes and 30 neurological disease-related genes (Fig 5c). The top 10 enriched KEGG terms are demonstrated in Fig 5d. A majority of enriched KEGG terms, such as Cholinergic synapse and Glutamatergic synapses, have been reported dysfunction in neurological disorders and diseases (Moretto, et al., 2018; Tata, et al., 2014). Interestingly, we find that GNAI2 and GNA11 are two mostly shared genes linking profenamine with neurological diseases. These two genes are recently discoveried involving in the pathological pathways of neurological diseases (Zhang et al., 2017; Zhang et al., 2018). Mirtazapine is associated with 9 cancers, including Colorectal Neoplasms, Breast Neoplasms, Neuroblastoma, Glioma, Urinary Bladder Neoplasms, Stomach

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Neoplasms, Esophageal Neoplasms, Lung Neoplasms and Prostatic Neoplasms. Fig S5e shows the shared meta-subnetwork that includes 97 edges connecting 21 mirtazapine's target genes with 33 cancer-related genes and 54 unknown genes. As expected, KEGG enrichment demonstrates that these genes are involved in cancer-related signaling pathways and play important roles in various cancers (Fig S5f). Intriguingly, DRD4 and GRB2 are two mostly shared genes mediating the associations between mirtazapine and cancers (Fig S5e). These two genes are involved in the oncogenesis of multiple types of cancers (Ijaz et al., 2018; Weissenrieder et al., 2019).

For MDrODi, some interesting cases are also observed. For example, GRACILE syndrome, a metabolic disease, is associated with 13 drugs as recorded in CTD database. A shared meta-subnetwork including 66 genes and 61 edges are obtained for this disease (Fig 5e). This meta-subnetwork contains 14 drug target genes and 13 GRACILE syndrome-related genes. The mostly shared gene is ATP5B, and the mostly significant enriched KEGG term is Oxidative phosphorylation (Fig 5f). This is expected as GRACILE syndrome is a fatal inherited disorder caused by a mutation in an oxidative phosphorylation related gene, BCS1L (https://ghr.nlm.nih.gov/condition/gracile-syndrome). It is also not surprising that the neurological diseases related genes are also enriched in this meta-subnetwork as patients with GRACILE syndrome had severe neurological problems (https://www.sciencedirect.com/topics/medicine-and-dentistry/gracile-syndrome). Another example is the Keratoconus (148300), an ophthamological disease, which is associated with three drugs, including acetaminophen, valproic acid and theophylline. Keratoconus and these three drugs share a meta-subnetwork consisting of 54 genes and 51 edges (Fig S5g). This meta-subnetwork includes 7 Keratoconus-related genes, 11 drug target genes and 36 unknown genes. It is expected that HDAC2 has high weights with its partner genes in this meta-subnetwork as it is involved in notch signaling pathway which is downregulated in keratoconus (You, et al., 2018).

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Consistent with the associations between collagen genes and keratoconus (Bykhovskaya, et al., 2016), a novel collagens coding gene, COL1A1, is captured in this meta-subnetwork as a highly shared gene. The enriched KEGG terms of this meta-subnetwork include Hippo signaling pathway (Fig S5h) which has been reported involving in keratoconus corneas (Kabza et al., 2017). Collectively, these results indicate that KDDANet can discover the shared genes mediating multiple KDDAs. The highly shared unknown genes can served as potential candidate targets for drug repurposing.

An online web server for KDDA decoding

To help biomedical researchers to understand the pathogenesis of disease and promote drug development, we develop an online web server, http://www.kddanet.cn, for facilitating the researchers to explore the subnetwork of genes mediating KDDAs. Our website is divided into three parts: I) querying and browsing subnetwork of genes mediating individual KDDA (iKDDA); II) querying and browsing meta-subnetwork of shared genes mediating multiple KDDAs connecting multiple diseases to one drug (MDiODr) and III) querying and browsing meta-subnetwork of shared genes mediating multiple KDDAs connecting multiple drugs to one disease (MDrODi). For Part I, subnetworks of genes mediating 52878 and 51745 KDDAs for SDrTDi and SDiTDr contexts are provided for querying, visualizing, and downloading, respectively. For Part II, 12386 and 12189 shared meta-subnetworks are contained in our web server for SDrTDi and SDiTDr, respectively. For Part III, 773 shared meta-subnetworks are contained in our web server for both contexts. We design a bulk download interface and tutorials for facilitating user to explorer KDDANet resulting subnetworks. Overall, we provide a simply and easy-to-use web server for user to decode the hidden genes mediating KDDAs and thus help researchers to understand the pathogenesis of disease, accelerate drug repurposing and disease treatment. In summary, we present a novel computational tool KDDANet and an online web source to decode molecular

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mechanism of KDDA that has broad utility and application value in biomedical studies.

Discussion

To facilitate drug repurposing, various computational tools have been developed to uncover novel drug-disease associations (Lu and Yu, 2018). However, the molecular mechanism mediating KDDAs have been still not well explored. This results in that unveiling the hidden genes mediating KDDAs become a great challenge for guiding drug repurposing. In this work, we develop a novel computational tool, KDDANet, which integrates minimum cost network flow optimization, depth first searching and graph clustering algorithm to reveal hidden genes and modules mediating KDDA. KDDANet allows for a global and systematic exploration of the hidden genes mediating KDDA. We apply KDDANet to unravel the subnetworks of genes mediating 53124 KDDAs. The comprehensive and system-level evaluations fully demonstrate the powerful prediction capability and general applicability of KDDANet. Case studies on both AD and obesity show that the subnetworks of genes identified by KDDANet are reliable and useful. Further validated by integrating analysis of genomic, transcriptomic, pharmacogenomic and survival data on primary tumors and cancer cell lines highlights that KDDANet captures novel candidates from interactome mediating the associations between drugs and different types of cancer. Based on these, we conclude that the inferred subnetworks mediating KDDA can serve as genome-wide molecular landscapes for guiding drug repurposing and disease treatment. Insights learned from our predictions will also enable to help researchers to design repurposing drugs to reverse disease phenotypes via targeting key genes in the subnetwork mediating KDDA. An important capability of KDDANet method is that it can reveal the shared genes mediating multiple KDDAs. This provides more valuable guides for drug repositioning and disease treatment since the shared genes mediating multiple KDDAs are closely linked to the molecular basis of drug repurposing. We further construct a user-friendly online web server, which allows

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

users to explore the subnetwork of genes mediating individual KDDA and the meta-subnetworks mediating multiple KDDAs.

The hidden genes mediating KDDA predicted by KDDANet highlight the power of integrative approaches to illuminate underexplored molecular processes mediating KDDA. In the future, the application value of our KDDANet tool in drug repurposing can be further improved from three aspects as follow: Firstly, the gene interactome used in KDDANet do not contain enough interactions between genome elements. In further studies, we will integrate other non-coding genome elements, especially long non-coding RNA and microRNA for constructing a comprehensive interactome. Secondly, integrating the biological networks from other omics layers, such as epigenomics, might have also further enhance the accuracy of KDDANet in discovering subnetworks and key genes mediating KDDA, and help us better to understand KDDA at multi-omics levels. The intrinsically ability of KDDANet to analyze large-scale heterogeneous interactome data containing tens of thousands of nodes and edges make it can well be suited to analyzing the accumulating data from 'multi-omics' technologies and biomedical research. Finally, KDDANet cannot carry out predictions for KDDA when a drug's target genes are unknown or when a disease has not been related to any known gene. For this, we plan to calculate the similarity scores between drugs and the similarity scores between diseases, and then to integrate drugs without any target gene and diseases without any related gene to our flow network model by similarity scores.

Methods

1 Datasets

Five different types of gene networks, including HumanNet, HINT+HI2012, iRefIndex, MultiNet and STRINGv10, are used in our current study (Lee et al., 2011; Leiserson et al., 2015), in which the nodes are represented by gene ID and connected by bidirectional edges (Dataset 1). Drugs and their target genes are obtained from

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

DrugBank 5.0.9 database (http://www.drugbank.ca/). In this study, we select 4861 drugs with at least one known target that is contained in the gene networks for further analysis. In total, 2196 Known Drug Target Genes (KDTGs) included in the gene networks are connected to these drugs by 12014 interactions (Dataset 2). Known Disease Related Genes (KDRGs) and classification of diseases are obtained by manually collecting from the previous study (Goh, et al., 2007) and the related literature. We focus on 1441 diseases with at least one related gene which is included in the gene networks for our study. In total, 16712 associations link these diseases to 1521 genes which are exist in the gene networks (Dataset 3). The KDDAs are extracted from Comparative Toxicogenomics Database (Davis, et al., 2009) (Dataset 4). In this study, 53124 KDDAs are analyzed in which the drug has at least one target gene and the disease has at least one related gene contained in HumanNet. For simplicity and consistency, we convert different types of drugs, diseases, and gene nomenclatures to DrugBank drug ID, OMIM disease ID and NCBI Entrez gene ID for subsequent modeling and analysis.

The primary RNA-seq datasets of Alzheimer's disease (AD) patients, obesity patients and normal individuals are downloaded from NCBI GEO Datasets under accession number of GSE53697, GSE81965 and GSE63887. After the SRA files are gathered, the archives are extracted and saved in FASTQ format using the SRA Toolkit. RNA-seq reads are trimmed using Trimmomatic software with the following parameters "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" (Version 0.36) (Bolger et al., 2014), and are further quality-filtered using FASTX Toolkit's fastq_quality_trimmer command (Version 0.0.13) with the minimum quality score 20 and minimum percent of 80% bases that has a quality score larger than this cutoff value. The high-quality reads are mapped to the hg38 genome by HISAT2, a fast and sensitive spliced alignment program for mapping RNA-seq reads, with -dta paramenter (Daehwan et al., 2015). PCR duplicate reads are removed using Picard tools and only uniquely mapped reads are kept for further analysis. The

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression levels of genes are calculated by StringTie (Mihaela et al., 2015) (Version v1.3.4d) with -e -B -G parameters using Release 29 (GRCh38.p12) gene annotations downloaded from GENCODE data portal (https://www.gencodegenes.org/). To obtain comparable expression abundance estimation for each gene, reads mapped to hg38 are counted as FPKM (Fragments Per Kilobase Of Exon Per Million Fragments Mapped) based on their genome locations. Differential expression analysis of genes is performed by DESeq2 using the reads count matrix produced from a python script "prepDE.py" provided in StringTie website (http://ccb.jhu.edu/software/stringtie/).

TCGA cancer genomics datasets are directly downloaded via the UCSC Xena project data portal (https://xenabrowser.net/datapages/). As the DNA methylation datasets are quantified as beta value in the DNA probe level, we map Illumina Human Methylation 450 probe ID to gene name using HumanMethylation450 annotation file. If a gene is mapped by multiple probes, we consider the averaged signals of these probes as the methylation level of this gene. We use Wilcox signed rank test for differential analysis of gene expression and DNA methylation of KDDANet resulting subnetwork genes mediating the associations between drugs and cancer. For this analysis, we only use the tumor samples which have adjacent normal tissue samples as control. Genomics of Drug Sensitivity in Cancer 1000 (GSDC1000) cancer cell line pharmacogenomic datasets are downloaded from GDSC website (https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html). Since this website provides only the CpG island methylation data, we download the beta value matrix of probe-level DNA methylation from NCBI GEO Dataset under accession number of GSE68379 and then convert it to gene-level beta value matrix using the same method as TCGA DNA methylation data. Cancer Cell Line Encyclopedia (CCLE) pharmacogenomic datasets are directly downloaded from Board Institute data portal (https://portals.broadinstitute.org/ccle/data).

2 The computational procedure of KDDANet in SDrTDi context

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.1 Construction of a unified flow network model

For each query drug and all its related diseases, the unified flow network model in SDrTDi context is built by integrating the query drug and all its related diseases into the gene network based on known drug-target relations and gene-disease associations. As shown in Figure 1, for the given query drug, we use it as source node and integrate it into gene network by introducing the directed edges from it point to its target genes. For each its related disease, we map the disease to gene network by introducing the directed edges from its related genes point to it. We incorporate a sink node *T* and introduce a directed edge pointing from each disease to it. With these definitions, a unified flow network model is constructed as a complex heterogeneous graph $G = (V, E)$, where V is the set of vertices and E is the set of edges. This graph includes two types of edges (bidirectional and directed) and three types of nodes (drugs, genes, and diseases). Each edge is assigned with a weight and a capacity. Flow goes from a source node to a sink node through the graph edges. The assigning scheme of weight and capacity is illustrated as follow:

2.2 Weight and capacity assigning scheme for network edges

Edges between gene nodes. Edges between gene nodes are weighted (W_{ij}) to reflect the probability that two genes g_i and g_j are functionally linked in the biological processes. The weight value between g_i and g_j is derived from a Bayesian statistics approach by integrating diverse functional genomics datasets (Lee et al., 2011). Briefly, each experimental dataset is evaluated for its ability to reconstruct known cellular pathways by measuring the likelihood that pairs of genes are functionally linked conditioned on the experimental evidence, calculated as a log likelihood score *LLS*

$$LLS = \ln\left\{\frac{P(L|E)/P(\sim L|E)}{P(L)/P(\sim L)}\right\} \quad (1)$$

Where $P(L|E)$ and $P(\sim L|E)$ are the frequencies of linkages L observed in the given

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

experiment *E* between annotated genes operating in the same pathway and in different pathways, respectively, while *P(L)* and *P(~L)* represent the prior expectation (i.e., the total frequency of linkages between all annotated human genes operating in the same pathway and operating in different pathways, respectively). The weight value (W_{ij}) of a given functional gene link derived from multiple experimental data sets is calculated as the weighted sum (*WS*) of individual *LLS* as:

$$WS = L_0 + \sum_{i=1}^{N} \frac{L_i}{D \cdot i}, \ for\ all\ L \geq T \tag{2}$$

Where L_0 represents the best *LLS* score among all *LLSs* for that gene pair, *D* is a free parameter for the overall degree of independence among the experimental data sets, and *i* is the order index of the data sets after rank-ordering the *N* remaining *LLS* scores for the given gene pair, starting from the second highest *LLS* score and descending in magnitude. The values of two free parameters *D* and *T* are chosen by systematically testing values of *D* and *T* to maximize overall predictive performance of gene function.

Edges between drug and drug's target gene nodes: Edges between each drug and drug's target gene nodes are weighted (w_{Si}) to reflect the reliability of the interaction between drug and target protein based on experimental and computational evidence. The weighting scheme is based on the predicted score of drug and target protein interaction (Yu et al. 2012). The weight value (W_{Si}) is calculated as:

$$W_{Si} = \frac{P_i}{\sum_{j \in T} P_j} \tag{3}$$

Where *T* denotes the set of each drug's targets, P_i denotes the predicted score between drug *S* and target *i*, P_j denotes the predicted score between drug *S* and target *j*.

Edges between disease-related gene and disease nodes. Edges between disease-related gene and disease nodes are weighted (w_{jd}) to reflect the reliability of

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the linkage between disease and gene based on experimental and computational evidence. The weighting scheme is based on the predicted score of gene-disease association derived from MAXIF algorithm (Chen et al., 2011). We calculate the weight value (W_{jd}) as follow:

$$W_{jd} = \frac{F_i}{\sum_{j \in D} F_j} \tag{4}$$

Where D denotes the set of genes linked to disease d, F_i denotes the predicted score between gene i and disease d, P_j denotes the predicted score between gene j and disease d.

Edges between disease and sink nodes. For each edge linking each disease d to sink node T, we assign it a same weight value $W_{dT} = 1/N$, where N denotes the number of diseases linking to sink node.

We further define for each edge in this heterogeneous network a capacity value that limits the flow quantity. For each edge connecting the query drug S to its target gene i, we assigned it a capacity C_{Si} equal to W_{Si}. For each edge linking the disease-related gene j to disease d, we assign it a capacity C_{jd} equal to $W_{jd.}$ For each edge linking the disease d to sink node T, we assign it a capacity C_{dT} equal to $W_{dT.}$ For other edges, we assign them a capacity $C_{ij} = 1$.

