frontiers in Cell and Developmental Biology

METHODS
published: 24 June 2015
doi: 10.3389/fcell.2015.00028

How to build personalized multi-omics comorbidity profiles

Mohammad Ali Moni[1, 2, 3*†] *and Pietro Liò*[1]

[1] Computer Laboratory, University of Cambridge, Cambridge, UK, [2] Department of Computer Science and Engineering, Pabna University of Science and Technology, Pabna, Bangladesh, [3] Bone Biology, Garvan Institute of Medical Research, The University of New South Wales, Sydney, NSW, Australia

OPEN ACCESS

Edited by:
Christine Nardini, Partner Institute for Computational Biology, China

Reviewed by:
Anshu Bhardwaj, Council of Scientific and Industrial Research, India
Maria Secrier, European Molecular Biology Laboratory, Germany
Chenggang Yu, Henry M. Jackson Foundation for the Advancement of Military Medicine, USA

****Correspondence:***
Mohammad Ali Moni, Bone Biology, Garvan Institute of Medical Research, The University of New South Wales, 384 Victoria street, Durlinghurst, Sydney, NSW 2010, Australia
m.moni@garvan.org.au

†Present Address:
Mohammad Ali Moni, Garvan Institute of Medical Research, University of New South Wales, Sydney, Australia

Specialty section:
This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology

Received: *30 September 2014*
Accepted: *27 April 2015*
Published: *24 June 2015*

Citation:
Moni MA and Liò P (2015) How to build personalized multi-omics comorbidity profiles. Front. Cell Dev. Biol. 3:28. doi: 10.3389/fcell.2015.00028

Multiple diseases (acute or chronic events) occur together in a patient, which refers to the disease comorbidities, because of the multi ways associations among diseases. Due to shared genetic, molecular, environmental, and lifestyle-based risk factors, many diseases are comorbid in the same patient. Methods for integrating multiple types of omics data play an important role to identify integrative biomarkers for stratification of patients into groups with different clinical outcomes. Moreover, integrated omics and clinical information may potentially improve prediction accuracy of disease comorbidities. However, there is a lack of effective and efficient bioinformatics and statistical software for true integrative data analysis. With the availability of the wide spread huge omics, phenotype and ontology information, it is becoming more and more practical to help doctors in clinical diagnostics and comorbidity prediction by providing appropriate software tool. We developed an R software **POGO** to compute novel estimators of the disease comorbidity risks and patient stratification. Starting from an initial diagnosis, omics and clinical data of a patient the software identifies the association risk of disease comorbidities. The input of this software is the initial diagnosis of a patient and the output provides evidence of disease comorbidities. The functions of **POGO** offer flexibility for diagnostic applications to predict disease comorbidities, and can be easily integrated to high–throughput and clinical data analysis pipelines. **POGO** is compliant with the Bioconductor standard and it is freely available at www.cl.cam.ac.uk/∼mam211/POGO/.

Keywords: comorbidity, multi-omics, ontology, multiplex network, data integration

Introduction

Exploring disease-disease associations by using multi-omics and clinical information is expected to improve our current knowledge of disease relationships, which may lead to further improvements in disease diagnosis, prognosis and treatment (Park et al., 2009). Recent research has increasingly demonstrated that many seemingly dissimilar diseases have common molecular mechanisms and strong associations among them (Yu and Wang, 2015). Because of the associations among diseases, multiple diseases (acute or chronic events) occur together in a patient, which is called disease comorbidities. Comorbidities relationships exist among diseases whenever they impact the same patients significantly more than expected by chance (Žitnik et al., 2013). It represents the co–occurrence of diseases or presence of different illness or medical conditions simultaneously or one after another in the same patient (Hidalgo et al., 2009; Park et al., 2009). The set of sequential disease associations, which refers to disease trajectories, uncovers time based disease comorbidity associations. They can also form the basis for understanding mathematical properties of

co-morbidity networks (Hidalgo et al., 2009; Jensen et al., 2014). Comorbidity associations can be due to direct or indirect causal relationships and the shared risk factors among them (Tong and Stevenson, 2007). If two diseases have comorbidity association, the incidence of one of them in an individual may increase the likelihood of another disease occurring. Certain diseases, such as diabetes and obesity often co-occur in the same patient, sometimes one being considered a significant risk factor for the other (Lee et al., 2008). Disease comorbidities are increasingly placing a greater burden on individuals, societies and health care services. It is an important factor for better risk stratification of patients and treatment planning.

Diseases with similar molecular, environmental, and lifestyle risk factors may be comorbid in individuals or may be risk factors for another disorder (Davis et al., 2010). Shared genetic, environmental and lifestyle factors have similar consequences, increasing the co-occurrence of associated diseases in the same individual. So, a person diagnosed for a combination of disorders and exposed to particular environmental, lifestyle and genetic risk factors may be at a increased risk of developing several other genetically and environmentally associated diseases (Barabási et al., 2011). It is now well accepted that phenotypes are determined by genetic material under environmental influences. For instance, many well-known and influential lifestyle factors such as smoking, diet, and alcohol intake are actively related to diabetes type 1 and type 2, and obesity (Astrup, 2001). Moreover, many complex diseases, such as cancer and diabetes, are affected by an integrated effect of environment and epistasis among many genes (Davis et al., 2010).

Recent evidence has exhibited that microRNAs play key roles in the evolution and progression of human diseases. Functionally related microRNAs tend to be associated with phenotypically similar diseases (Lu et al., 2008). Recently, genome-wide association studies (gwas) proved to be useful as a method for exploring phenotypic associations with diseases (Lewis et al., 2011). Single-nucleotide polymorphisms (SNPs), a variation of a single nucleotide, are assumed to play a major role in causing phenotypic differences between individuals. It has become possible to assess systematically the contribution of common SNPs to complex diseases. Copy number variations (CNVs; which involve loss, duplication or rearrangement of long stretches of DNA in individual's genome) can cause various phenotypic abnormalities (Zhang et al., 2009). CNVs are significantly associated with the risk of complex human diseases including inflammatory autoimmune disorders, diabetes etc. (Bae et al., 2011). The development of type 2 diabetes has also been known to be influenced by molecular, lifestyle and environmental factors (Kahn et al., 2006).

Most of the research works focussed on a particular data type, for example gene expression, to find profiles that are associated with particular disease, prognosis and drug response. The integrative analysis of various omics data has become increasingly widespread because each approach has intrinsic caveats. For instance, important information may be missing because of false negatives or may be misleading because of false positives. In addition, by analyzing different types of data in isolation we may miss important information that results from the coordinated activity of biological components at various levels. Some studies indicated that these limitations can be mitigated by integrating two or more omics datasets. Several studies (Goh et al., 2007; Lee et al., 2008; Lu et al., 2008; Hu and Agarwal, 2009; Liu et al., 2009; Park et al., 2009; Schadt, 2009; Jiang et al., 2010; Suthram et al., 2010) reported on the role of a single omic or phenotypic measure to represent disease-disease associations (such as shared pathways or gene ontology). But, one needs to study diverse sources of evidence including miRNA-based relationships, shared environmental factors, ontology, SNPs, CNVs and phenotypic manifestations for better understanding.

Since, diseases may share many different types of associations with varying levels of risk for disease comorbidities, a singular view of associations between diseases is not enough to predict comorbidities. As more and more ontology, phenotype, omics and environmental data sets become publicly available, it is beneficial to improve our understanding of human diseases and diseases comorbidities based on these new system-level biological data. Combination of multiple types of omics, phenotype and ontology data identifies integrative biomarkers for the stratification of patients with clinical outcome. Further, behavioral and environmental aspects should also be considered in order to realize disease-disease associations. Therefore, it is clear that method and tool for stratifying patients and prediction of disease comorbidities in order to reliably predict prognosis or success of treatments are of critical importance in the field of medicine. We propose a computational framework that integrates all available, heterogeneous and relevant data including miRNA-target interactions, miRNA-disease association, phenotype similarities of diseases, GO (gene ontology), SNPs, CNVs and known disease-environmental associations to capture the complex relationships between phenotypes, genotypes and clinical comordibidity. Therefore, the underlying goal of this chapter is to integrate diverse sets of omics, environmental and phenotypic data, and to develop the comprehensive models of interaction between the disease associated factors for the prediction of the patient specific disease comorbidity, and to develop comorbidity map.