2.3 Minimum cost flow optimization algorithm

We aimed for a solution that will (I) pick the subset of the query drug's target genes most likely to modulate all its related diseases by the disease-related genes without limiting it a priori to known mediating genes, (II) identify and rank intermediary genes that are likely to be part of cellular pathways connecting the query drug's target genes to all its related diseases but escaped detection by experiments, (III) give preference to proteins that lie on highest-probability paths connecting the query drug to all its related diseases without imposing constraints on the network topology.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The rationality for proposing this idea including two aspects: I) this need less computational time than that finding the highest probability subnetwork connects a query drug to each its related diseases at a time; II) this can effective find the shared genes mediating multiple KDDAs. We formulate this goal as a minimum cost flow optimization problem (Yeger-Lotem, et al., 2009; Huang, et al., 2011; da Rocha, et al., 2016). Cost was defined as the negative log of the probability of an edge. Hence, minimizing the cost gives preference to highest-probability paths. Given the unified flow network, this problem can be expressed as a linear programming formula that minimize the total cost of the flow network while diffusing the most flow from source node to sink node. This resembles water finding the path of least resistance through a complex landscape. Let W_{ij}, F_{ij} and C_{ij} refer to the weight, flow and capacity from node *i* to node *j*, respectively. The linear programming formula can be written as follow.

Minimize $$\sum_{i\in V, j\in V}\left(-\log\left(W_{ij}\right) * F_{ij}\right) - \left(\gamma * \sum F_{Sv}\right) \quad (5)$$

Subject to $$\sum_{j\in V} F_{ij} - \sum_{j\in V} F_{ji} \quad \forall i \in V - \{S, T\} \quad (6)$$

$$F_{Sv} - \sum_{i\in D} F_{iT} = 0 \quad (7)$$

$$0 \leq F_{ij} \leq C_{ij} \quad (8)$$

Where *V* denotes a set of nodes included in the flow network; *S* denotes the query drug and *v* denotes its target gene. The parameter gamma (γ) controls the size and the quality of the optimized subnetwork. The first component of this formula, $\sum_{i\in V, j\in V}\left(-\log\left(W_{ij}\right) * F_{ij}\right)$, ensure minimizing the network cost that given priority to obtain highest probability gene subnetwork, at the same time, the second component, $-(\gamma * \sum F_{Sv})$, indicates maximizing the total flow across entire network. This optimization problem can be efficiently solved using primal simplex method provided in Mixed Integer Linear Programming (MILP) solver (http://lpsolve.sourceforge.net/). The solution $argmin_{f_{ij}>0} \sum_{i\in V, j\in V}\left(-\log\left(w_{ij}\right) * fij - \gamma * fSv\right.$ get the highest probability subnetwork mediating the associations between query drug and all its related diseases.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.4 Depth first searching and Markov clustering (MCL)

Once the highest probability subnetwork mediating the query drug (disease) with all its related diseases (drugs) is obtained, we implement depth first searching (Asano et al., 2014) on this subgraph to find the subnetwork made up of all paths linking the query drug (disease) to each it's related disease (drug). All genes in the solution are ranked by the amount of flow they carry. The more flow that passes through a protein, the more important it is in mediating KDDA. After obtaining the subnetwork mediating individual KDDA, Markov clustering (MCL) (https://micans.org/mcl/) is employed to further discover gene modules mediating KDDA by using the flow quantities through edges of subnetwork as weight values.

3 Application KDDANet to SDiTDr context

To apply KDDANet in SDiTDr context, the query disease, and a set of all its related drugs are mapped into gene network by disease-related genes and drug target genes. For constructing flow network model, the weights and capacities of network edges can be assigned using the similar method as described in SDrTDi context by substituting query drug with query disease, substituting query drug related diseases with query disease related drugs, substituting query drug's target genes with query disease related genes and substituting disease related genes with drug's target genes. Implementing Minimum cost flow optimization, Depth first searching and Markov clustering (MCL) on the unified flow network is same as SDrTDi context.

4 Performance evaluation

The predictive performance of KDDANet is evaluated as follow: I) Based on the hypothesis that the larger functional similarity between a gene and known KDDA genes, the higher probability this gene is positive one mediating KDDA, we compile a standard set of positive and negative KDDA genes for each KDDA resulting

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

subnetwork to unbiasedly evaluate the performance of KDDANet using the following strategy:

I) For each gene in the KDDA resulting subnetwork, we first calculate the mean functional similarity scores of it with KDTGs and KDRGs by our previous published method using Gene Ontology (GO), KEGG pathways and InterPro annotation as functional terms (Yu et al., 2019).

II) We perform a permutation test by randomly producing KDTGs and KDRGs 1000 times to compute the empirical significance level of functional similarity. We select the genes having significant functional similarities with both KDTGs and KDRGs from KDDA resulting subnetwork as positive KDDA genes using the criterion that the similarity score is larger than 95th percentiles of the simulated background distributions. The other genes are considered as negative KDDA genes.

Based on this gold standard, Receiver operating characteristic (ROC) and precision recall (PR) curves are produced and the areas of under curve of ROC and PR (AUROC and AUPRC) are calculated for gene list ranked by flow amount of each KDDANet resulting subnetwork. For comparison with PPA, SNPLS, comCHIPER and DGPsubNet, we use the default parameters as mentioned in their published papers. We consider that all genes contained in a co-module mediating KDDA of each drug and disease pair in the co-module. The genes are ranked by probability score mediating a KDDA. For PPA and SNPLS, the probability score of a gene *g* mediating the association between drug *i* and a type of cancer *j* is defined as averaged prediction score across all cell lines of this type of cancer (Kutalik, et al., 2008; Chen and Zhang, 2016). For comCHIPER, the probability score of a gene *g* mediating the association between drug *i* and disease *j* is defined as the sum of the products of posterior indicator probabilities across all co-modules (Zhao and Li, 2012). For DGPsubNet, we defined the probability score of a gene *g* mediating the association between drug *i* and disease *j* as the calculated *z-score* z_{ij} (Wang, et al., 2014). The areas of under curve of ROC and PR (AUROC and AUPR) are calculated using R package of PRROC. KDDA gene subnetwork visualization is carried out using Cytoscape software

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(https://cytoscape.org/). KEGG pathway enrichment analysis is performed by clusterProfiler R package (Yu, et al., 2012). The visualization of all results is carried out in R software.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author Contributions

H.Y. and L.L. conceive the original research plans, analyze the data, develop software and write the article; L.L and H.Y develop the web server. H.Y., supervised the experiments; J.Z., C.L., and M.C. provide helpful suggestions and computational support. All authors revise and confirm the paper.

Acknowledgements

This work is supported by Project funded by China Postdoctoral Science Foundation (Grant No. 2018M642441) and High-Performance Computing Platform in Center of Cryo-Electron Microscopy of Zhejiang University.

Figure Legends

Fig 1. Schematic illustration of KDDANet computational pipeline in SDrTDi context. a) Mapping query drugs and all its related diseases into the weighted functional gene interaction network through known drug-target associations and gene-disease associations; b) Constructing a unified flow network model for each query drug and all its related diseases (multiple KDDAs); c) Identifying the highest probability gene subnetwork mediating multiple KDDAs by minimum cost flow optimization; d) Identifying subnetwork mediating individual KDDA by depth first searching; e) Identifying gene interaction modules mediating individual KDDA by Markov clustering (MCL).

Fig 2. Performance evaluation of KDDANet method. a) The density curve of AUROC and AUPRC of KDDANet with different γ setting and permutation test in SDrTDi context; b) The density curve of AUROC and AUPRC of KDDANet with different γ setting and permutation test in SDiTDr context; c) Comparison of AUROC and AUPRC of KDDANet with PPA, SNPLS, comCHIPER and DGPsubNet; d) AUROC and AUPRC of KDDANet with different types of gene interaction networks.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig 3. Mechanistic relevance of KDDANet prediction results. a) KDDANet resulting subnetwork mediating phylloquinone (DB01022)-alzheimer's disease (AD, 104300) association; b) Top 10 enriched KEGG terms of subnetwork genes mediating phylloquinone-AD association; c) Expression level of STAT3, GNAI1 and GC in normal individuals and AD patients, ** *p*-value < 0.01 calculated by Mann-Whitney U test; d) KDDANet resulting subnetwork for heparin (DB01109)-obesity (601665) association; e) Top 10 enriched KEGG terms of subnetwork genes mediating heparin-obesity association; f) ROC and PR curves of KDDANet gene subnetwork mediating heparin-obesity association. In the subnetwork, the size of a gene node is proportional to its network degree; The thickness of a network edge is proportional to its flow amount. Fragments Per Kilobase Of Exon Per Million Fragments Mapped, FPKM.

Fig 4. KDDANet provides novel molecular insights on the associations between drugs and cancer. a) KDDANet resulting subnetwork mediating sotalol (DB00489)-prostate cancer (176807) association; The size of a gene node is proportional to its network degree; The thickness of a network edge is proportional to its flow amount. b) Top 10 enriched KEGG terms of subnetwork genes mediating sotalol-prostate cancer association. c) Gene expression-based survival analysis for CASP2, SPARC and GRB2 in patients with prostate cancer, obtained by GEPIA online analysis (http://gepia.cancer-pku.cn/about.html). d) KDDANet gene subnetwork mediating nebularine (DB04440)-lung cancer (211980) association; e) Top 10 enriched KEGG terms of subnetwork genes mediating nebularine-lung cancer association; f) Gene expression-based survival analysis for BYSL and BOP1 in patients with lung cancer (obtained by GEPIA online analysis (http://gepia.cancer-pku.cn/about.html)) and their expression levels in lung cancer tumor samples and adjacent normal tissue samples, *** *p*-value < 0.001 calculated by Wilcox signed rank test.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig 5. KDDANet uncovers the shared genes mediating multiple KDDAs. a) Boxplot demonstrates the distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDiODr and MDrODi in SDrTDi context. b) Boxplot demonstrates the distributions of weight values of KDDANet meta-subnetwork and random meta-subnetwork for MDiODr and MDrODi in SDiTDr context. c) Shared meta-subnetwork mediating DB00392(profenamine)-neurological diseases associations in SDrTDi context; d) Top 10 enriched KEGG terms of shared meta-subnetwork genes mediating DB00392-neurological diseases associations; e) Shared meta-subnetwork mediating the associations between GRACILE syndrome (603358) and 13 drugs; d) Top 10 enriched KEGG terms of shared meta-subnetwork genes mediating the associations between GRACILE syndrome (603358) and multiple drugs; In the meta-subnetwork, the triangle node denotes disease related genes; the diamond node denotes drug's target genes, the size of a gene node is proportional to its network degree; the thickness of a network edge is proportional to its weight value.

Supplementary Figure Legends

Fig S1. Performance evaluation of KDDANet method. a) Scatter plot demonstrates the distribution of overlap ratio between KDTGs and KDRGs; The x-axis denotes the total number of KDTGs and KDRGs mediating a KDDA. b) Colored scatter plot demonstrates the Minimum (Min), 25th Quantile (Q1), 50th Quantile (Q2), 75th Quantile (Q3) and Maximum (Max) of overlap ratio distribution in different types of diseases; The size and color of point denotes gene number and overlap ratio, respectively; c) Boxplot demonstrates the enrichments of KDTGs in KDDANet resulting subnetwork with different γ settings in SDrTDi context, using random permutation as control (Rd); d) Similar to as c), demonstrating the enrichments of KDRGs; e) Similar to c), demonstrating the enrichments of KDTGs in SDiTDr context; f) Similar to d), demonstrating the enrichments of KDRGs in SDiTDr context. g) Boxplot demonstrates the enrichment of KTKGs in SDrTDi context, using random permutation

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

as control (Rd); h) Similar to g), demonstrating the enrichment of KTKGs in SDiTDr context; i) Similar to g), demonstrating the enrichment of NTKGs; j) Similar to h), demonstrating the enrichment of NTKGs; k) The KDDANet recovery ratio of KDTGs, KDRG and KTKGs with different γ setting in SDrTDi context; l) The KDDANet recovery ratio of KDTGs, KDRGs and KTKGs with different γ setting in SDrTDi context; m) Line chart demonstrates the percentage of KDTGs and KDRGs that are incorporated into the KDDANet resulting subnetwork, as well as the percentage of low probability edges with weight smaller than 0.3 in SDrTDi context (HCI, high confidence interaction; LCI, low confidence interaction); n) Similar to m), demonstrating the percentages in SDiTDr context.

Fig S2. Robustness and general applicability of KDDANet method. a) and b) AUROC and AUPRC of KDDANet across different types of diseases in SDrTDi context; c and d) Similar to a) and b) in SDiTDr context; Values in parentheses denotes the number of KDDAs; e) The KDDANet recovery ratio of KDTGs, KDRGs and KTKGs using different types of networks in SDiTDr context. f) Similar to e) in SDrTDi context.

Fig S3. Mechanistic relevance of KDDANet prediction results. a) Enrichment Score (ES) heatmap of 53 KEGG pathways on 19 different types of diseases in SDrTDi context; b) Similar to a), in SDiTDr context. A two-step procedure is used to calculate ES. First, for each disease type, we calculate the frequency of KDDA resulting subnetworks enriched on 53 KEGG pathways and obtain a frequency matrix. Then, for each pathway, we normalize the frequency matrix to a *z*-score matrix. The value in *z*-score matrix is defined as ES; c) ROC and PR curves of KDDANet resulting subnetwork mediating phylloquinone-alzheimer disease (AD) association; d) Pathway-gene relationship network demonstrating the enriched KEGG terms and their related genes for module M1 and M2 mediating phylloquinone-AD association; e) Averaged expression level of module M1 and M2 genes in normal individuals and AD patients, * *p*-value < 0.05 calculated by Mann-Whitney U test; f) Similar to d)

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pathway-gene relationship network demonstrating the enriched KEGG terms and their related genes for module M1, M2 and M3 mediating heparin-obesity association; g) Averaged expression level of module M1, M2 and M3 genes in normal individuals and obesity patients, ** *p*-value < 0.01, * *p*-value < 0.05 calculated by Mann-Whitney U test; h) Expression level of GADPH, COL1A2, FN1, SPARC, POMC and CTSD in normal individuals and obesity patients; *** *p*-value < 0.001, ** *p*-value < 0.01 calculated by Mann-Whitney U test. Fragments Per Kilobase Of Exon Per Million Fragments Mapped, FPKM. In the pathway-gene relationship network, the size of a KEGG term node is proportional to its *p*-value in enrichment analysis; The color of a gene node denotes its fold change of expression level between normal individuals and patients.

Fig S4. KDDANet provides novel molecular insights on KDDAs linking drug with cancer. a) The ratio of TCGA tumor samples harboring SNVs or CNVs of KDDANet resulting subnetwork genes mediating the associations between drugs and cancer and randomly selected genes, and the overall statistic difference of mRNA expression and DNA methylation of KDDANet resulting subnetwork genes mediating the associations between drugs and cancer (randomly selected genes) between tumor samples and normal samples, respectively; * *p*-value < 0.05 calculated by Mann-Whitney U test. b) The ratio of cancer cell lines harboring SNVs or CNVs of KDDANet resulting subnetwork genes mediating the associations between drugs and cancer and randomly selected genes, * *p*-value < 0.05 calculated by Mann-Whitney U test; c) Boxplots demonstrate the overall statistic difference between IC50 values of anti-cancer drugs on cell lines harboring SNVs or CNVs of KDDANet resulting subnetwork genes mediating the associations between drugs and cancer and cell line without SNVs and CNVs of these genes (random selected genes), *** *p*-value < 0.001, ** *p*-value < 0.01 calculated by Mann-Whitney U test; d) Representative examples demonstrate the difference between IC50 values of anti-cancer drugs on cell lines harboring SNVs or CNVs of KDDANet resulting subnetwork genes mediating the

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associations between drugs and cancer and cell line without SNVs and CNVs of these genes, ** *p*-value < 0.01, * *p*-value < 0.05 calculated by Mann-Whitney U test; e) Scatter plot demonstrates the correlation between expression abundances / methylation levels of KDDANet resulting subnetwork genes mediating the associations between drugs and cancer (randomly selected genes) with half maximal inhibitory concentration (IC50) values of anti-cancer drugs in both GSDC1000 and CCLE cancer cell lines, respectively; PPC, Pearson Correlation Coefficients; *p*-value is calculated by cor.test function in R software. f) Representative examples demonstrate the correlation between the expression abundances / methylation levels of KDDANet resulting subnetwork genes mediating the associations between drugs and cancer and IC50 values of anti-cancer drugs in both GSDC1000 and CCLE cancer cell lines, respectively. g) Pathway-gene relationship network demonstrating the enriched KEGG terms and their related genes for for M1 and M3 mediating sotalol (DB00489)-prostate cancer (176807) association. h) Averaged expression level of module M1, M2 and M3 genes in prostate cancer tumor samples and adjacent normal tissue samples, * *p*-value < 0.05 calculated by Wilcox signed rank test; i) Expression level of YWHAE, SPARC and GRB2 in prostate cancer tumor samples and adjacent normal tissue samples, ** *p*-value < 0.01 calculated by Wilcox signed rank test.