In the case of a complex or even in an unknown case of diseases, physicians may get assistance to take decision quickly and efficiently by using effective software tool. We developed an R software tool POGO to compute statistically significant associations among diseases, to predict disease comorbidity risk and to develop comorbidity maps, which are useful for the physicians and informative for the patients. To perform the computation of the comorbidity risk, this software uses clinical, gene expression, miRNA, SNPs, CNVs, ontology, phenotypic, and environmental data. The inputs of this software is the initial diagnostic result of the patient. The goal of this software is to construct comorbidity maps that incorporate disease interactions, omics, phenotypic and ontology information, and environmental influences. It is a user-friendly and interactive personalised disease and disease comorbidity prediction software. It provides different comorbidity assessment and stratification; integration of omics information with

POGO output data could be used to predict more accurate survival probability of patients. The functions included in POGO offer flexibility for applications, and can be easily integrated into highthroughput analysis pipelines for translation medicine.

Implementation

POGO provides a number of processing options to find comorbidity maps of a patient. R bioconductor annotation data packages "org.Hs.eg.db," "HPO.db," and "GO.db" are used for the annotation and mapping between gene symbol, Entrez id, HPO term, OMIM id and GO term (Gentleman et al., 2004). POGO is dependent on "DOSE" and "GOSemSim" bioconductor packages for the mapping with different annotation (Yu and Wang, 2015). We used the mapping manually constructed by Goh et al. (2007) and Park et al. (2009) to convert OMIM IDs to ICD-9 codes. A set of differential expressed gene symbols/Entrez ids/OMIM id/miRNA ids/HPO terms/GO terms/3 or 5 digit ICD-9-CM code of any disease can be used as input of POGO functions. Flow diagram of POGO software is shown in **Figure 1**.

GO–disease Association

GO enables us to analyse disease association by adopting semantic similarity measures to expand our knowledge of the relationships among different diseases. We downloaded the ontology file and annotations of Homo sapiens from the Gene Ontology database[1] in April 2014. In total, we collected 171,888 annotations between 13,166 genes and 10,787 GO terms. We developed a function `comorbidityGO` for the computation of GO based disease comorbidity in an ontology sense. It is a GO-based enrichment analysis function to measure association among diseases and to explore their functional associations from gene sets. We implemented a semantic similarity measurement to quantify the association between gene ontology and their associated diseases. The semantics of GO terms are encoded into a numeric format and the different semantic contributions of the distinct relations are considered. Moreover, hypergeometric test is applied to a gene set to calculate the significance of a GO term, and the significant GO term sets are selected according to their *p*-values. Gene set enrichment analysis are used for predicting

[1]http://www.geneontology.org.

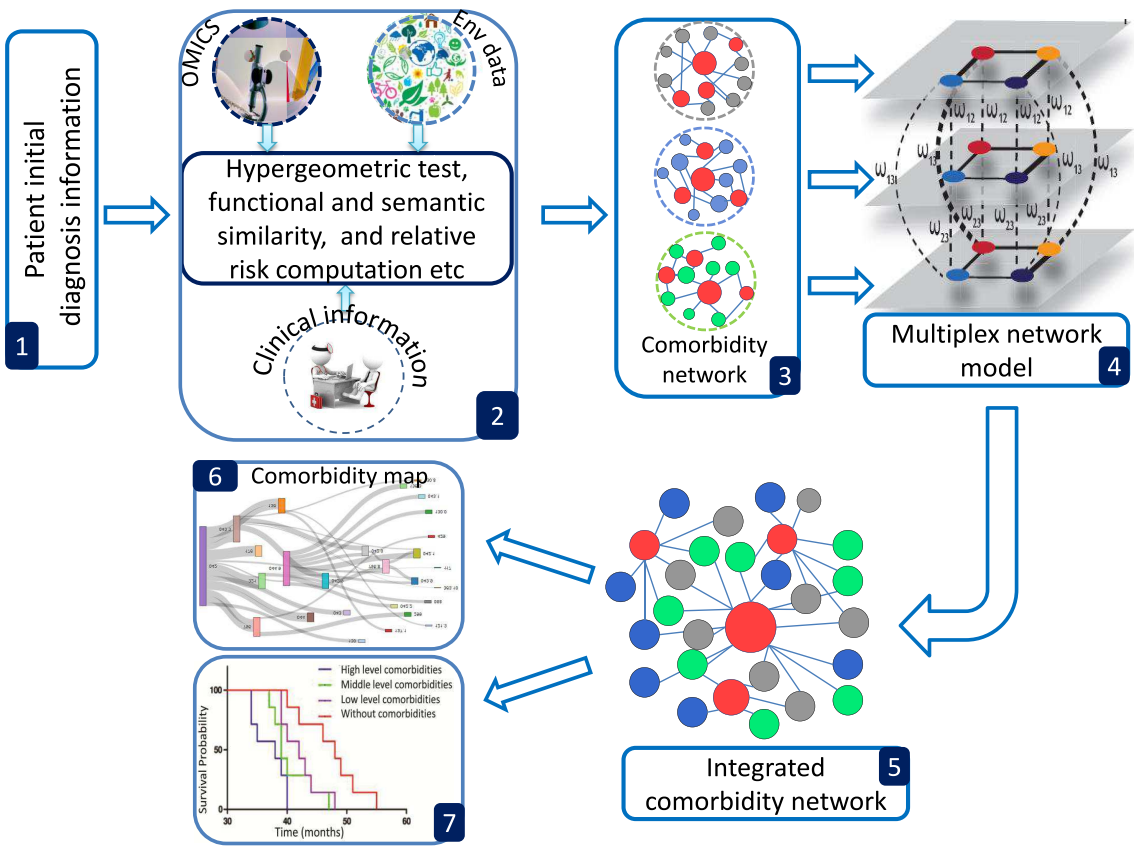


FIGURE 1 | Overview framework of POGO software. (1) POGO takes as input preliminary diagnosis data of a patient and check the validation of the input. (2) It preprocesses and updates required databases, performs statistical computation (hypergeometric and semantic similarity tests), and calculates relative risk between diseases. (3) Comorbidity scores and disease network are provided as a result to the user. (4) Multiplex model is applied for data integration to produce integrated comorbidity network as (5). (6, 7) Visualization of the comorbidity map and survival probability of patient considering comorbidity. Env is used to indicate environment.

the significance of gene–disease and disease–disease associations. `comorbidityGO` function operates by using either of the following input: GO id, disease OMIM id, a list of gene symbols, Entrez gene ids or ICD-9 code of the patient disease. This function provides disease comorbidity associations and network based on the GO. `comorbidityGO` requires two parameters: id list and id type. An example and its output is given in **Figure 2**.

```
> comorbidityGO( "189907" , "OMIM" )

OMIM           GO EVIDENCE ONTOLOGY  PATH   SYMBOL ENTREZID ICD9CM
189907 GO:0000122      IEA       BP 04950     TCF2     6928    250
189907 GO:0001714      IEA       BP 04950   BCKDHB      594  270.3
189907 GO:0005634      IDA       CC 04950     TCF2     6928    189
189907 GO:0044212      IEA       MF 04950     TCF2     6928  593.9
... ...
```

Phenotype–disease Association

POGO integrated HPO database that has integrated HPO terms to represent patients phenotypic abnormalities (Robinson et al., 2008). The OMIM (McKusick, 2007) is also incorporated with POGO, and associated to HPO by annotations from http://www.human-phenotype-ontology.org. The associations are generated using the information about the phenotypes of a particular syndrome and the corresponding genes that are known to cause this syndrome when mutated. With the development of omics techniques, the number of uncovered gene-phenotype associations has increased notably over the last few years. In our approach, phenotypes are linked with diseases through associating phenotype-gene with gene-disease bipartite graphs by applying neighborhood-based methods. All the paths from a phenotype to a disease are explored by considering causative genes to assign a weight based on frequency and linked the phenotype to the disease in a new phenotype-disease bipartite graph. Then, we introduced a Bidirectionally-induced Importance Weight prediction method to phenotype-disease bipartite graph in order to approximate the weights of the edges of diseases with phenotypes, by considering link information from both sides of the phenotype-disease bipartite graph. The construction of the phenotype network is based on the phenotypic similarity score among different disease phenotypes. In the phenotype network, the association between any two different disease phenotypes was fixed when their phenotypic similarity score exceeded the significance threshold. For visualization, POGO includes links between disease pairs for which the co-occurrence is notably greater than the random expectation based on phenotype prevalence of the diseases. The function `comorbidityHPO` of POGO package is able to take input an OMIM id/3 or 5 digit ICD-9-CM code of a disease or a list of gene symbols/Entrez ids and provides comorbidity pattern of diseases based on the phenotype disease associations. `comorbidityHPO` requires two parameters: id list and id type. An example and its output is given in **Figure 3**.