Fig S5. KDDANet uncovers the shared genes mediating multiple KDDAs. a) Boxplot demonstrates the distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDiODr in different types of diseases in SDrTDi context; b) Boxplot demonstrates the distributions of weight values in KDDANet meta-subnetwork and random meta-subnetwork for MDrODi in different types of diseases in SDrTDi context; c) Similar to a) in SDiTDr context; d) Similar to b) in SDiTDr context; e) Shared meta-subnetwork mediating DB00370-cancer associations; f) Top 10 enriched KEGG terms of shared genes of meta-subnetwork mediating DB00370-cancer association; g) Shared meta-subnetwork mediating the associations

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

between Keratoconus (148300) and multiple drugs; h) Top 10 enriched KEGG terms of shared genes of meta-subnetwork mediating the associations between Keratoconus (148300) and multiple drugs.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Alaimo, S., and Pulvirenti, A. (2019) Network-Based Drug Repositioning: Approaches, Resources, and Research Directions. *Methods in Molecular Biology*, **1903**, 97-113.

Alisi, L.*, et al.* (2019) The Relationships Between Vitamin K and Cognition: A Review of Current Evidence, *Frontiers in Neurology*, **10**, 239.

Asano T. et al. (2014) Depth-First Search Using O(n) Bits. *Lecture Notes in Computer Science*, **8889**.

Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: Identifying and developing new uses for existing drugs, *Nature Reviews Drug Discovery*, **3**, 673-683.

Bao, J.M., et al. (2015) AGE/RAGE/Akt pathway contributes to prostate cancer cell proliferation by promoting Rb phosphorylation and degradation, American *Journal of Cancer Research*, 5, 1741-1750.

Bolger, A.M., Marc, L., and Bjoern, U. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114-2120.

Bradley, D. (2005) Why big pharma needs to learn the three 'R's. *Nature reviews Drug discovery*, **4**, 446.

Brizzolara, A., Benelli, R., Venè, R., Barboro, P., Poggi, A., Tosetti, F., and Ferrari, N. (2017) The ErbB family and androgen receptor signaling are targets of Celecoxib in prostate cancer. *Cancer Letters*, **400**, 9-17.

Buonanno, A., and Fischbach, G.D. (2001) Neuregulin and ErbB receptor signaling pathways in the nervous system. *Current Opinion in Neurobiology*, **11**, 287-296.

Bykhovskaya, Y., Margines, B. and Rabinowitz, Y.S. (2016) Genetics in Keratoconus: where are we?, *Eye and Vision*, **3**, 16.

Chen, J. and Zhang, S. (2016) Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data, *Bioinformatics*, **32**, 1724-1732.

Chen, X.Q., Sawa, M., and Mobley, W.C. (2018) Dysregulation of neurotrophin signaling in the pathogenesis of Alzheimer disease and of Alzheimer disease in Down syndrome, *Free Radical Biology & Medicine*, **114**, 52-61.

Chen, Y., Jiang, T., and Jiang, R. (2011). Uncover disease genes by maximizing information flow in the phenome-interactome network. Bioinformatics 27, i167-176.

Cho, A.*, et al.* (2016) MUFFINN: cancer gene discovery via network analysis of somatic mutation data, *Genome Biology*, **17**, 129.

Choi, M., Kim, H., Yang, E.J., and Kim, H.S. (2020) Inhibition of STAT3 phosphorylation attenuates impairments in learning and memory in 5XFAD mice, an animal model of Alzheimer's disease, *Journal of Pharmacological Sciences*, 143, 290-299.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Daehwan, K., Ben, L., and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357-360.

da Rocha, E.L., et al. (2016) NetDecoder: a network biology platform that decodes context-specific biological networks and gene activities, *Nucleic Acids Research*, **44**, e100.

Davis, P.M., Cynthia G*, et al.* (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks, *Nucleic Acids Research*, **37**, D786.

Eke, I. and Cordes, N. (2015) Focal adhesion signaling and therapy resistance in cancer, *Seminars in Cancer Biology*, **31**, 65-75.

Farrow, J.M., Yang, J.C. and Evans, C.P. (2014) Autophagy as a modulator and target in prostate cancer, *Nature Reviews Urology*, **11**, 508-516.

Gannon, H.S., Zou, T., Kiessling, M.K., and Gao, G.F. (2018) Identification of ADAR1 adenosine deaminase dependency in a subset of cancer cells. *Nature Communications*, **9**, 5450.

Ghofrani, H.A., Osterloh, I.H., and Grimminger, F. (2006) Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. *Nature Reviews Drug Discovery*, **5**, 689-702.

Gomez, J.M. (2008) Growth hormone and insulin-like growth factor-I as an endocrine axis in Alzheimer's disease. *Endocrine, Metabolic & Immune Disorders Drug Targets*, **8**, 143-151.

Goh, K.I.*, et al.* (2007) The human disease network, *Proceedings of the National Academy of Sciences of the United States of America,* **104**, 8685-8690.

Heidegger, I., Kern, J., Ofer, P., Klocker, H., and Massoner, P. (2014) Oncogenic functions of IGF1R and INSR in prostate cancer include enhanced tumor growth, cell migration and angiogenesis. *Oncotarget*, **5**, 2723-2735.

Huang, J.*, et al.* (2011) eResponseNet: a package prioritizing candidate disease genes through cellular pathways, *Bioinformatics*, **27**, 2319-2320.

Ijaz, M., Wang, F., Shahbaz, M., Jiang, W., Fathy, A.H., and Nesa, E.U. (2018) The Role of Grb2 in Cancer and Peptides as Grb2 Antagonists. *Protein and Peptide Letters*, **24**, 1084-1095.

Isoppo de Souza, C., Rosa, D.D., Ettrich, B., Cibeira, G.H., Giacomazzi, J., Tusset, P., Ashton-Prolla, P., Medeiros, L.R., Caleffi, M., Neto, E.C., et al. (2012) Association of adipokines and adhesion molecules with indicators of obesity in women undergoing mammography screening. *Nutrition & Metabolism*, **9**, 97.

Kaapu, K.J.*, et al.* (2015) Sotalol, but not digoxin is associated with decreased prostate cancer risk: A population-based case-control study, *International Journal of*

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cancer, **137**, 1187-1195.

Krance, S.H., Wu, C. Y., Zou, Y., Mao, H., Toufighi, S., He, X., Pakosh, M., and Swardfager, W. (2019). The complement cascade in Alzheimer's disease: a systematic review and meta-analysis, *Molecular Psychiatry*.

Kabza, M., Karolak, J.A., Rydzanicz, M., Szcześniak, M.W., Nowak, D.M., and Ginter-Matuszewska, B. (2017) Collagen synthesis disruption and downregulation of core elements of TGF-β, Hippo, and Wnt pathways in keratoconus corneas. *European Journal of Human Genetics*, **25**, 582-590.

Kahn, B.B. and Flier, J.S. (2000) Obesity and insulin resistance, *The Journal of Clinical Investigation*, **106**, 473-481.

Kutalik, Z., Beckmann, J. and Bergmann, S. (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data, *Nature Biotechnology.*, **26**, 531.

Langin, D., et al. (2005) Adipocyte lipases and defect of lipolysis in human obesity, *Diabetes*, **54**, 3190-3197.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome research 21, 1109-1121.

Leiserson, M., Vandin, F., Wu, H. et al. (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, **47**, 106–114.

Lotfi Shahreza, M., Ghadiri, N., Mousavi, S.R., Varshosaz, J., and Green, J.R. (2018) A review of network-based approaches to drug repositioning, *Briefings in Bioinformatics*, **19**, 878-892.

Lu, L. and Yu, H. (2018) DR2DI: a powerful computational tool for predicting novel drug-disease associations, *Journal of Computer-Aided Molecular Design*, **32**, 633-642.

Moretto, E., et al. (2018) Glutamatergic synapses in neurodevelopmental disorders, *Progress in Neuropsychopharmacology & Biological Psychiatry*, **84**, 328-342.

Moridi, M., Ghadirinia, M., Sharifi-Zarchi, A., and Zare-Mirakabad, F. (2019). The assessment of efficient representation of drug features using deep learning for drug repositioning. *BMC bioinformatics*, **20**, 577.

Mihaela, P., Pertea, G.M., Antonescu, C.M., Tsung-Cheng, C., Mendell, J.T., and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**, 290-295.

Neschadim, A., Summerlee, A.J. and Silvertown, J.D. (2015) Targeting the relaxin hormonal pathway in prostate cancer, *International Journal of Cancer*, **137**, 2287-2295.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ogretmen, B. (2018) Sphingolipid metabolism in cancer signalling and therapy, *Nature Reviews Cancer*, **18**, 33-50.

Oliveira, T.G., and Di Paolo, G. (2010) Phospholipase D in brain function and Alzheimer's disease. *Biochimica et Biophysica Acta*, **1801**, 799-805.

Olsson, U., et al. (2001) Changes in matrix proteoglycans induced by insulin and fatty acids in hepatic cells may contribute to dyslipidemia of insulin resistance, *Diabetes*, **50**, 2126-2132.

Ozvegy-Laczka, C., et al. (2005) Tyrosine kinase inhibitor resistance in cancer: role of ABC multidrug transporters, *Drug Resistance Updates*, **8**, 15-26.

Pardeshi, R., Bolshette, N., Gadhave, K., Ahire, A., Ahmed, S., Cassano, T., Gupta, V.B., and Lahkar, M. (2017) Insulin signaling: An opportunistic target to minify the risk of Alzheimer's disease. *Psychoneuroendocrinology*, **83**, 159-171.

Paul, S.M.*, et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge, *Nature Review Drug Discovery*, **9**, 203-214.

Reichenbach, N., Delekate, A., Plescher, M., Schmitt, F., Krauss, S., Blank, N., Halle, A., and Petzold, G.C. (2019) Inhibition of Stat3-mediated astrogliosis ameliorates pathology in an Alzheimer's disease model, *EMBO Molecular Medicine*, **11**, e6995.

Santilli, F., Vazzana, N., Liani, R., Guagnano, M.T., and Davi, G. (2012) Platelet activation in obesity and metabolic syndrome. *Obesity Reviews*, **13**, 27-42.

Scheckel, C., Drapeau, E., Frias, M.A., Park, C.Y., Fak, J., Zucker-Scharff, I., Kou, Y., Haroutunian, V., Ma'ayan, A., Buxbaum, J.D., et al. (2016) Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain, *eLife*, **5**, e10421.

Silver, M., Janousova, E., Hua, X., Thompson, P.M., and Montana, G. (2012) Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. *NeuroImage*, **63**, 1681-1694.

Tai, I.T., and Tang, M.J. (2008) SPARC in cancer biology: its role in cancer progression and potential for therapy. *Drug Resistance Updates*, **11**, 231-246.

Taylor, V.H. and Macqueen, G.M. (2010) The Role of Adipokines in Understanding the Associations between Obesity and Depression, *Journal of Obesity*, **748048**.

Tata, A.M., et al. (2014) Cholinergic system dysfunction and neurodegenerative diseases: cause or effect?, *CNS & Neurological Disorders-Drug Targets*, **13**, 1294-1303.

Väremo, L., Henriksen, T.I., Scheele, C., Broholm, C., Pedersen, M., Uhlén, M., Pedersen, B.K., and Nielsen, J. (2017) Type 2 diabetes and obesity induce similar transcriptional reprogramming in human myocytes. *Genome Medicine*, **9**, 47.

Väremo, L., Scheele, C., Broholm, C., Mardinoglu, A., Kampf, C., Asplund, A., Nookaew,

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

I., Uhlén, M., Pedersen, B.K., and Nielsen, J. (2015) Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. *Cell Reports*, **11**, 921-933.

Viollet, B., et al. (2009) Targeting the AMPK pathway for the treatment of Type 2 diabetes, *Frontiers in Bioscience*, **14**, 3380-3400.

Vlassenko, A.G., and Raichle, M.E. (2015) Brain aerobic glycolysis functions and Alzheimer's disease. *Clinical and Translational Imaging*, **3**, 27-37.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546-1558.

Wang, L.*, et al.* (2014) Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks, *CPT Pharmacometrics System Pharmacology*, **3**, e146

Weissenrieder, J.S., Neighbors, J.D., Mailman, R.B., and Hohl, R.J. (2019) Cancer and the Dopamine D(2) Receptor: A Pharmacological Perspective. *The Journal of Pharmacology and Experimental Therapeutics*, **370**, 111-126.

Weng, L., Zhang, L., Peng, Y., and Huang, R.S. (2013) Pharmacogenetics and pharmacogenomics: a bridge to individualized cancer therapy. *Pharmacogenomics*, **14**, 315-324.

Yan, S.D., Bierhaus, A., Nawroth, P.P., and Stern, D.M. (2009) RAGE and Alzheimer's disease: a progression factor for amyloid-beta-induced cellular perturbation? *Journal of Alzheimer's Disease*, **16**, 833-843.

Wu, C., Kang, J.E., Peng, L.J., Li, H., Khan, S.A., Hillard, C.J., Okar, D.A., and Lange, A.J. (2005). Enhancing hepatic glycolysis reduces obesity: differential effects on lipogenesis depend on site of glycolytic modulation. *Cell Metabolism*, **2**, 131-140.

Xuan, P., Cao, Y., Zhang, T., Wang, X., Pan, S., and Shen, T. (2019). Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics*, **35**, 4108-4119.

Yeger-Lotem, E., et al. (2009) Bridging the gap between high-throughput genetic and transcriptional data reveals cellular pathways responding to alpha-synuclein toxicity, *Nature Genetics*, **41**, 316.

Yildirim, M.A.*, et al.* (2007) Drug-target network, *Nature Biotechnology*, **25**, 1119-1126.

You, J., et al. (2018) RNA-Seq analysis and comparison of corneal epithelium in keratoconus and myopia patients, *Scientific Reports*, **8**, 389.

Yu, G., et al. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters, Omics: A Journal of Integrative Biology, **16**, 284-287.

Yu, H., Chen, J., Xu, X., Li, Y., Zhao, H., Fang, Y., Li, X., Zhou, W., Wang, W., and Wang, Y.

bioRxiv preprint doi: https://doi.org/10.1101/749762; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(2012) A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS ONE*, **7**, e37608.

Yu, H., Lu, L., Jiao, B., and Liang, C. (2019) Systematic discovery of novel and valuable plant gene modules by large-scale RNA-seq samples. *Bioinformatics*, **35**, 361-364.

Zhang, H., Wang, T., Han, Z., Wang, L., Zhang, Y., Wang, L., and Liu, G. (2020). Impact of Vitamin D Binding Protein Levels on Alzheimer's Disease: A Mendelian Randomization Study. *Journal of Alzheimer's Disease*, **74**, 991-998.

Zhang, M., Mu, H., Shang, Z., Kang, K., Lv, H., Duan, L., Li, J., Chen, X., Teng, Y., Jiang, Y., et al. (2017). Genome-wide pathway-based association analysis identifies risk pathways associated with Parkinson's disease. *Neuroscience*, **340**, 398-410.