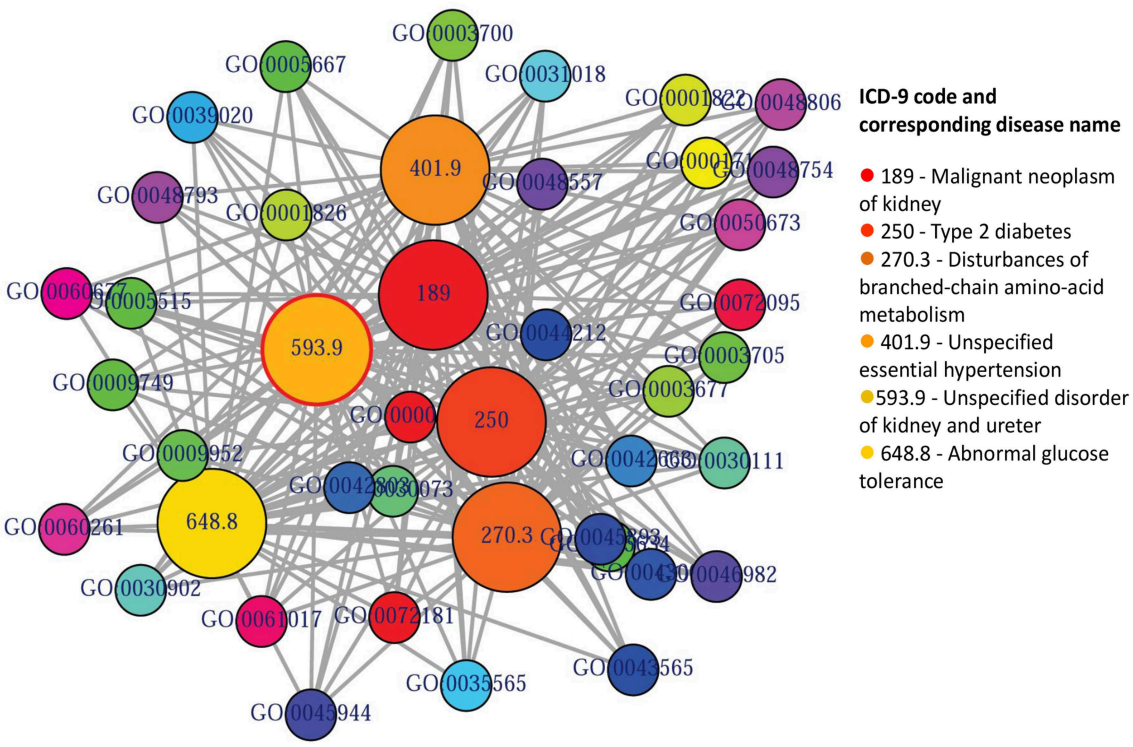


FIGURE 2 | Output figure and statistics of `>comorbidityGO ("189907", "OMIM")`. The OMIM disease id of the "Diabetes mellitus, insulin-dependent" is 189907, which is used as input to the `comorbidityGO`. We show disease comorbidity for the Diabetes mellitus through the GO-disease associations. The size of the nodes represents the degree of associations. ICD-9 codes are used to represent disease categories.

```
> comorbidityHPO( "79001" , "Entrez" )

ENTREZID    SYMBOL  OMIM    PATH    GO
79001       VKORC1  122700  NA      GO:0005789
79001       VKORC1  122700  NA      GO:0005789
79001       VKORC1  607473  NA      GO:0005789
79001       VKORC1  608547  NA      GO:0047057
79001       VKORC1  608547  NA      GO:0047057
... ...

HPID        HPName
HP:0000118  Phenotypic abnormality
HP:0012200  Abnormality of prothrombin
HP:0001892  Abnormal bleeding
HP:0003256  Abnormality of the coagulation cascade
HP:0010989  Abnormality of the intrinsic pathway
... ...
```

Disease–SNPs Association

At present there are only a few databases of genetic variations associated with diseases. Despite the needs for analyzing SNP and disease association, most of the existing databases are based only on functional variants at specific locations on the genome, or deal with only a few genes correlated with disease. There is no integrated resource to widely support genes, SNPs, and disease associated information. Therefore, we integrated data from different databases (dbSNP Sherry et al., 2001, HGVbase Fredman et al., 2002, JSNP Hirakawa et al., 2002, GAD Becker et al., 2004 and OMIM McKusick, 2007) and literature Yang et al., 2008 for studying SNPs-diseases associations. We integrated the information to present the interrelationships among SNPs located in genes, genes associated with diseases, and SNPs associated with diseases. It can aid the understanding of the genes which cause diseases and the impact of SNPs on diseases. For associated information among genetic variation and diseases, we built a database, SNP, which is a combined database of genes, genetic variation and diseases for the utilization in POGO. Two diseases are connected if they share at least one SNP that is statistically significant dysregulated to the disease related gene. Our software is designed to capture the relationships between SNPs associated with disease and disease-causing genes. POGO computes disease-disease association by adopting semantic similarity measures and hypergeometric test. Neighborhood based benchmark method is used to identify the comorbidity pattern among diseases (Goh et al., 2007). We built the associated network as a bipartite graph; each common neighbor node is selected based on the Jaccard coefficient method (Goh et al., 2007). `comorbiditySNP` function of POGO takes as input any of these three options: a list of gene symbols, a list of Entrez gene ids, SNPs ids or an OMIM id. This function provides disease comorbidity associations and network based on the SNPs-gene-disease associations. `comorbiditySNP` requires two parameters: id list and id type. An example and its output is given in **Figure 4**.

```
> inputList<-c("TNFSF11", "TNFRSF11B", "TNFRSF11A", "A2M", "TGFBR3")
> comorbiditySNP(inputList, "Symbol")

SYMBOL      OMIM    ENTREZID    PATH
TNFRSF11A   174810  8792        04060
TNFRSF11A   602080  8792        04060
TNFRSF11A   603499  8792        04060
TNFRSF11B   239000  4982        04060
TNFRSF11B   239000  4982        04060
... ...

GO          SNPID       DiseaseName
GO:0043123  rs884205    Bone mineral density
GO:0002250  rs3018362   Pagets disease
GO:0043123  rs694419    Serum albumin level
GO:0007165  rs2062375   Osteoporosis
GO:0007165  rs12679857  Type 1 diabetes
... ...
```

Disease–environment Association

The analysis of environment-disease associations is important to investigate the molecular mechanism of a disease. POGO integrated "etiome," human disease etiological factors database (Liu et al., 2009), and developed a function `comorbidityENV` to predict the comorbidity risk based on disease environment association (Liu et al., 2009). Integrating genetic, nutritional, behavioral and environmental factors results in the "etiome," which they defined as the comprehensive compendium of disease etiology (Liu et al., 2009). They used natural language processing to look for annotations in articles, and thus creating associations between diseases and environmental information. "etiome" has been developed with the identified 3342 environment related factors that are associated with 3159 complex diseases (Liu et al., 2009). They also identified 1100 genes associated with 1034 diseases from the genetic association studies database GAD (Becker et al., 2004). GAD has 863 diseases information with both genetic and environmental etiological factors. By using all these information, POGO is able to develop comorbidity map by incorporating relations between the diseases themselves as well as relations to environmental factors. This software identifies the disease–disease associations using the associations among environment and their associated diseases. Hypergeometric test is used for extracting associations among environment and diseases; graph topological structure is used to measure the similarity between diseases (Wang et al., 2007). `comorbidityENV` function takes as input any of the following options: a list of gene symbols, a list of Entrez gene ids or an OMIM id. This function provides disease comorbidity associations and network based on the gene-environment-disease associations. `comorbidityENV` requires two parameters: id list and id type. An example and its output is given in **Figure 5**.

```
> comorbidityENV( "SDHB" , "Symbol" )
SYMBOL  OMIM    ENTREZID PATH   GO          EVIDENCE
SDHB    115310  6390     00020  GO:0005515  IPI
SDHB    115310  6390     00020  GO:0005515  IPI
SDHB    115310  6390     00020  GO:0005515  IPI
SDHB    612359  6390     05016  GO:0051539  ISS
SDHB    612359  6390     05016  GO:0051539  ISS
... ...

ONTOLOGY    DiseaseName     EnvironmentImpact
MF          Bone Neoplasms  Bone Cysts
MF          Bone Neoplasms  Bone Marrow Transplantation
MF          Bone Neoplasms  HIV Infections
MF          Bone Neoplasms  Kidney Transplantation
MF          Bone Neoplasms  Heart Transplantation
... ...
```

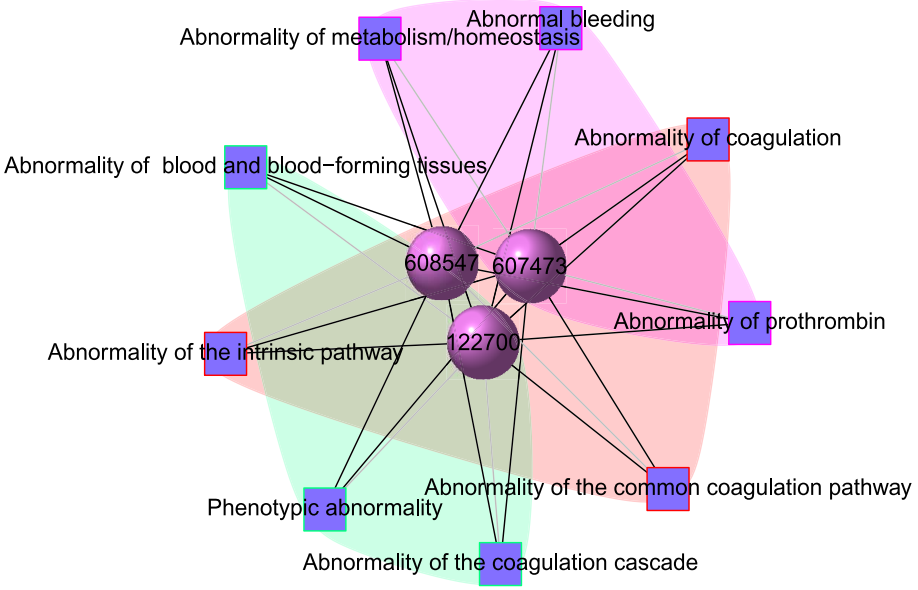


FIGURE 3 | Output figure and statistics of `> comorbidityHPO ("79001", "Entrez")`. The Entrez disease id "79001" is used as input to the `comorbidityHPO`. We show an example of disease comorbidity map for this gene through the phenotype-disease associations. Here the square nodes represent the phenotypes and spheres represent OMIM disease ids.

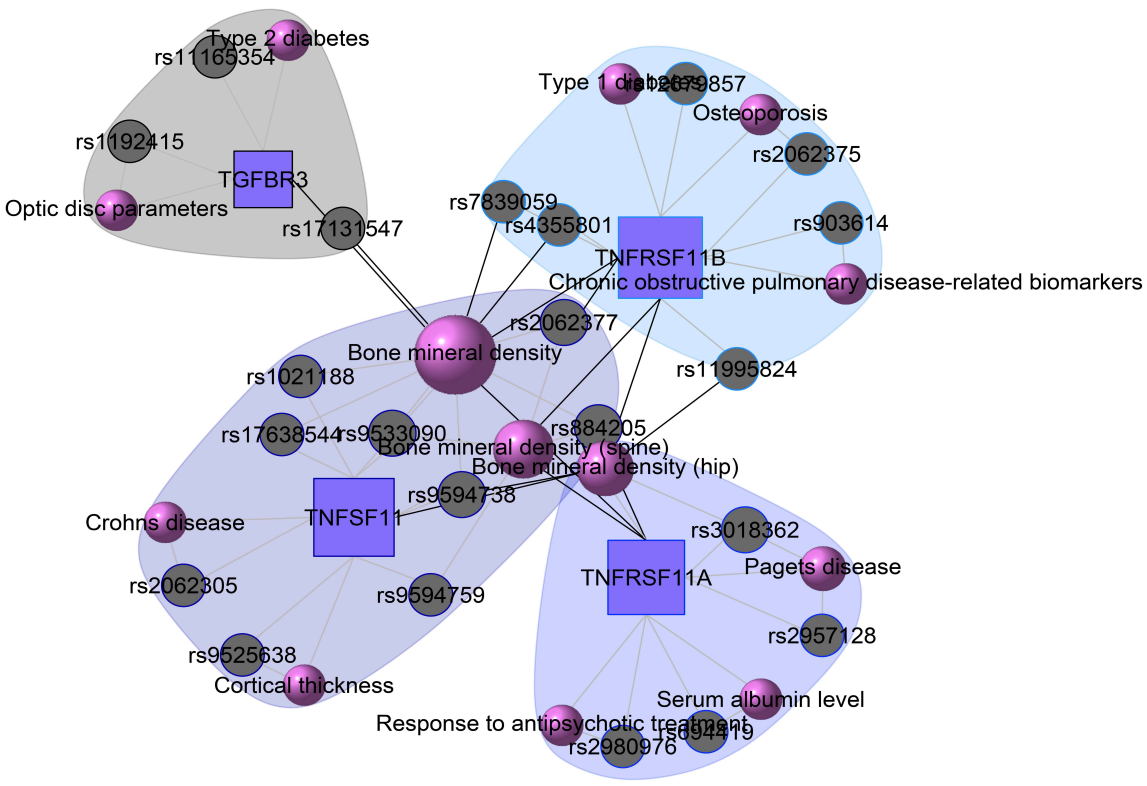


FIGURE 4 | Output figure and statistics of `>comorbiditySNP (c("TNFSF11", "TNFRSF11B", "TNFRSF11A", "A2M", "TGFBR3"), "Symbol")`. We show an example of disease comorbidity through the SNPs-gene-disease associations. Here the square nodes represent the genes symbols, circles represent SNPs ids, and spheres represent diseases names. The size of the nodes represents the degree of associations.