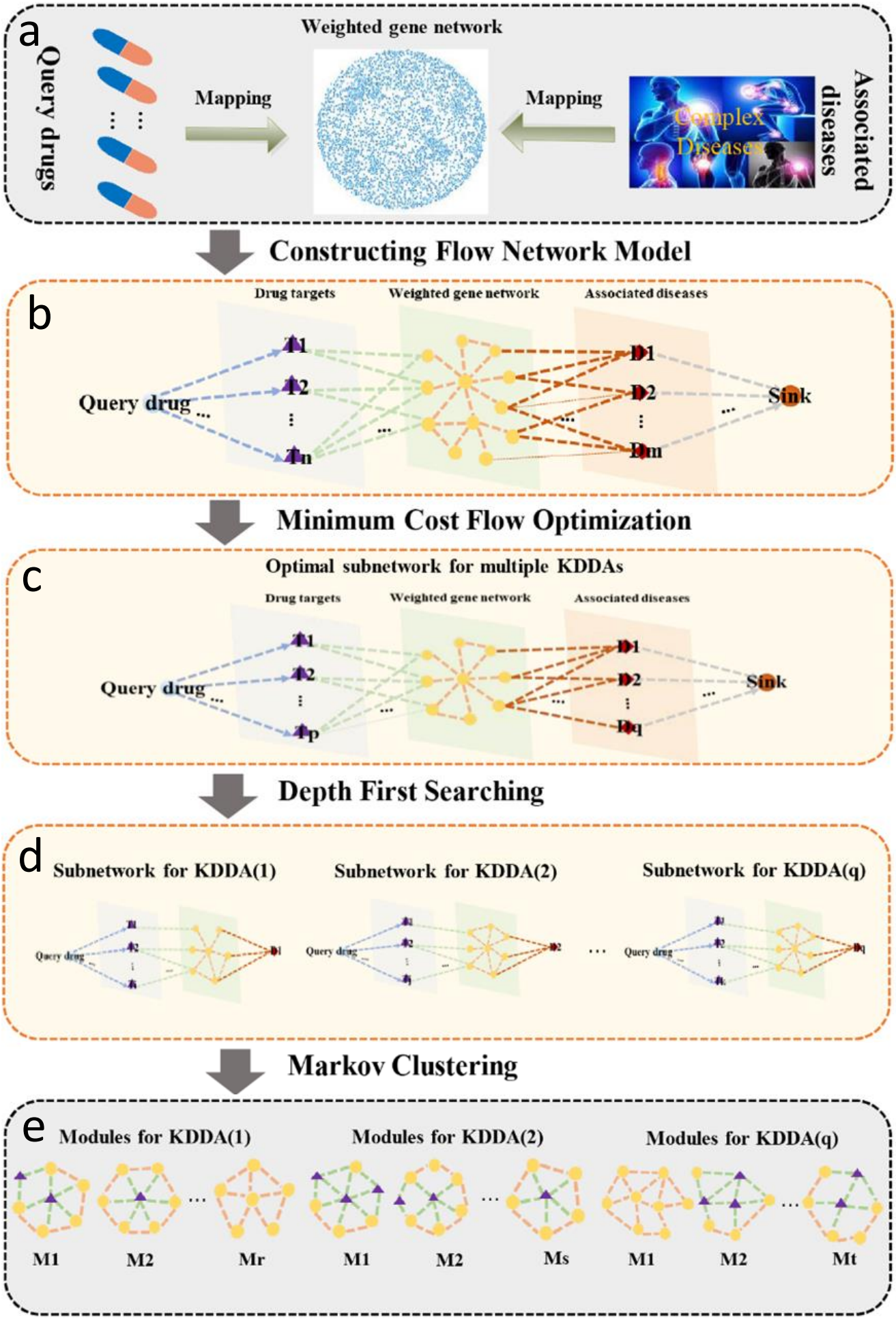
Zhang, Q., Ma, C., Gearing, M., Wang, P.G., Chin, L.S., and Li, L. (2018) Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer's disease. *Acta Neuropathologica Communications*, **6**, 19.

Zhao, S. and Li, S. (2012) A co-module approach for elucidating drug-disease associations and revealing their molecular basis, *Bioinformatics*, **28**, 955-961.

Zheng L. (2017) Plasminogen: a potential target gene for dietary supplements and biomarker of the early stage of obesity by fatigue mice, *Biomedical Research*, **28**, 4299-4304.

not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

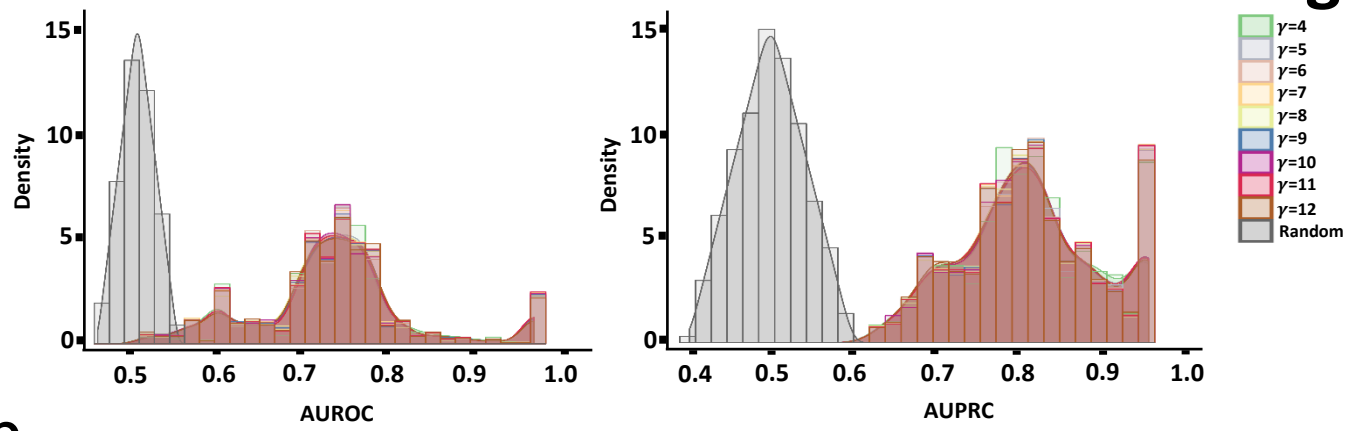
Fig 1



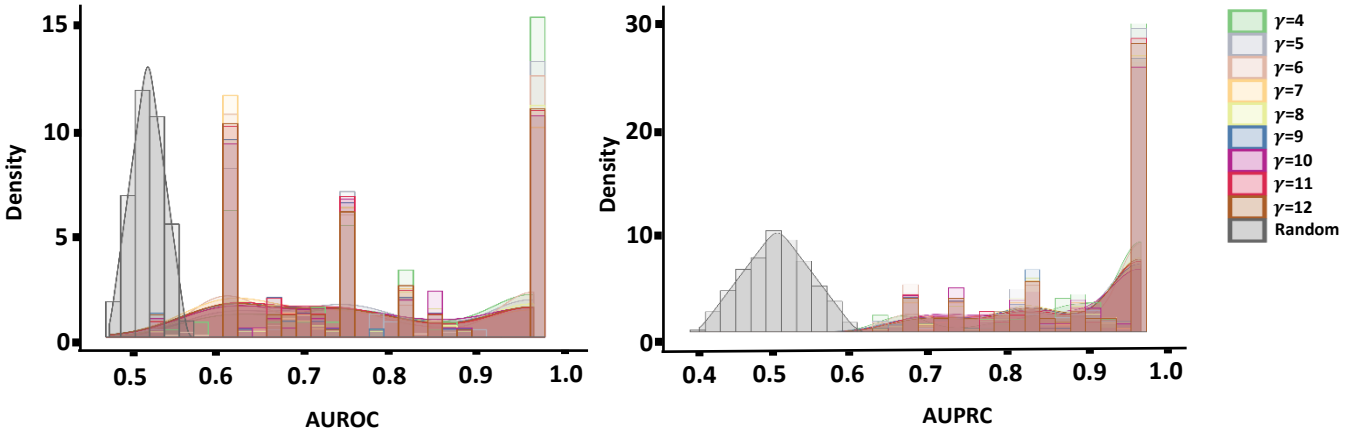
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available
under aCC-BY-NC-ND 4.0 International license.

Fig 2

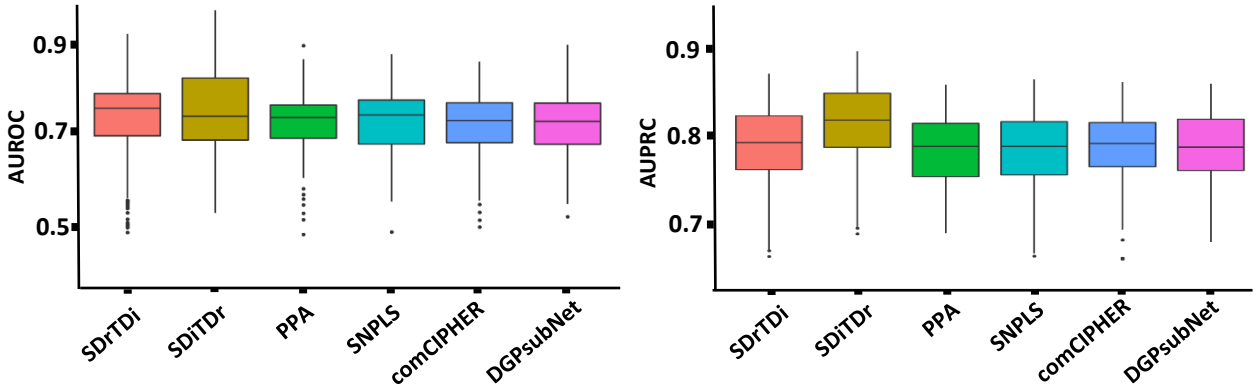
a



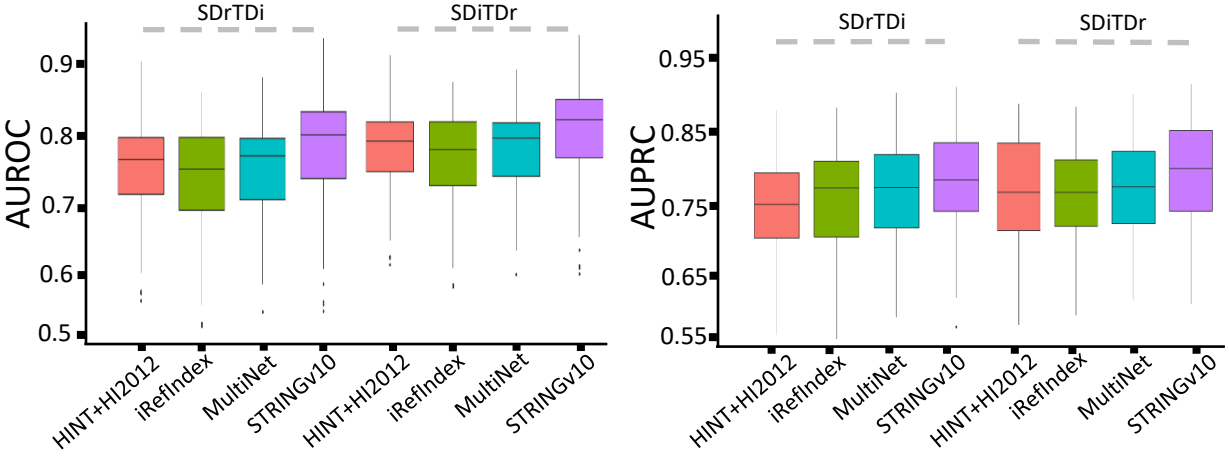
b



c

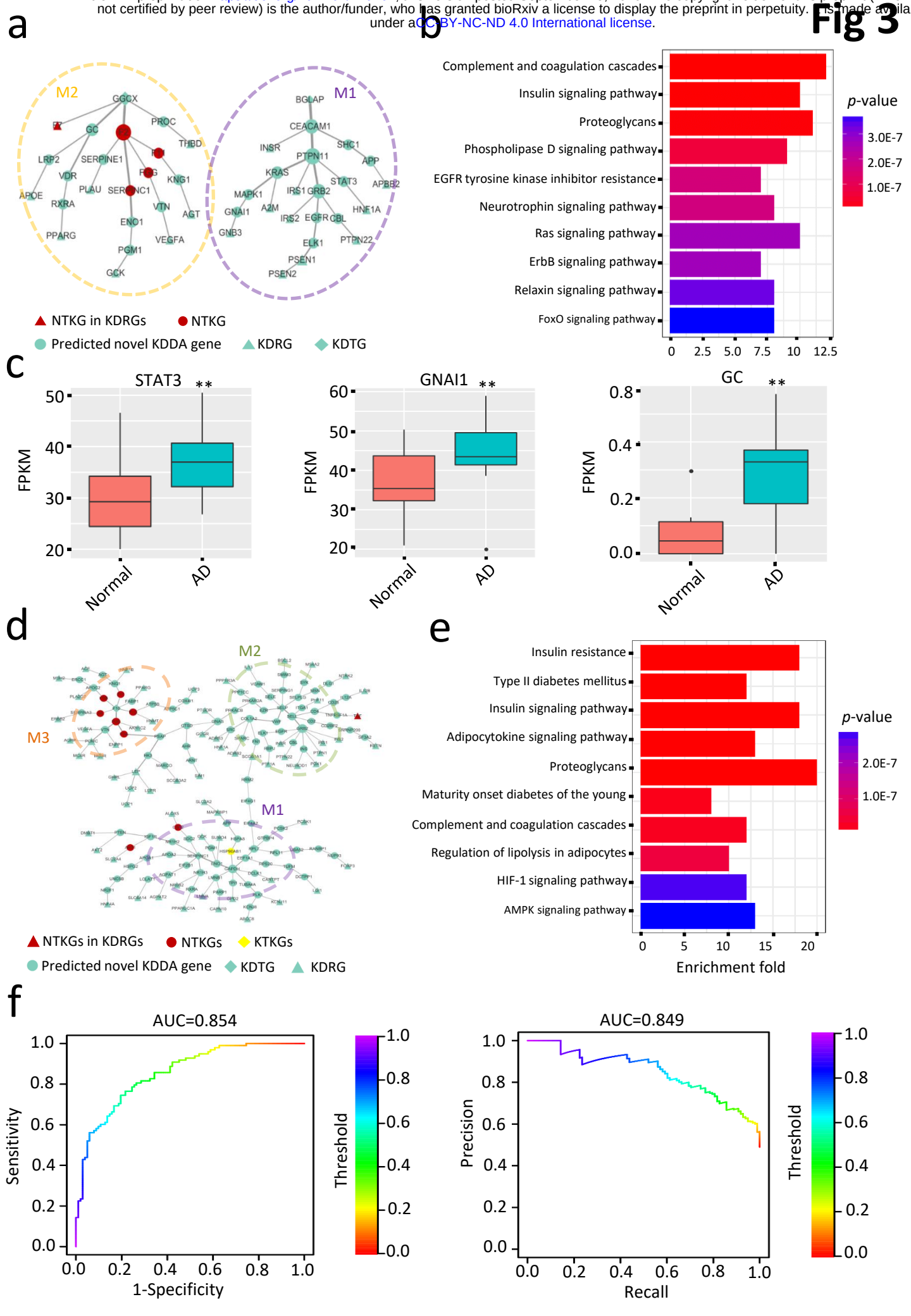


d



not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available
under a CC-BY-NC-ND 4.0 International license.

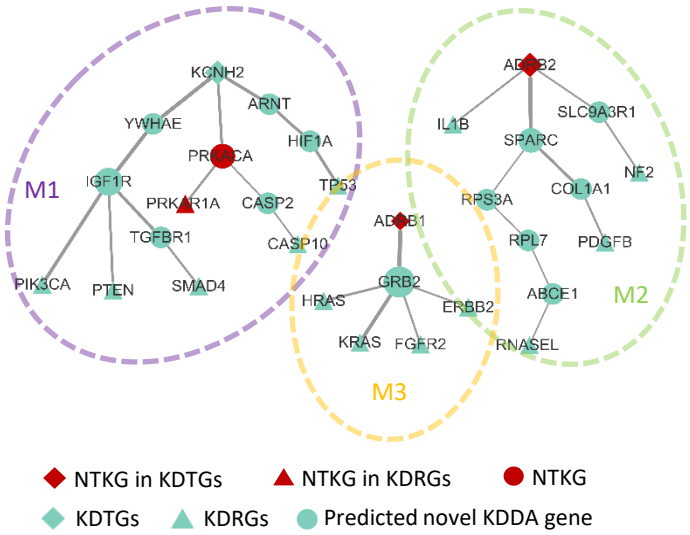
Fig 3



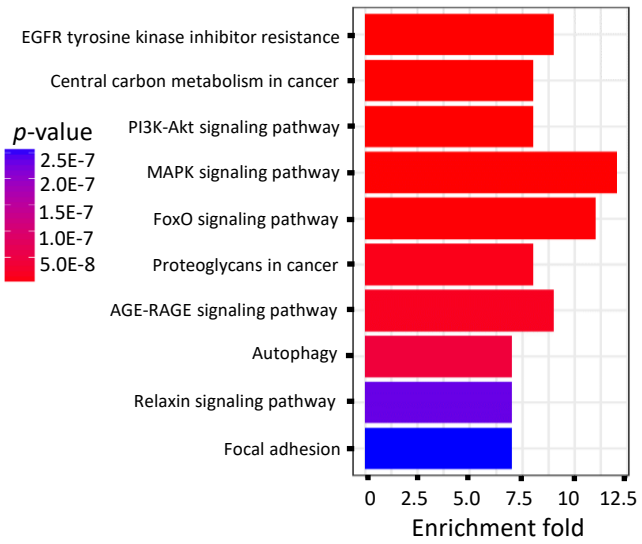
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available
under aCC-BY-NC-ND 4.0 International license.