miRNA–disease Association

MicroRNA (miRNA) performs its regulatory function through its target genes. Two diseases are connected if they share at least one gene and/or one miRNA that is statistically significant dysregulated (Goh et al., 2007). miRNAs with similar functions tend to be associated with diseases with similar phenotypes, and vice versa (Lu et al., 2008). Based on these hypothesis, we used a framework to identify miRNA-disease associations through the direct identified association from the miRNA-disease association database and indirect association from the combined database of miRNA-target and gene-disease associations. POGO makes use of microRNA-target databases, miR2Disease (Jiang et al., 2009), HMDD (Li et al., 2014), and gene-disease association databases, OMIM (McKusick, 2007), to explore the mRNA and miRNA association between diseases. We filtered out invalid miRNA-disease associations with incorrect disease names or miRNA names. We used National Library of Medicine[2] to obtain the correct disease names. We used miRBase to get the correct miRNA names (Kozomara and Griffiths-Jones, 2011). For a miRNA-disease pair, firstly, POGO maps the causal genes of the disease. It uses a *p*-value to measure the significance of the association between the miRNA and the disease. OMIM diseases ids are mapped with ICD-9-CM codes based on the literature (Park et al., 2009). Neighborhood based benchmark method is used to identify the comorbidity pattern among diseases. We build the associated network as a bipartite graph; each common neighbor node is selected based on the Jaccard coefficient method (Goh et al., 2007). `comorbiditymiRNA` function of POGO takes as input any of the following options: a list of gene/miRNA symbols, a list of Entrez gene ids, an ICD-9 code, an GO id or an OMIM id. This function provides disease comorbidity associations and network based on the disease-miRNA associations. `comorbiditymiRNA` requires two parameters: id list and id type. An example and its output is given in **Figure 6**.

```
> comorbiditymiRNA( "TNFRSF11A" , "Symbol" )

ENTREZID miRNAID        DiseaseName                          SYMBOL
8792     hsa-miR-432    Duchenne muscular dystrophy (DMD)    TNFRSF11A
8792     hsa-miR-324-3p primary biliary cirrhosis (PBC)      TNFRSF11A
8792     hsa-miR-324-3p lupus nephritis                      TNFRSF11A
8792     hsa-miR-432    miyoshi myopathy (MM)                TNFRSF11A
8792     hsa-miR-664    multiple sclerosis                   TNFRSF11A
8792     hsa-miR-432    nemaline myopathy (NM)               TNFRSF11A
... ...

GO          EVIDENCE    ONTOLOGY    OMIM    PATH
GO:0002250  IMP         BP          174810  5323
GO:0002250  IMP         BP          174810  5323
GO:0009897  IDA         CC          174810  4060
GO:0009897  IDA         CC          602080  5323
GO:0002250  IMP         BP          602080  4380
GO:0002250  IMP         BP          612301  5323
... ...
```

CNV–disease Association

Copy number variants are hypothesized to cause diseases through several mechanisms. Sometimes, the combination of two or more copy number variants can produce a complex disease. Additionally, complex diseases might occur when copy number variants are combined with other genetic and environmental factors (McCarroll and Altshuler, 2007). Diseases might be caused by copy number variants due to both additional copies of sequence (duplications) and losses of genetic material (deletions). We used Database Genomic Variants (DGV[3]) database and developed a function `comorbidityCNV` to predict the comorbidity risk based on CNVs-disease association (MacDonald et al., 2014). POGO makes use of DGV and OMIM (McKusick, 2007) to explore the genetic association between diseases. Two diseases are connected if they share similar copy number variations. OMIM diseases ids are mapped with ICD-9-CM codes based on the literature (Park et al., 2009). Neighborhood based benchmark method is used to identify the comorbidity pattern among diseases (Goh et al., 2007). We build the associated network as a bipartite graph; each common neighbor node is selected based on the Jaccard coefficient method (Goh et al., 2007). `comorbidityCNV` function of POGO takes as input any of the following options: a list of gene symbols, a list of Entrez gene ids or an OMIM id. This function provides disease comorbidity associations and network based on the disease-CNV associations. `comorbidityCNV` requires two parameters: id list and id type. An example and its output is given in **Figure 7**.

```
> comorbidityCNV("602228", "OMIM")

SYMBOL    OMIM ENTREZID  PATH   GO          EVIDENCE
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
TCF7L2  602228  6934    04310   GO:0005515       IPI
... ...

ONTOLOGY  CNV.ID   Chr  Start       End         VarSubtype
MF       nsv7211    10  108617417   118351740   Inversion
MF       nsv7553    10  114845707   114890646   Loss
MF    esv2074123    10  114876971   114877374   Deletion
MF      nsv24033    10  114877162   114877217   Loss
MF     nsv527837    10  114888608   114911079   Loss
... ...
```

Integrated Comorbidity Prediction Using Multiplex

As a single source of genomic data is prone to bias, incompleteness and noise, integration of different genomic data sources is designed to accomplish reliable disease comorbidities prediction. Systematic integration and comparison of multiple layers of information is required to provide deeper insights into biological systems. We incorporated a multiplex network model into POGO to integrate multiple omics, environmental and phenotypic information. To leverage the potential of multi-omics studies, exploratory data analysis methods that provide systematic integration and comparison of multiple layers of omics information are required. We applied our multiplex method of integrating different types of data

[2]http://www.nlm.nih.gov/.

[3]http://dgv.tcag.ca/.

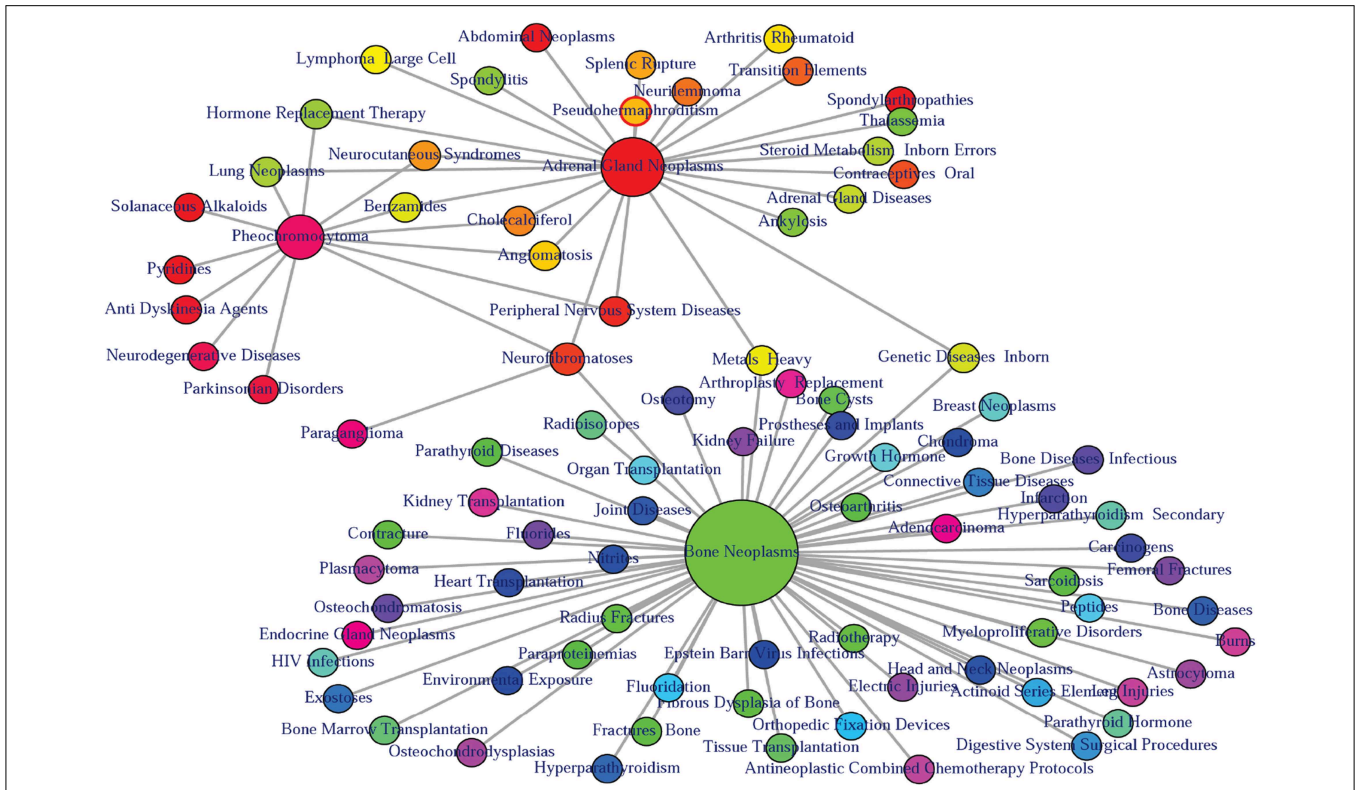


FIGURE 5 | Output figure and statistics of `>comorbidityENV("SDHB", "Symbol")`. The gene symbol "SDHB" is used as input to the `comorbidityENV`. We show disease comorbidity map for this gene input through the disease-environmental associations. The size of the nodes represents the degree of associations.

by modeling similarities between diseases in a multiplex network. The multiplex network allows us to model diseases by representing each data type as a layer in the multiplex. Importantly, this allows us to capture the interactions between the various types of data, such as the interdependence of mRNA expression and signaling pathways with clinical information of the disease comorbidities. We developed a function `comorbidityMultiplex` to predict the integrated comorbidity risk. `comorbidityMultiplex` function takes as input any number of layers information. This function provides integrated comorbidity associations and network. As an example of integrating with this function we considered three different types of data for three layers of our multiplex network: mRNA-disease, pathway-disease and clinical association information. An example and its output is given in **Figure 8**.

In this example, we considered association information of 10 diseases, which are the output of other functions of POGO. The ICD-9 code of the 10 diseases are 155, 157, 199, 286, 287, 571, 572, 574, 576, and 782. POGO identified disease-disease comorbidity associations network based on the gene-disease association and pathway-disease association, which are shown in **Figures 8A,B** respectively. It is notable that there is no shared pathway for the disease 572 with the 9 other diseases. The comorbidity network based on the clinical information is shown in **Figure 8C**. We used all these three association networks for the input of our multiplex network (see **Supplementary Tables S1–S3**). In this case, the multiplex network is comprised of three layers, each with 10 nodes. In each layer, each node has a weighted undirected edge connecting it to every other node in the same layer. In addition, each disease is connected to itself in every other layer

```
> input = c("S1.txt", "S2.txt", "S3.txt")
> sv<-c(1, 1, .5) #strength value of each layer
> comorbidityMultiplex(input, sv)
... ...
$aggG
        ICD.155 ICD.157 ICD.199 ICD.286 ICD.287 ICD.571 ICD.572 ICD.574
ICD.155     0.0    20.0    14.0    12.0     9.5    12.0    22.5     2.0
ICD.157    20.0     0.0    12.5     5.5    14.0     3.0     5.0     2.5
ICD.199    14.0    12.5     0.0     2.0     7.5     2.0     3.0     1.5
ICD.286    12.0     5.5     2.0     0.0    16.5     4.5     7.5     1.5
ICD.287     9.5    14.0     7.5    16.5     0.0     6.5     7.5     1.5
ICD.571    12.0     3.0     2.0     4.5     6.5     0.0    27.5     2.5
ICD.572    22.5     5.0     3.0     7.5     7.5    27.5     0.0     2.5
ICD.574     2.0     2.5     1.5     1.5     1.5     2.5     2.5     0.0
... ...
```

by the strength of interaction between the data types. So the multiplex network created using POGO is formed of three layers using the mRNA, pathway and clinical data. Each layer provided information on the same diseases. This result is a 30×30 multiplex matrix, since a multiplex matrix is formed of $n \times h$ rows and columns where n is the number of patients and h is the number of layers. Our software POGO can find the disease comorbidities by integrating all the descriptive layers, taking into account the properties of the multiplex. All these three categories association data are used as input of our multiplex network and predicted the integrated disease comorbidities network as shown in the **Figure 8D**.

Comorbidity Mapping

Patient medical records contain important clarification regarding the co-occurrences of diseases affecting the same patient. Two diseases are connected if they are co-expressed in a significant number of patients in a population (Hidalgo et al., 2009). To estimate the correlation starting from disease co-occurrence, we need to quantify the strength of the comorbidity risk. We used two comorbidity measures to quantify the strength of comorbidity associations between two diseases: (i) the Relative Risk (fraction between the number of patients diagnosed with both diseases and random expectation based on disease prevalence) as the quantified measures of comorbidity tendency of two disease pairs; and (ii) ϕ-correlation (Pearsons correlation for binary variables) to measure the robustness of the comorbidity association (Moni and Lio, 2014). We used the relative risk RR_{ij} and ϕ-correlation ϕ_{ij} of observing a pair of diseases i and j affecting the same patient. The RR_{ij} allows us to quantify the co-occurrence of disease pairs compared with the random expectation. When two diseases co-occur more frequently than expected by chance, we will get $RR_{ij} > 1$ and $\phi_{ij} > 0$. The two comorbidity measures are not completely independent of each other. We included links between disease pairs for which the co-occurrence is notably greater than the random expectation based on population prevalence of the diseases. Clinical information is from the http://www.icd9data.com in the ICD-9-CM format and collected from Hidalgo et al. (2009). The function `comorbidityMap` of POGO package is able to take input an OMIM id/3 or 5 digit ICD-9-CM code of a disease or a list of gene symbols/Entrez ids and provides comorbidity map of the patient based on the relative risk and ϕ-correlation. `comorbidityMap` requires two parameters: id list and id type. An example and its output is given in **Figure 9**.

Methods

Diseases are connected when they share at least one significant dysregulated gene/miRNA/SNP/CNV/GO/phenotype or environmental factor. Let a specific set of associated diseases D and a set of significant biomarker genes G, gene-disease associations attempt to find whether gene $g \in G$ is associated with disease $d \in D$. If G_i and G_j are the sets of significant up and down dysregulated genes associated with diseases i and j respectively then the number of shared dysregulated genes (n_{ij}^g) associated with both diseases i and j is as follows:

$$n_{ij}^g = N(G_i \cap G_j) \tag{1}$$

We calculated the similarity between a pair of diseases based on the number of entities (gene, SNP, CNV, miRNA, HPO or environmental factor) that shared between them. For an instance, in case of gene-disease association, we generated a list of genes known to be associated with each disease, and the disease similarity (association) was calculated based on how many genes are shared between a pair of diseases. The similarity is defined as

$$Sim(i, j) = \frac{N(G_i \cap G_j)}{\sqrt{N(G_i)} * \sqrt{N(G_j)}}, \tag{2}$$

where $N(G_i)$ and $N(G_j)$ are the number of genes linked to disease i and j respectively, and $N(G_i \cap G_j)$ is the number of genes associated to both disease i and j. SNP-sharing, CNV-sharing, miRNA-sharing, HPO-sharing and environmental factors were also generated with the same approach used for gene-sharing.

Hypergeometric test is implemented for enrichment analysis (Subramanian et al., 2005). It is used to assess whether the number of selected genes or ontology associated with disease is larger than expected. To determine whether any disease annotate a specified list of genes at frequency greater than what would be expected by chance, POGO calculates a p-value using the hypergeometric distribution. Significance of the enrichment analysis is assessed by the hypergeometric test and the p-value is adjusted by false discovery rate (FDR). The hypergeometric p-value is calculated using the following formula:

```
> comorbidityMap("042", "ICD9")
ICD.9.D1 ICD.9.D2 Prevalence.D1 Prevalence.D2 Co.occurrenceD1D2 RRij
"011"    "018"    16646         639           110               134.842507
"011"    "031"    16646         3693          807               171.170619
"011"    "042"    16646         1067          64                 46.984060
"011"    "112"    16646         141325        752                 4.168058
"011"    "117"    16646         9094          179                15.418178
 ... ...

 CI1          CI2         phi        t
131.740584  138.0174686 0.0334998   12.600646
170.628511  171.7144495 0.1024054   38.700702
 45.141791   48.9015140 0.0148728    5.591768
  4.153894    4.1822713 0.0118565    4.457522
 15.199244   15.6402660 0.0136184    5.120042
 ... ...
```