Fig 4

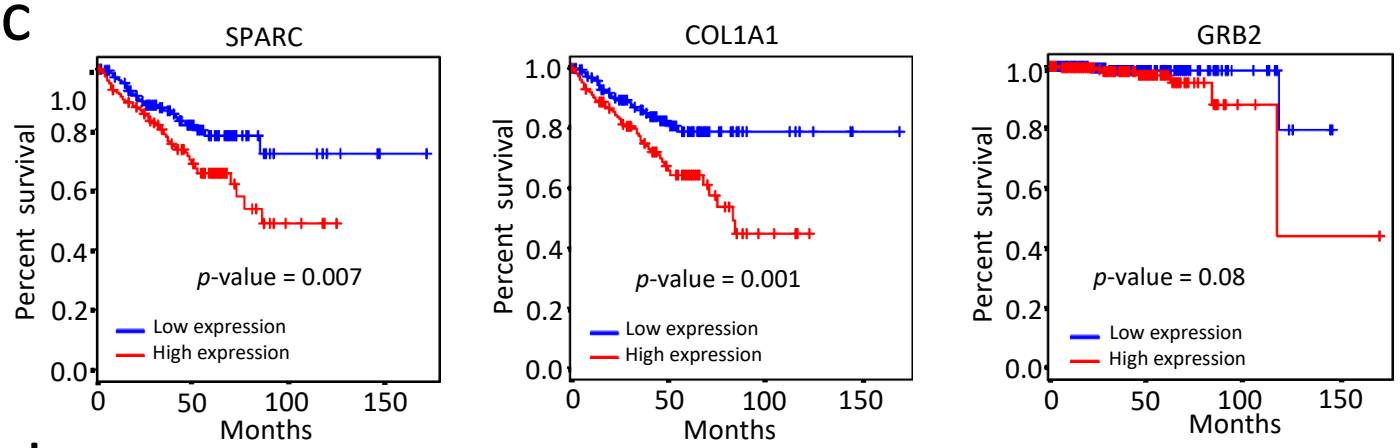
a



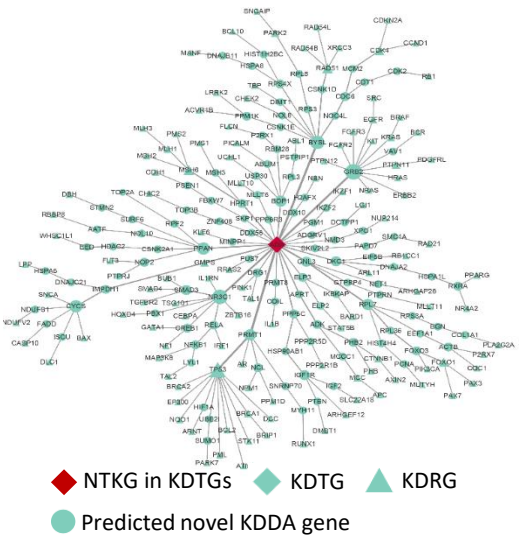
b



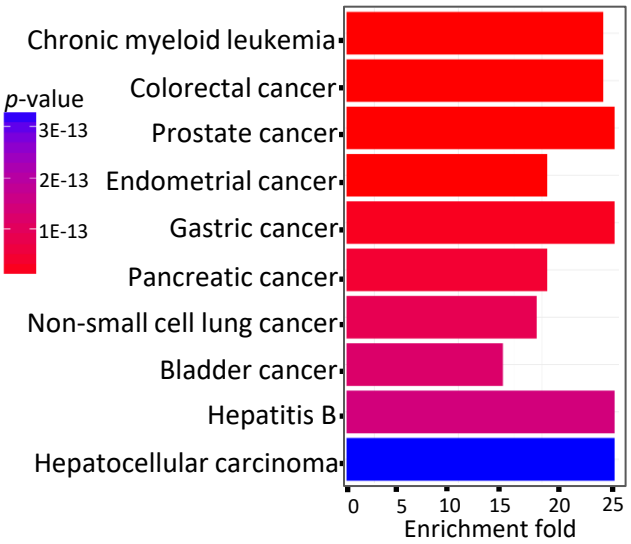
c



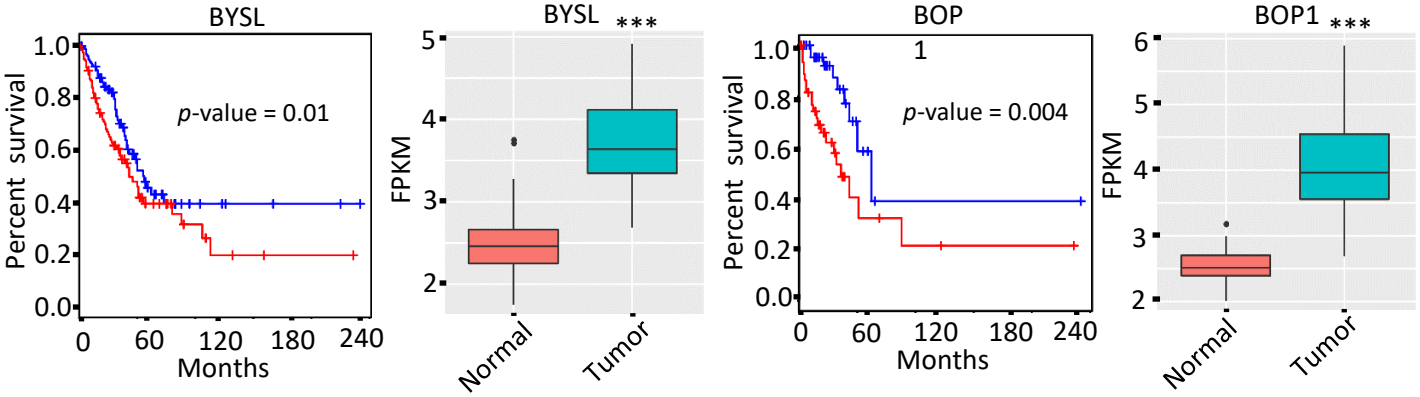
d



e

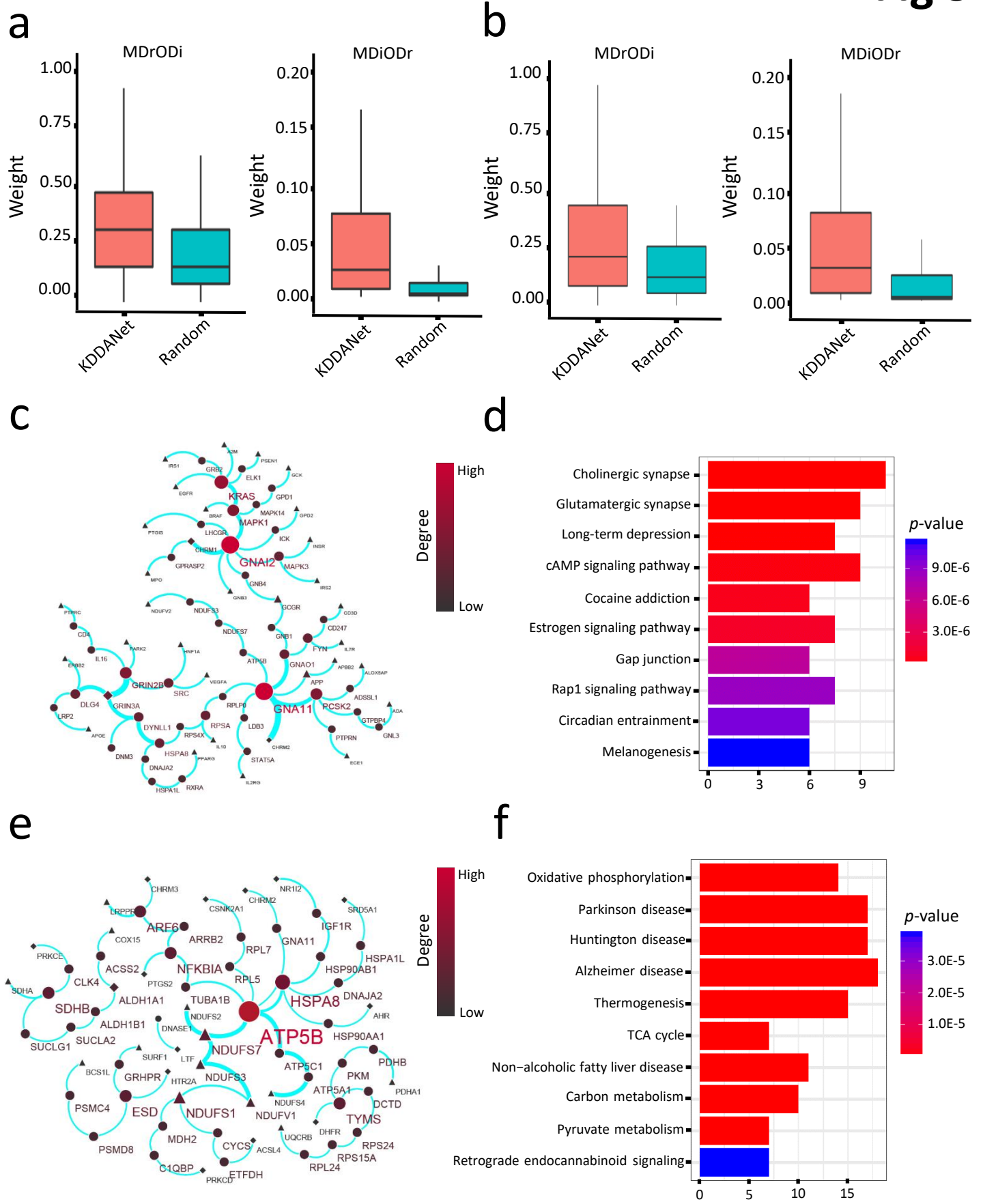


f



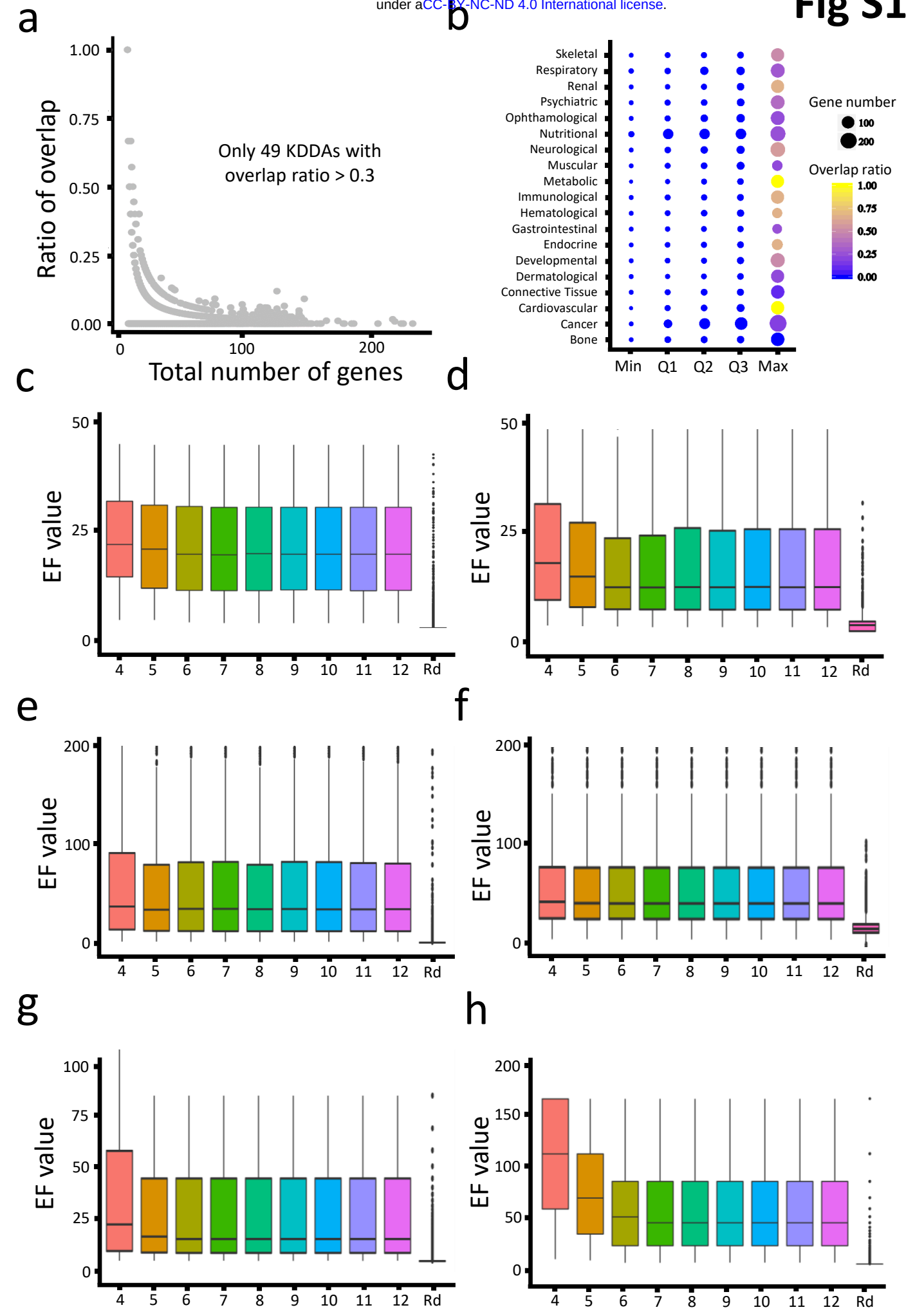
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

Fig 5



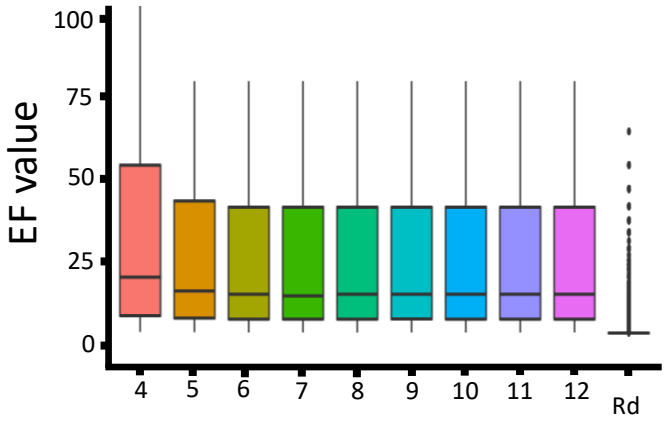
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