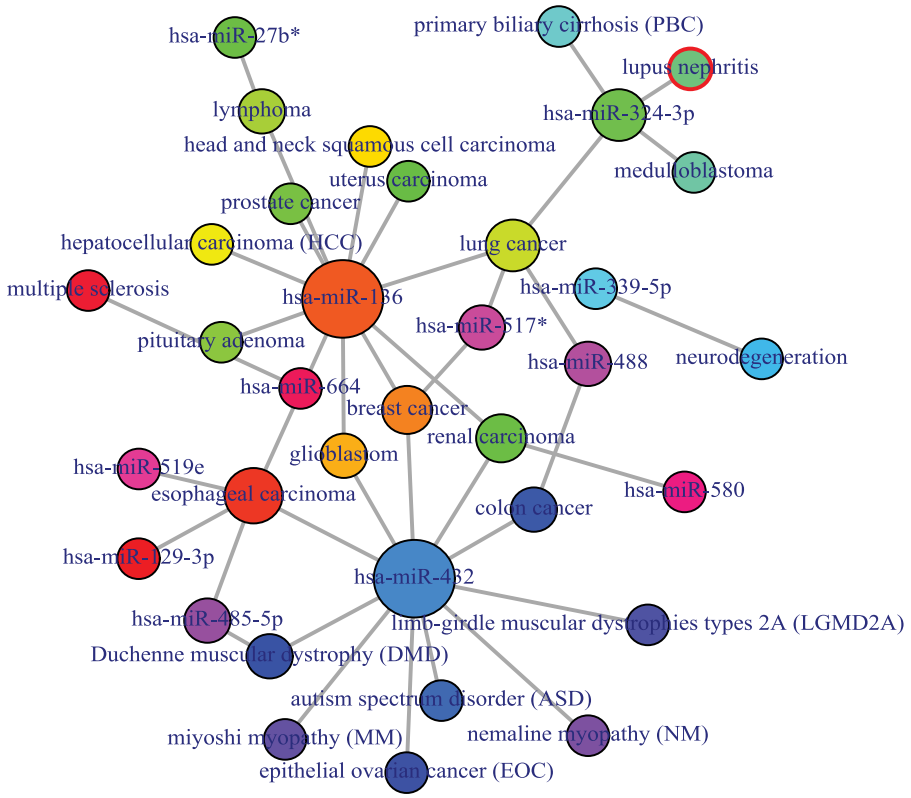


FIGURE 6 | Output figure and statistics of `> comorbiditymiRNA( "TNFRSF11A", "Symbol" ).` The gene Symbol TNFRSF11A is used as input to the `comorbiditymiRNA`. We show the comorbidities originated using the miRNA-disease associations information. The size of the nodes represents the degree of associations.

$$p - value = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad (3)$$

where N is the total number of reference genes, M is the number of genes that are associated to the disease of interest, n is the size of the list of genes of interest and k is the number of genes within that list which are associated to the disease. In case of GO term the p-value reports the likelihood of finding n genes annotated with a particular GO term in the set of interest by chance alone, given the number of genes annotated with that GO terms in the reference set. A biological process, molecular function or cellular location which are represented by a GO term is called enriched if the p-value is less than 0.05.

The co-occurrence indicates the number of common miRNAs/genes/ontology/SNPs/CNVs between two diseases. We applied the Jaccard index or Jaccard similarity coefficient, which is known as a standard method for comparing the similarity between two sets of entities. Each common neighbor is calculated based on the Jaccard Index method to calculate the strength of co-occurrence, where association score for a node pair is as:

$$Ass_{i,j} = \frac{N(G_i \cap G_j)}{N(G_i \cup G_j)} \quad (4)$$

We improved the performance of the association scores based on the Adamic and Adar measure (Adamic and Adar, 2003), which weights the impact of neighbor disease nodes inversely with respect to their total number of connections as follows:

$$AssScore(i,j) = \sum_{n \in N(G_i \cap G_j)} \frac{1}{log(degree(n))} \quad (5)$$

This inverse frequency technique is based on the principle that rare relationships are more specific and have more impact on the disease association.

Finally POGO calculates disease-disease interaction score. The score refers to the strength of the interaction between the diseases based on the protein interaction. The interaction score (ϕ_{ij}) is assigned for each disease pair i and j as follows:

$$\phi_{ij} = log(n_{ij}^{g} * N + Z) - \log(NG_i * NG_j + Z) \quad (6)$$

Here, NG_i and NG_j are the total number of genes for the disease, i and j, respectively. n_{ij}^{g} is the total number of common genes between the two diseases. N is the size of entire proteins involved in the disease protein network. Z is a constant ($Z = 1$) introduced to avoid out-of bound errors, if $NG_i = NG_j = n_{ij}^{g} = 0$.

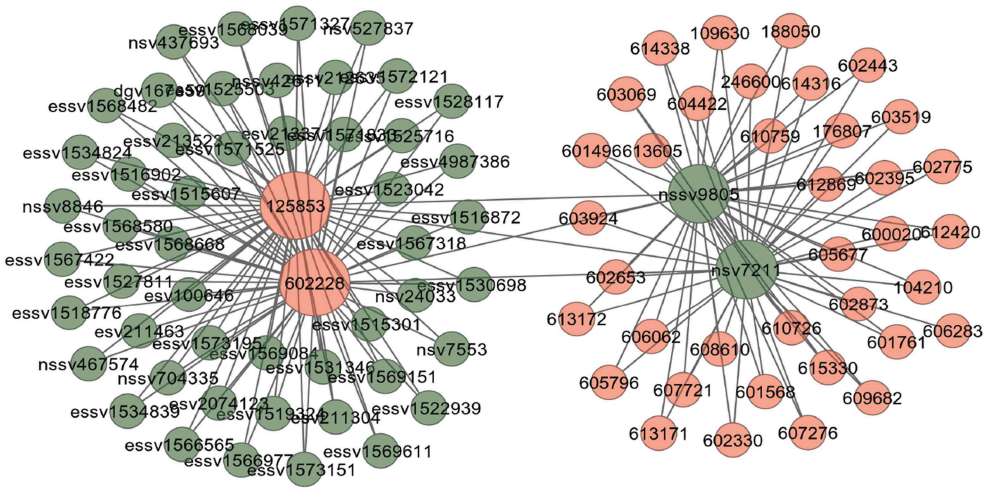


FIGURE 7 | Output figure and statistics of `> comorbidityCNV ("602228", "OMIM").` The OMIM disease id of the "Type 2 Diabetes mellitus" is 602228, which is used as input to the `comorbidityCNV`. We show disease comorbidity for the "Type 2 Diabetes mellitus" through the CNVs-disease associations. Here the light red color nodes represent the OMIM disease ids and light green color nodes represent the CNVs ids. The size of the nodes represents the degree of associations.