Fig S1

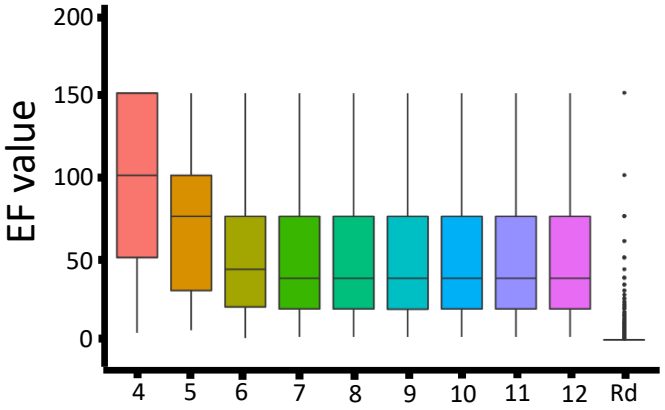


not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

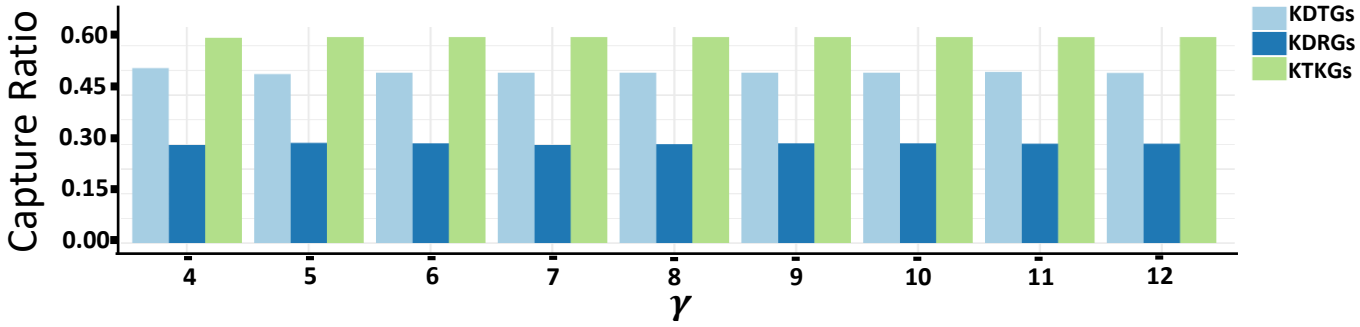
i



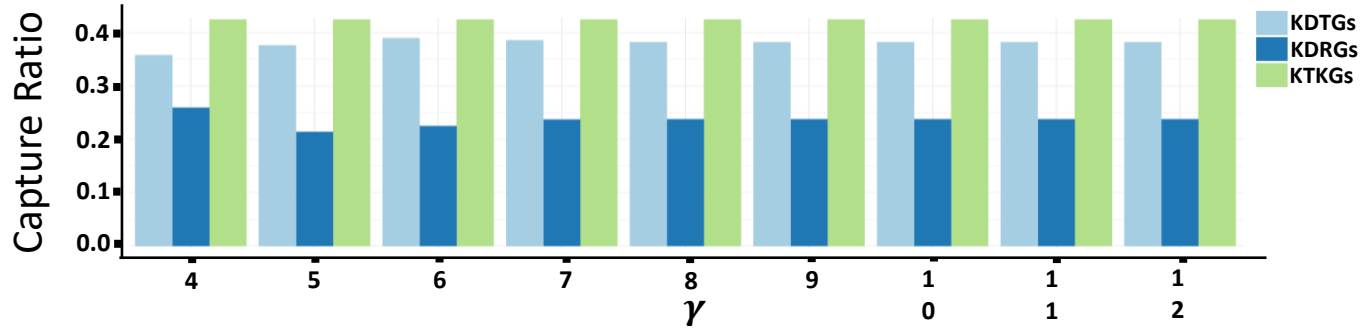
j



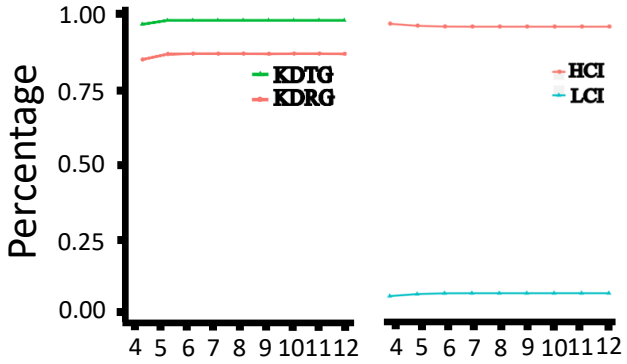
k



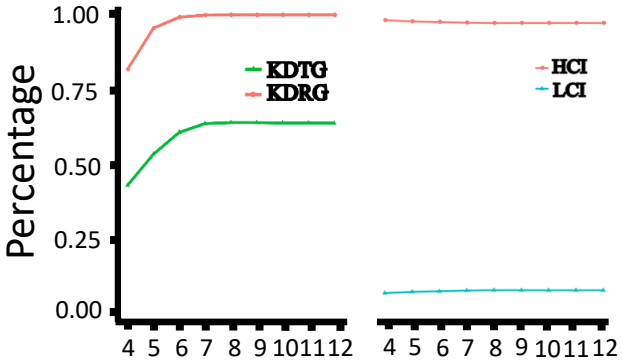
l



m

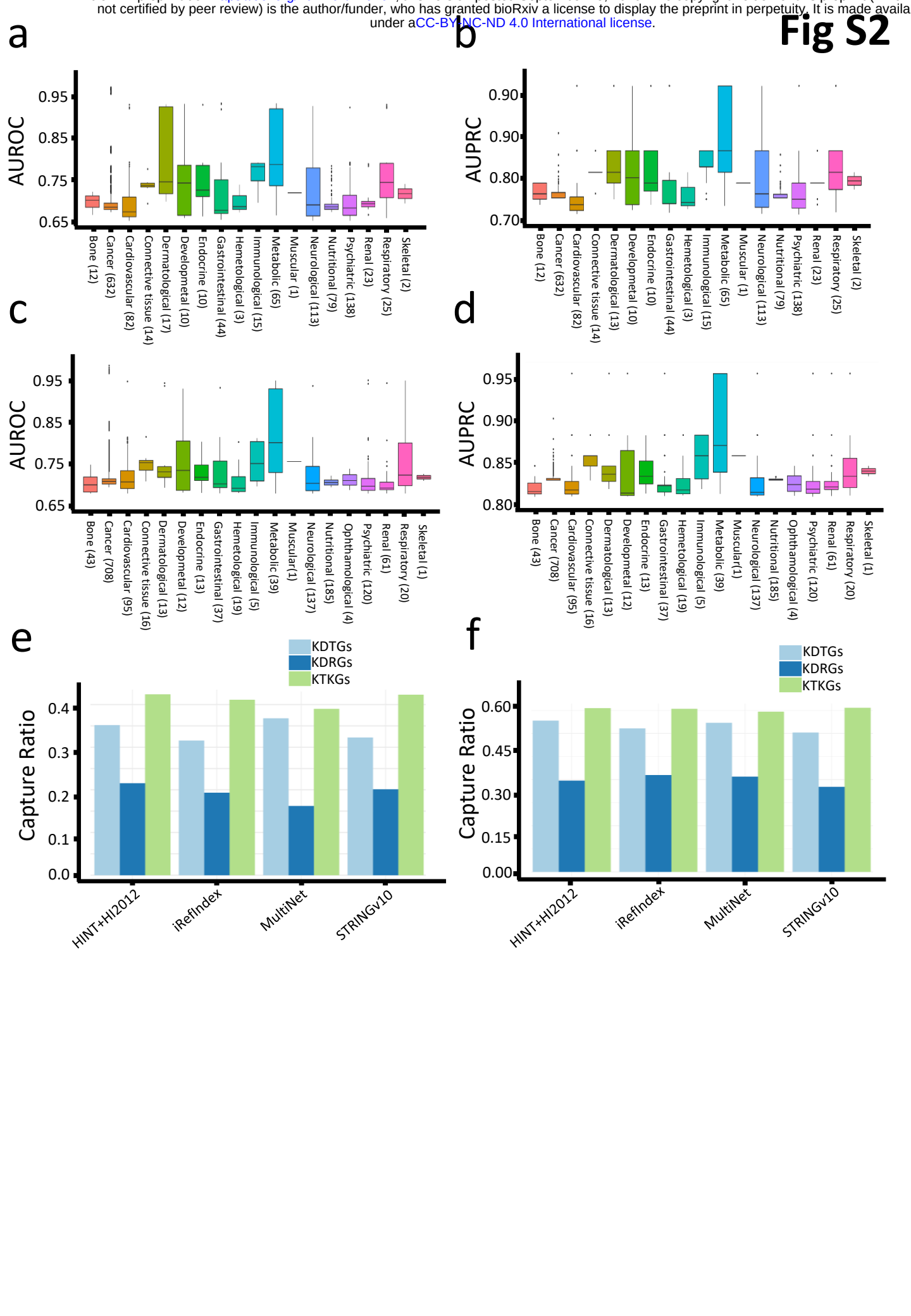


n



not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

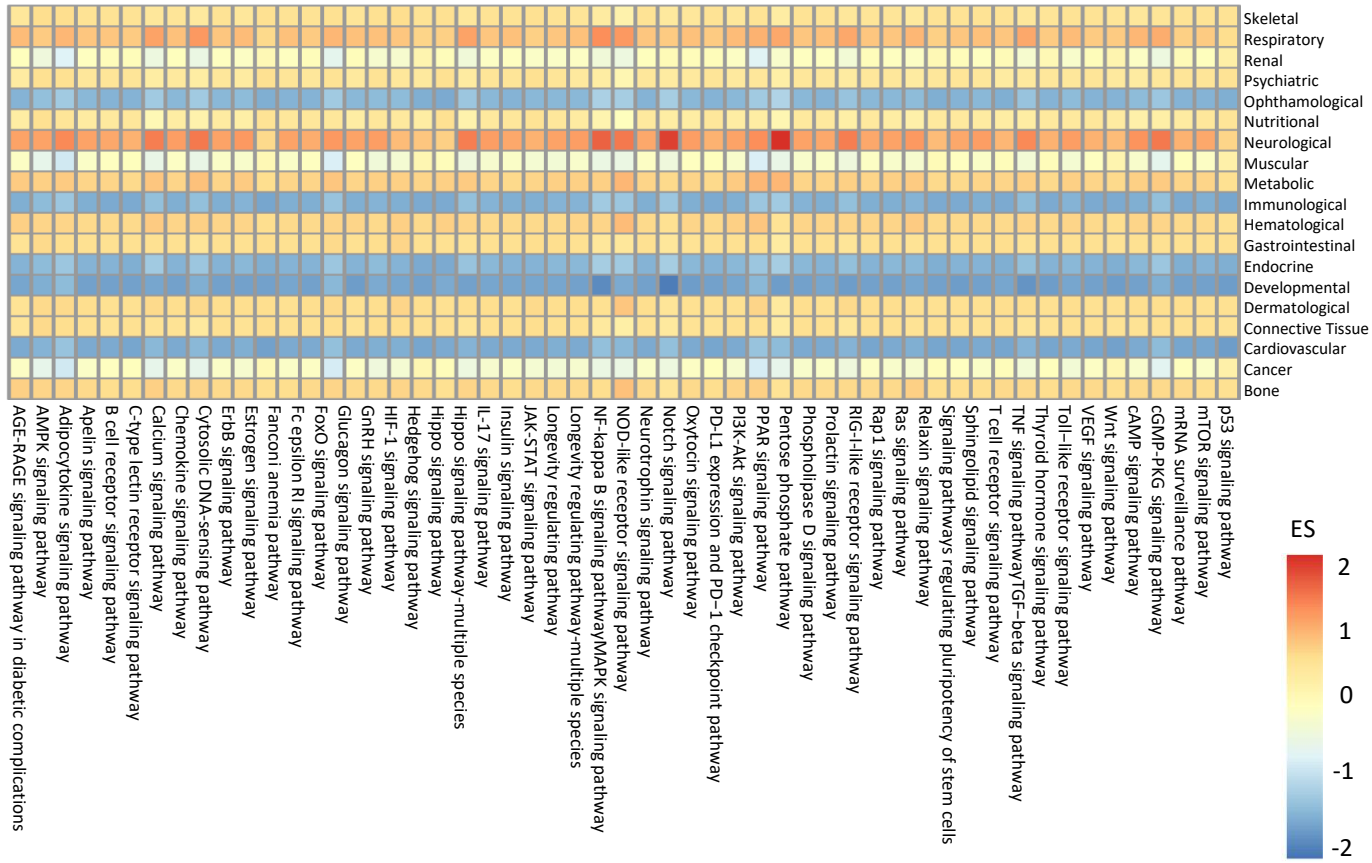
Fig S2



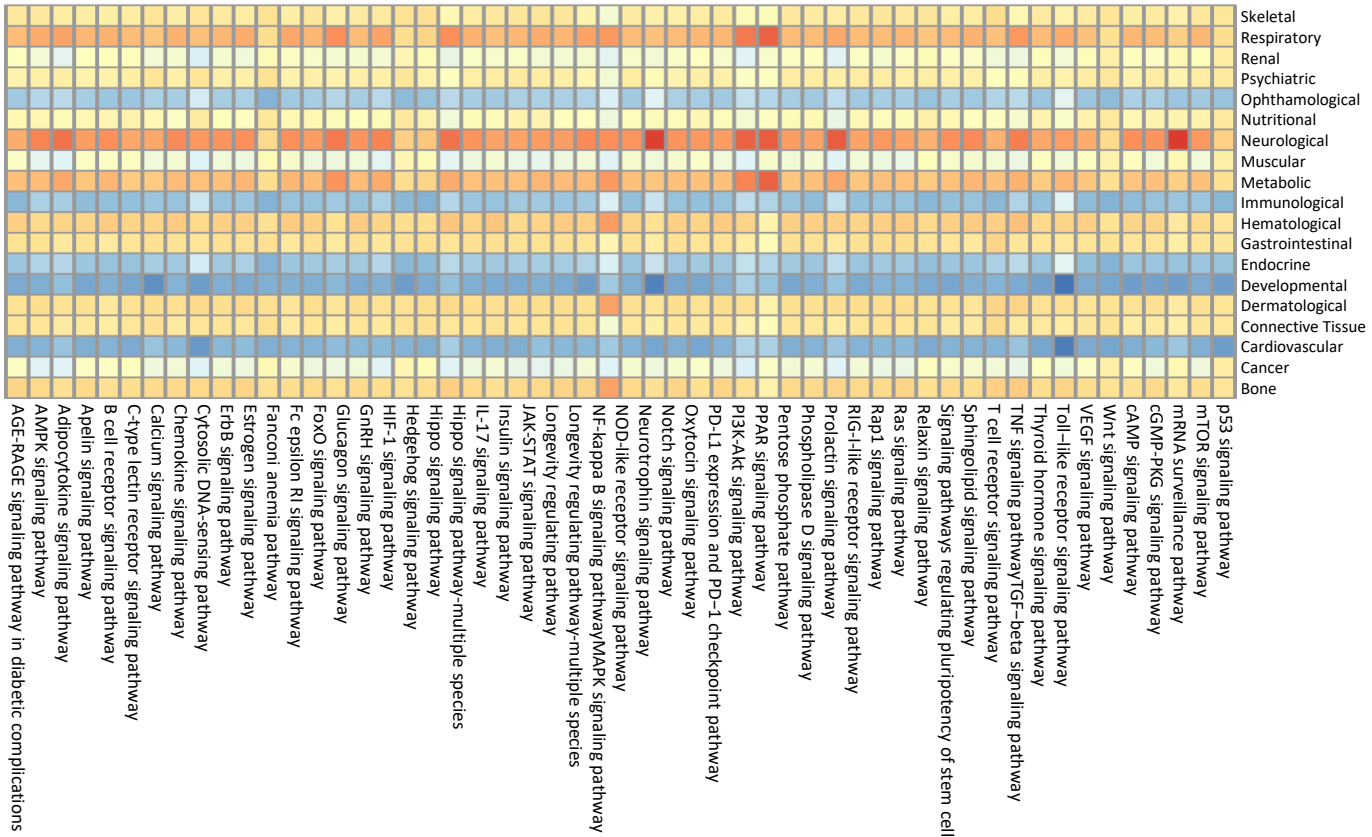
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

Fig S3

a

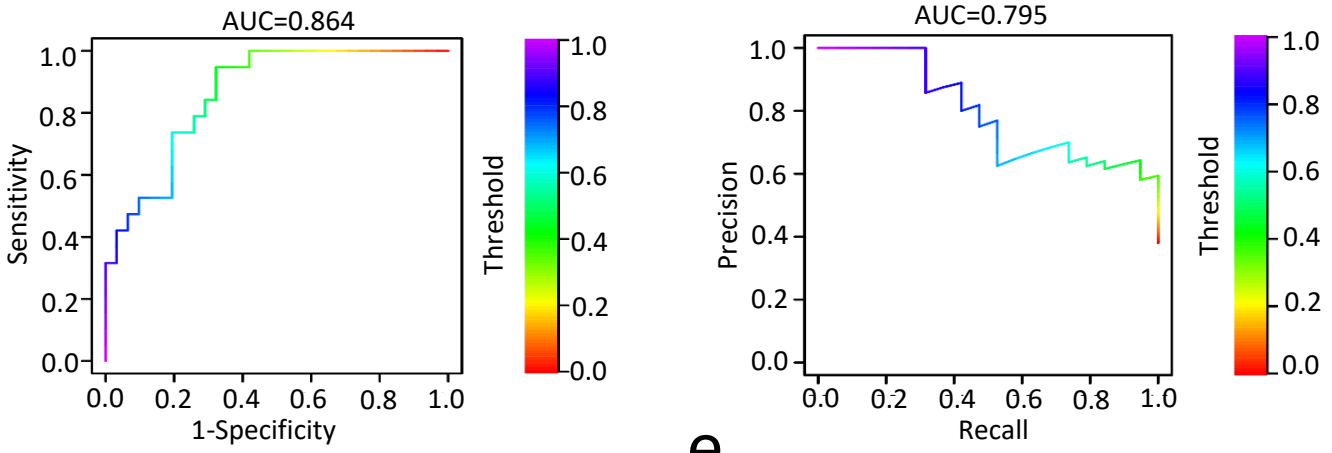


b

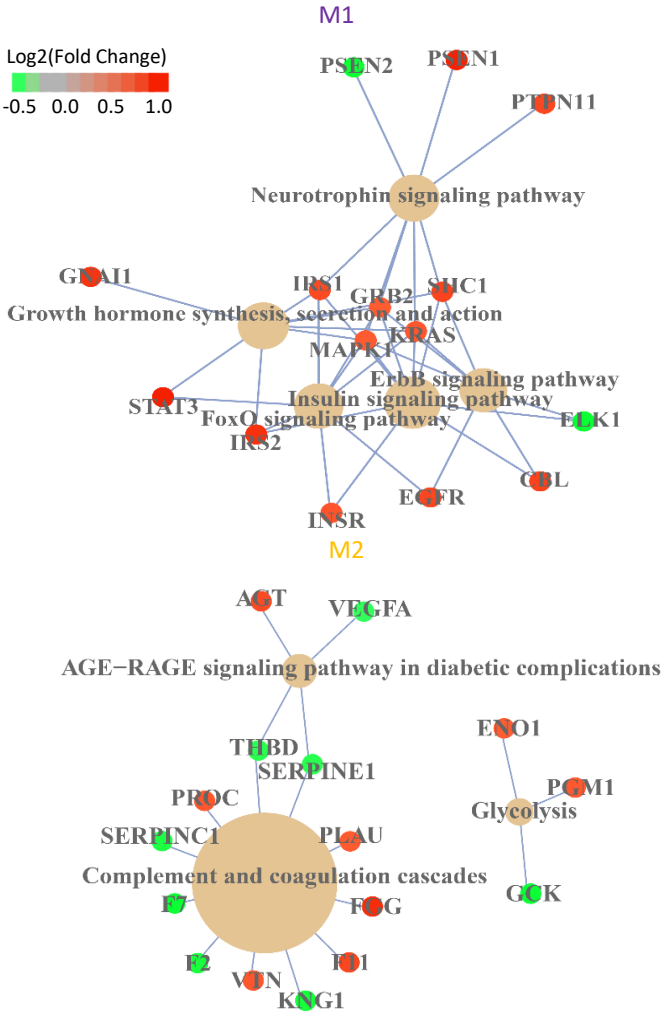


not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

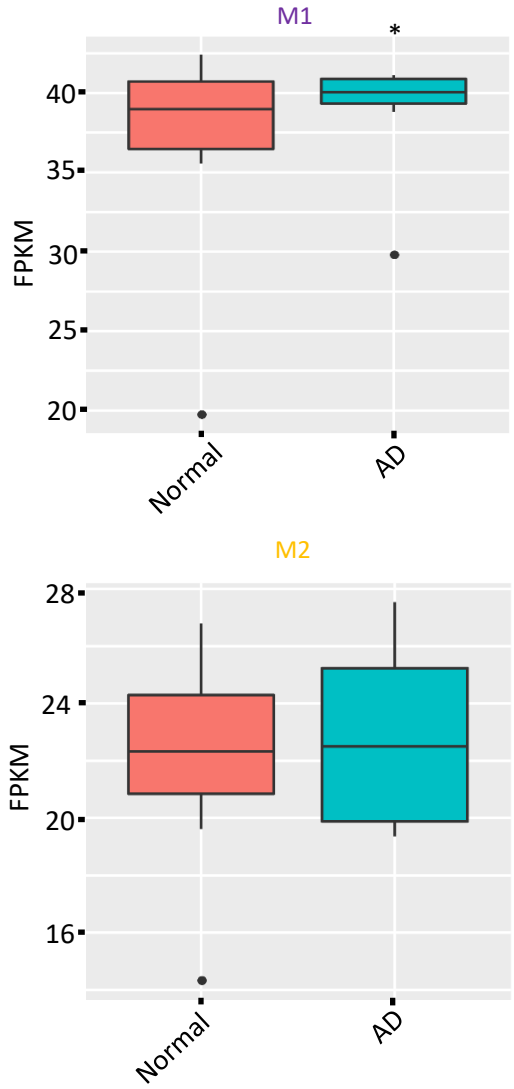
c



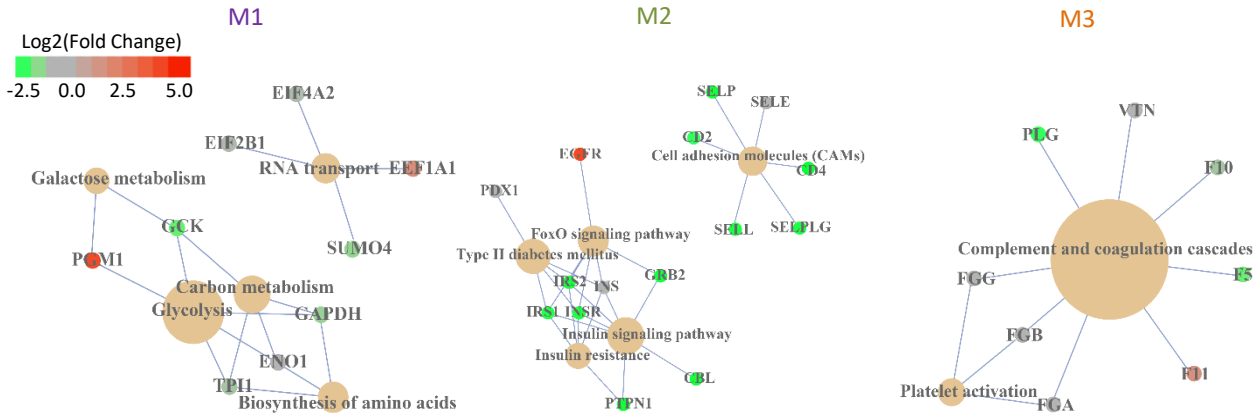
d



e

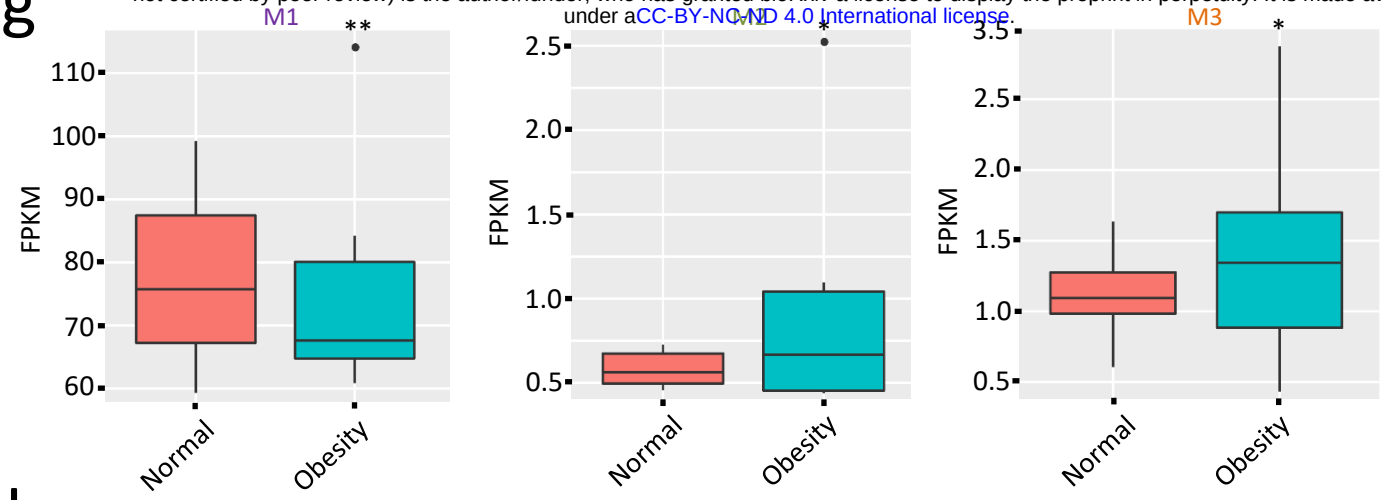


f

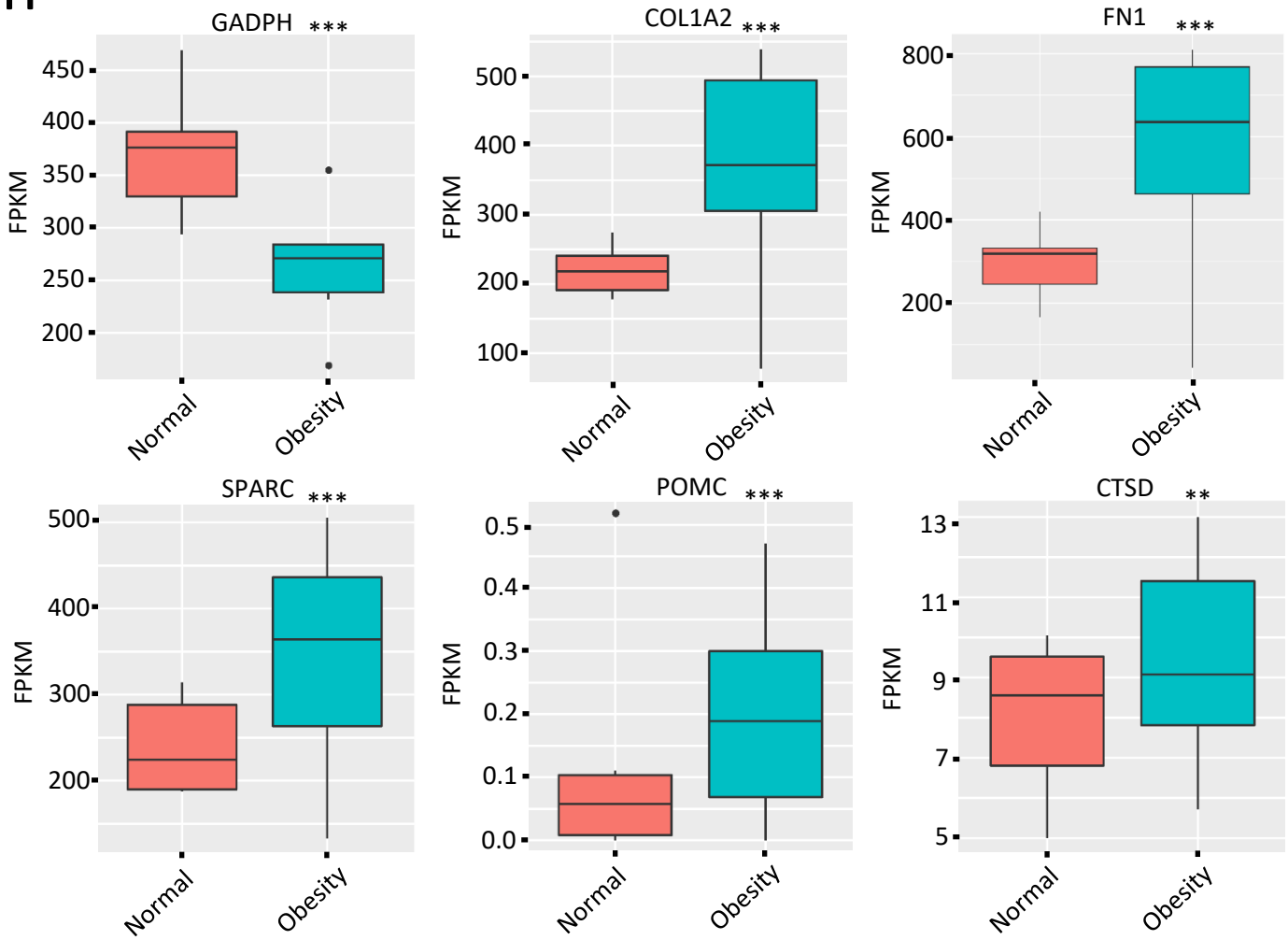


not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

g

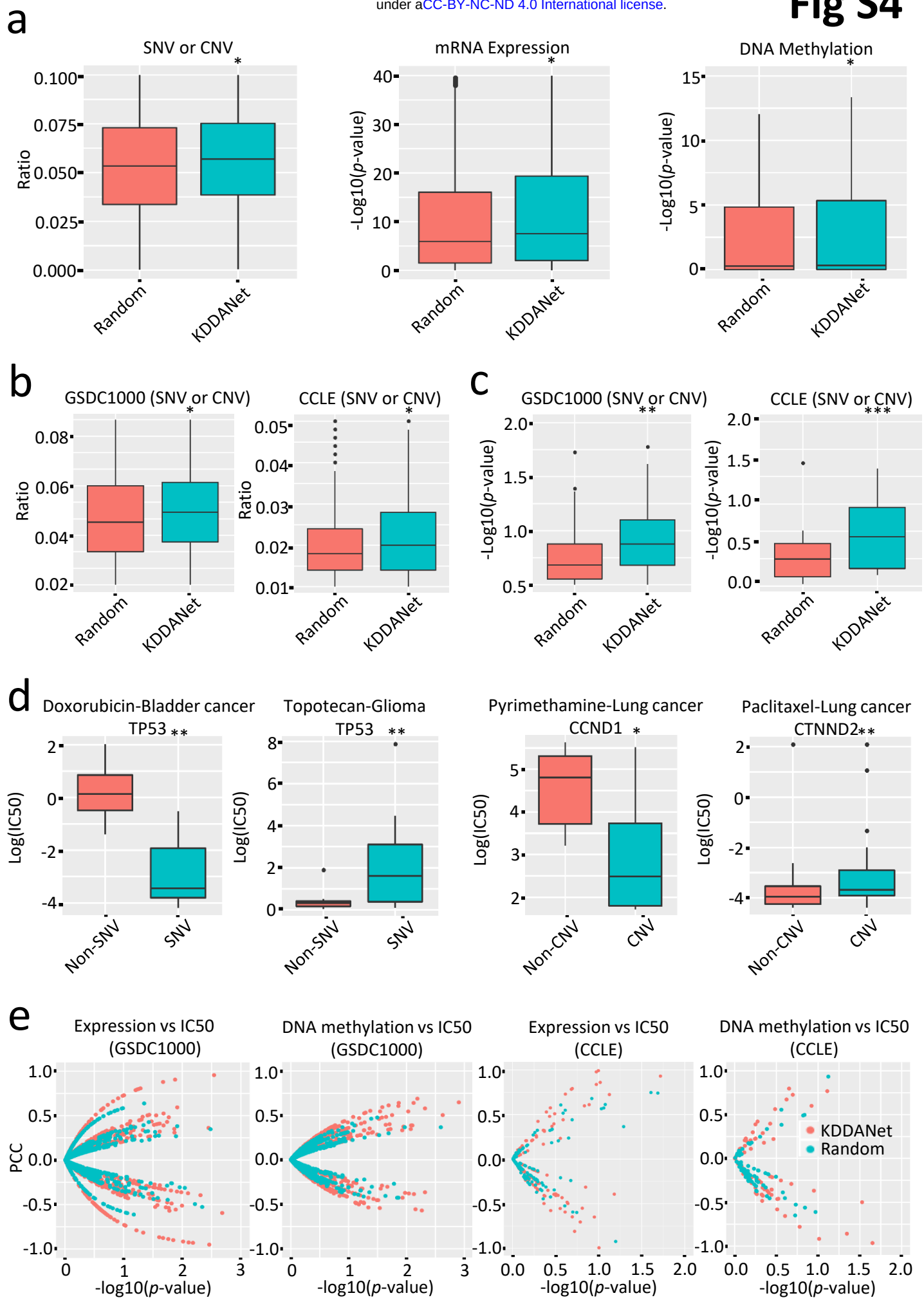


h



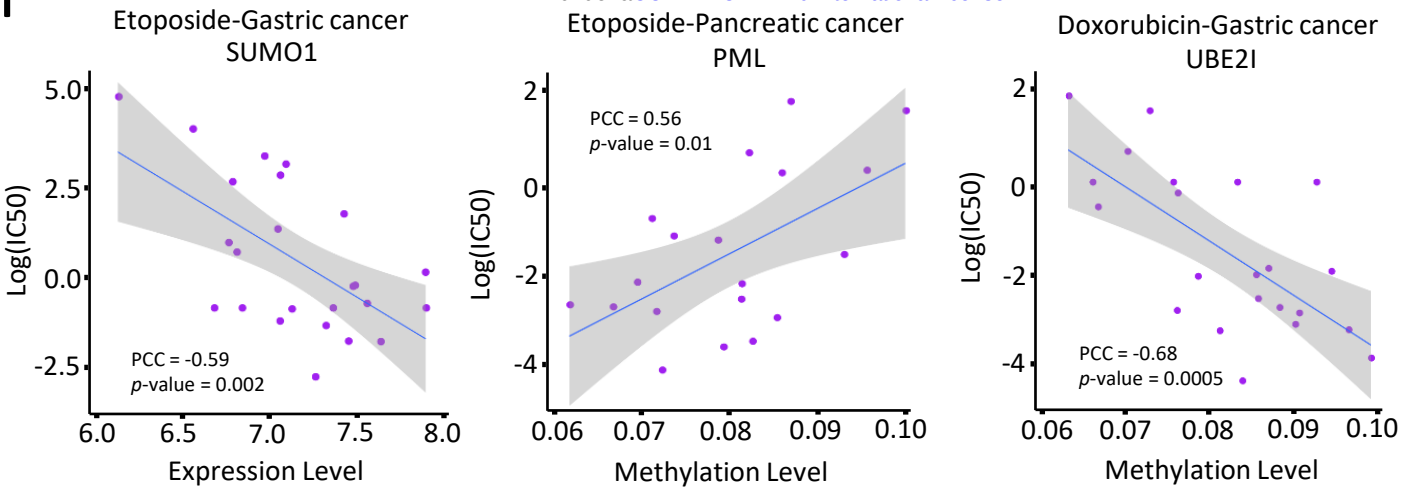
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available
under aCC-BY-NC-ND 4.0 International license.

Fig S4

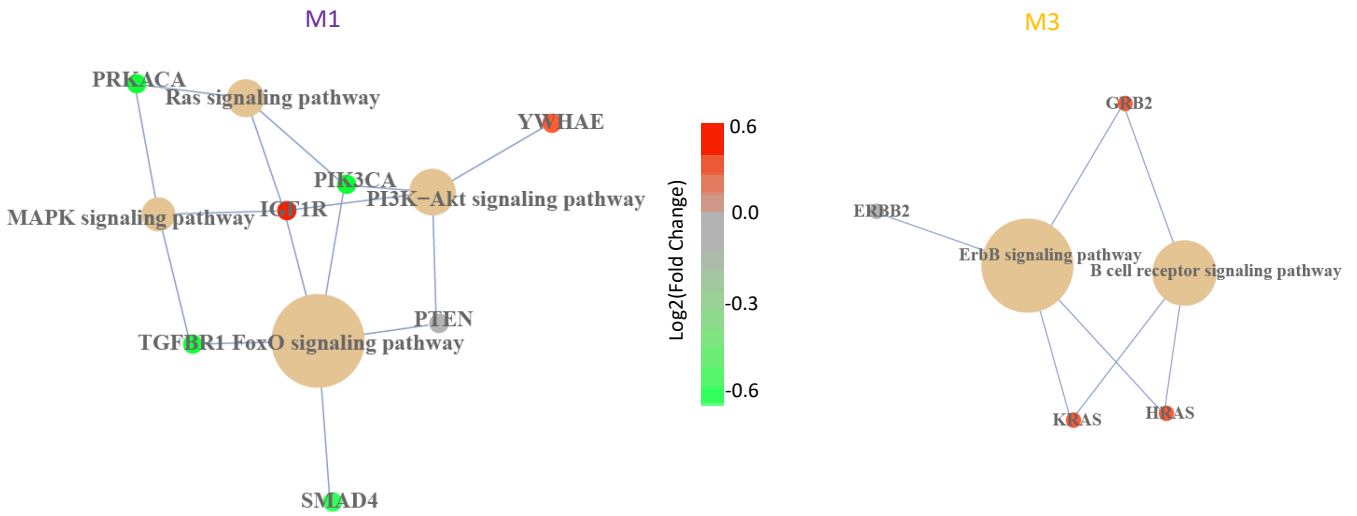


not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

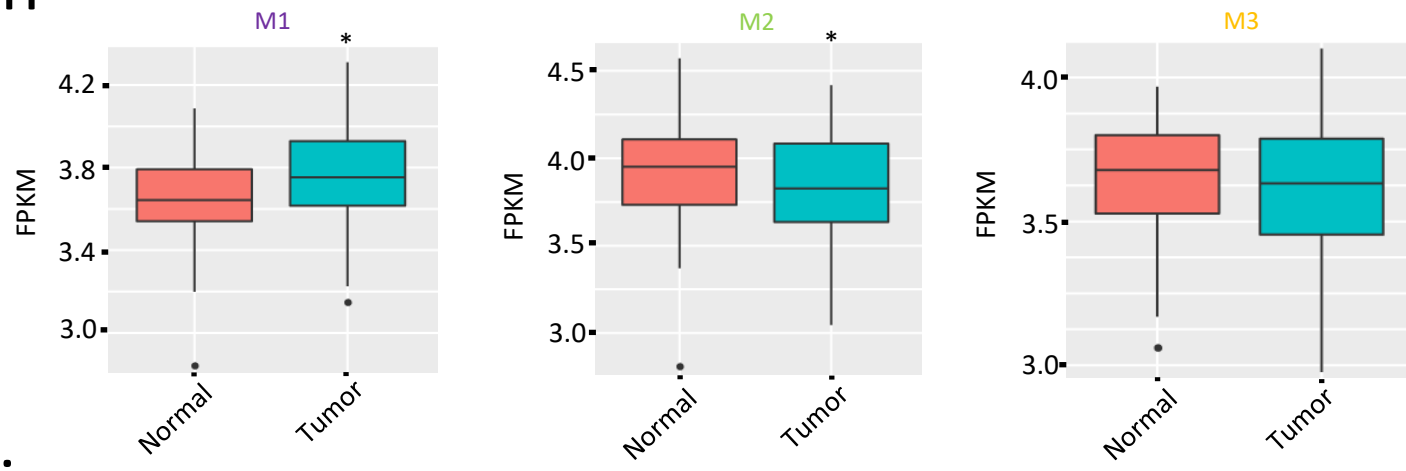
f



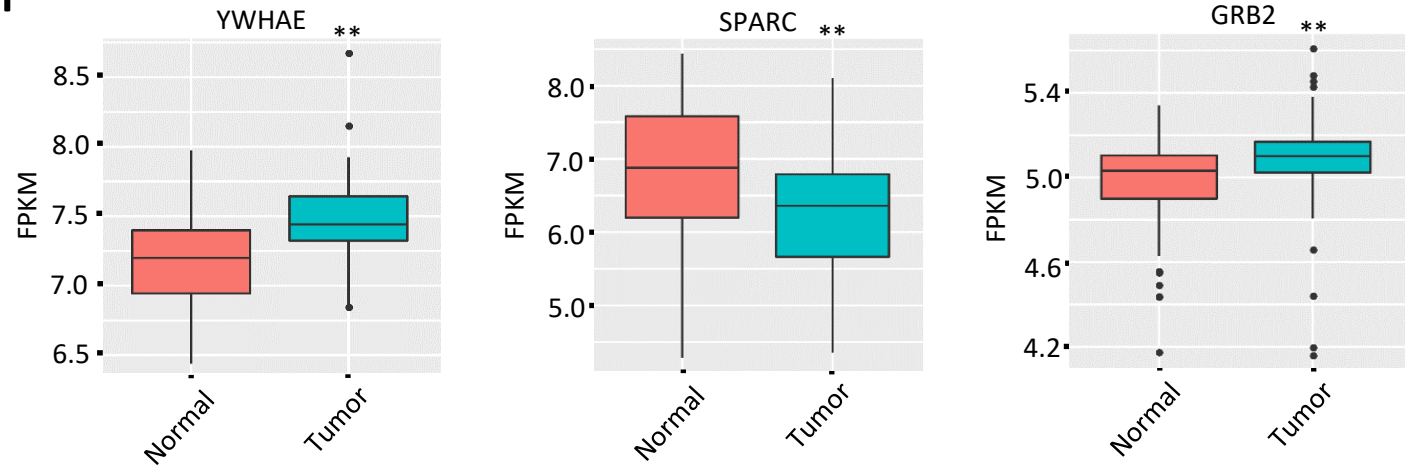
g



h

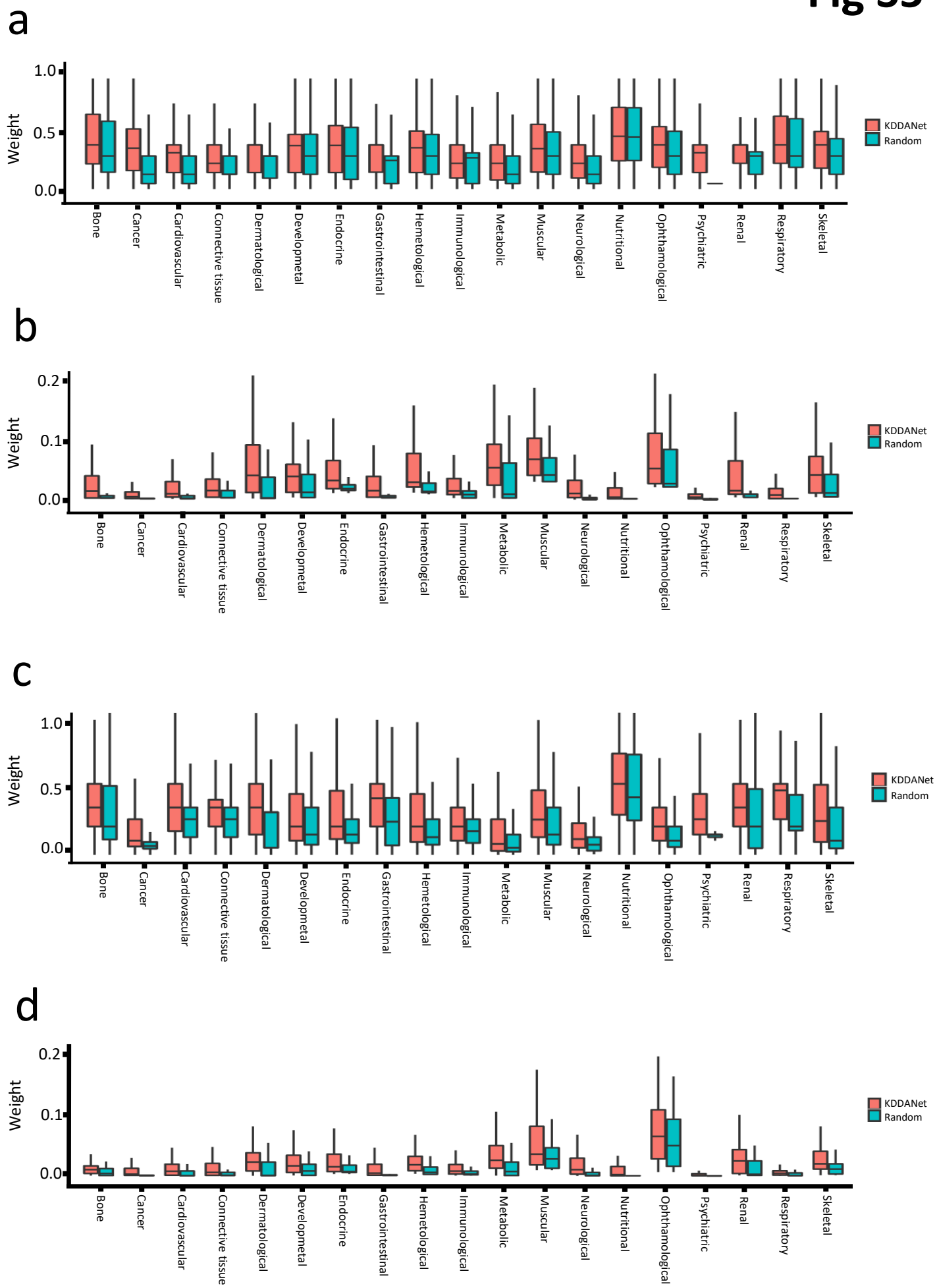


i



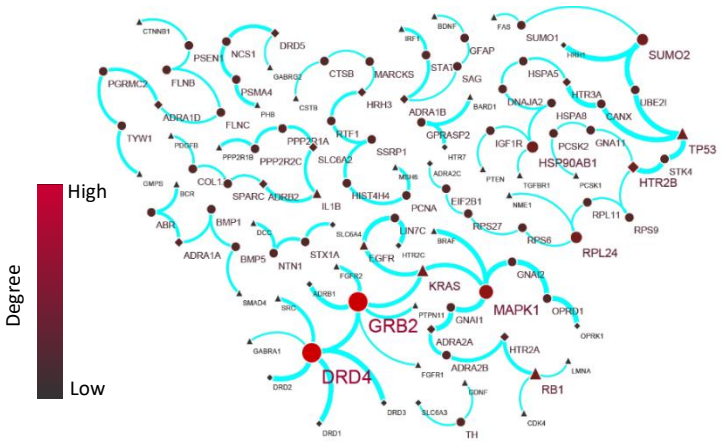
not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NC-ND 4.0 International license.

Fig S5

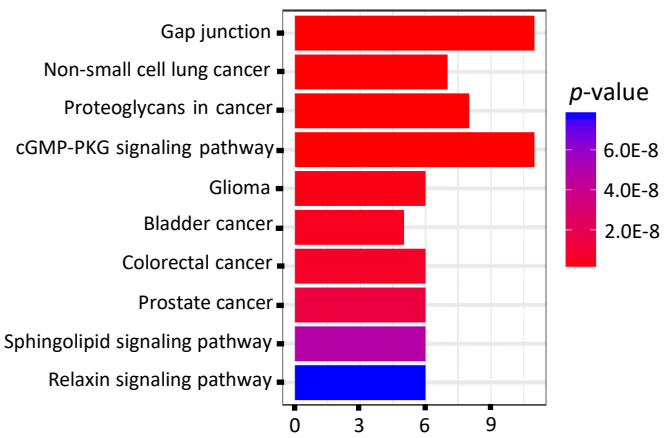


not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made availa
under aCC-BY-NCND 4.0 International license.

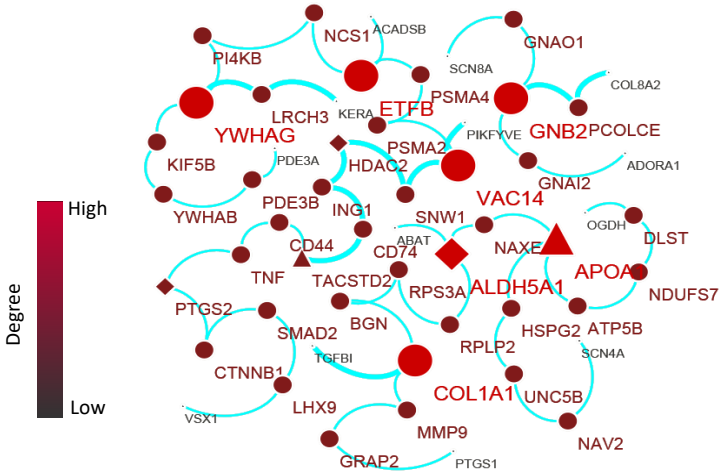
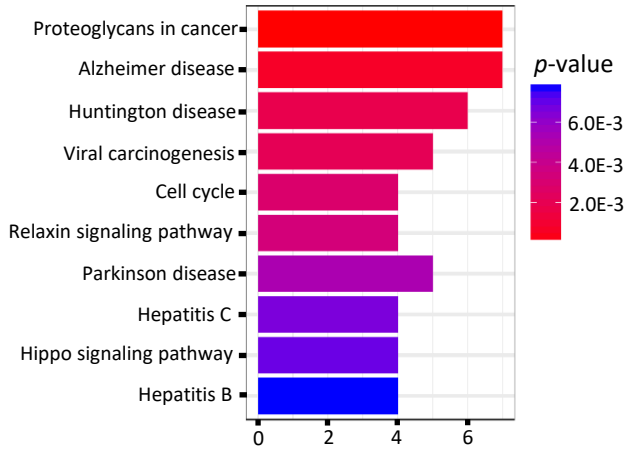
e



f



g



h