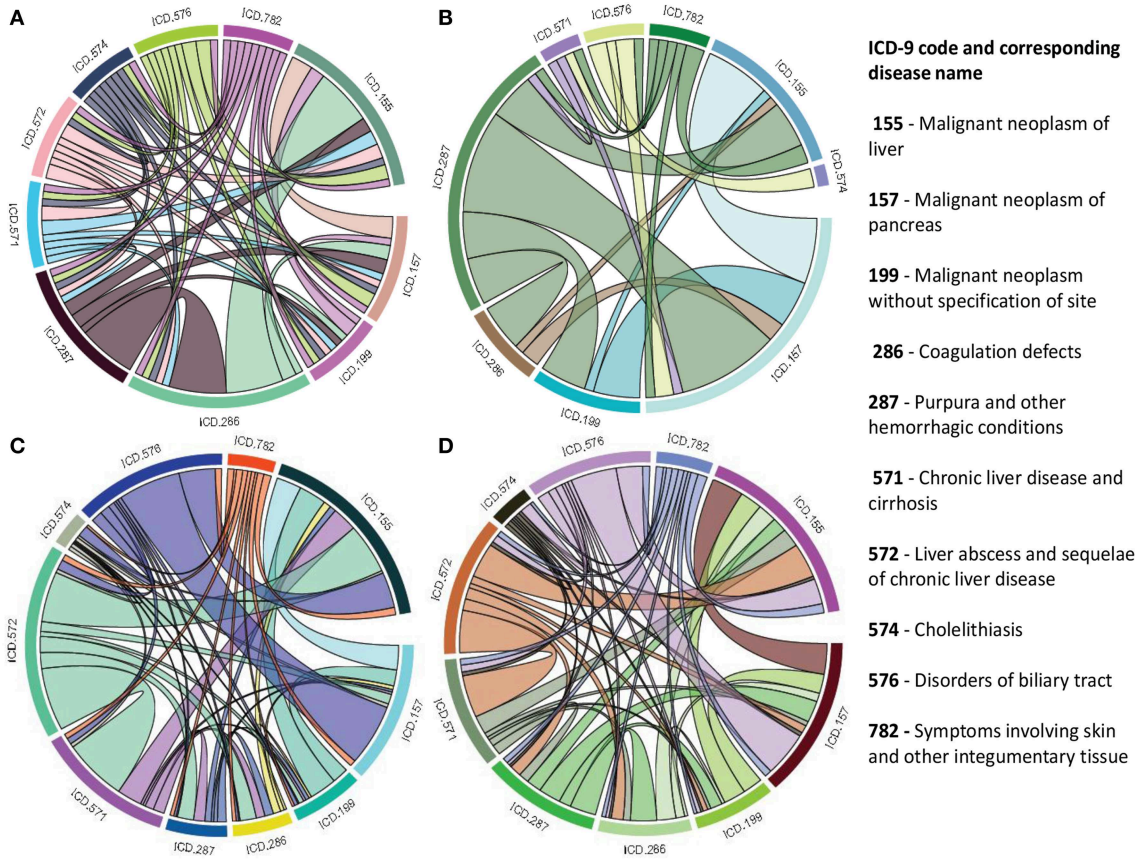
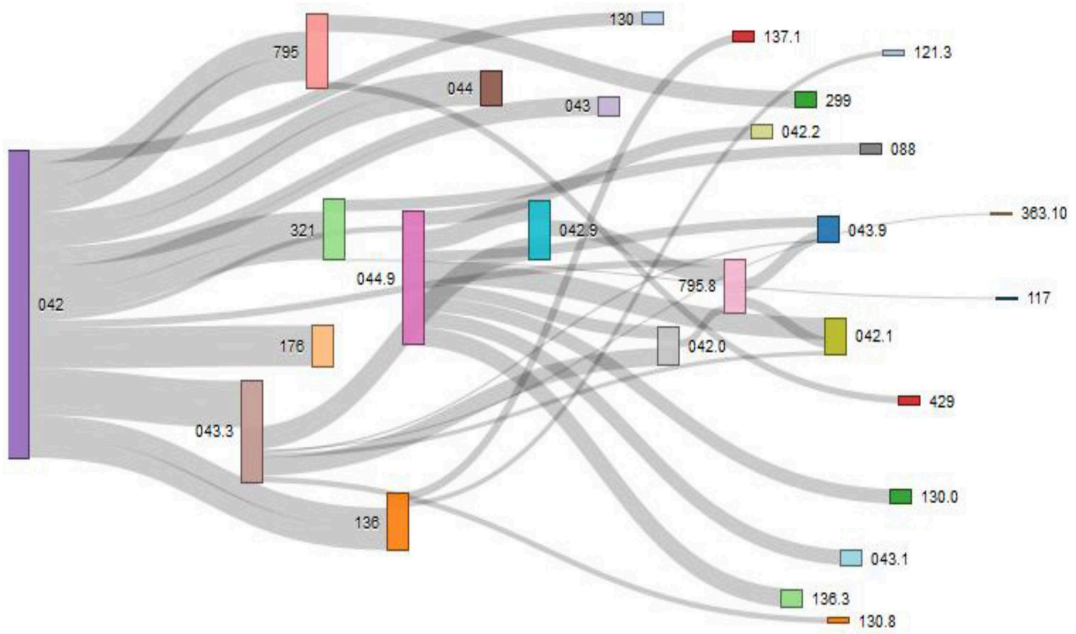


FIGURE 8 | Disease comorbidities network are constructed by applying the multiplex network model. Each disease is denoted by the ICD-9-CM code. **(A)** is a comorbidity association network based on the gene disease association data. **(B)** is a comorbidity association network based on the pathway disease association data. **(C)** is a comorbidity association network based on the clinical information. **(D)** is a comorbidity association network based on the integrated multiplex network output of the input of **(A–C)** as layers of the model.



ICD-9 code and corresponding disease name

042 - HIV infection
042.0 - HIV with specified infections
042.2 – HIV with specified malignant neoplasm
042.9 - Acquired immunodeficiency syndrome, unspecified
043 - HTLV-III/LAV infection
043.1 - HTLV-III/LAV infection causing specified diseases of the central nervous system
043.3 - HTLV-III/LAV infection causing other specified conditions,
043.9 Acquired immunodeficiency syndrome-related complex with or without other conditions
044 - Other HTLV-III/LAV conditions
044.9 - HTLV-III/LAV infection, not otherwise specified
088 - Arthropod-borne diseases
117 – Mycoses
121.3 - Fascioliasis
130 – Toxoplasmosis
130.0 - Meningoencephalitis due to toxoplasmosis
130.8 - Multisystemic disseminated toxoplasmosis
136 - Unspecified infectious & parasitic diseases
136.3 – Pneumocystosis
137.1 - Late effects of central nervous system tuberculosis
176 - Kaposi's sarcoma
299 - Pervasive developmental disorders
321 - Type 2 diabetes mellitus
363.10 - Disseminated chorioretinitis
429 - Ill-defined descriptions and complications of heart disease
795 - Nonspecific abnormal cytological, histological, immunological, and dna test findings
795.8 - Abnormal tumour markers

FIGURE 9 | Output figure and statistics of `>comorbidityMap ("042", "ICD9")`. The icd-9-CM code of the HIV is 042, which is used as input to the `comorbidityMap`. We show disease comorbidity for the HIV infection (042) with other diseases, whose ICD-9-CM codes are 042.0 (with specified infections), 042.1 (causing other specified infections), 042.2 (with specified malignant neoplasms), 042.9 (acquired immunodeficiency syndrome, unspecified), 043 (HTLV-III/LAV infection), 043.1 (HTLV-III/LAV infection causing specified diseases of the central nervous system), 043.3 (HTLV-III/LAV infection causing other specified conditions), 043.9 (acquired immunodeficiency syndrome-related complex with or without other conditions), 044 (Other HTLV-III/LAV conditions), 044.9 (HTLV-III/LAV infection, not otherwise specified), 088 (arthropod-borne diseases), 117 (mycoses), 121.3 (fascioliasis), 130 (toxoplasmosis), 130.0 (meningoencephalitis due to toxoplasmosis), 130.8 (multisystemic disseminated toxoplasmosis), 136 (unspecified infectious and parasitic diseases), 136.3 (pneumocystosis), 137.1 (late effects of central nervous system tuberculosis), 176 (Kaposi's sarcoma), 299 (pervasive developmental disorders), 321 (type 2 diabetes mellitus), 363.10 (disseminated chorioretinitis), 429 (ill-defined descriptions and complications of heart disease), 795 (nonspecific abnormal cytological, histological, immunological, and dna test findings), and 795.8 (abnormal tumor markers). POGO uses color rectangle to classify different disease codes and the size of the rectangle is used to represent the severity of that disease.

The expected result of ϕ_{ij} is positive, when the disease pair is over-represented and negative, when the disease pair is under-represented. Co-occurrence also indicates the number of shared patients. So, we used weighting scheme to avoid the bias based on disease prevalence. The mutual information weight $W(d_i, d_j)$ between two diseases d_i and d_j is defined as

$$W(d_i, d_j) = log\left(\frac{p(d_i, d_j)}{p(d_i) * p(d_j)}\right) \quad (7)$$

where the numerator is the observed co-occurrence (joint probability) and the denominator is the random expectation of co-occurrence (product of marginal probabilities).

The use of semantic similarity between biological processes to estimate disease association could enhance the identification and characterization of disease association besides identifying novel biological processes involved in the diseases. Graph-based methods using the topology of GO graph structure is used to compute semantic similarity. We adapted the approach for computing the functional similarity of GO terms from Wang et al. (2007, 2010). Semantic values of GO term are measured according to the DAG of corresponding disorders. Semantic similarity for any pair of GO term is calculated based on disease semantic value. Formally, a GO term a can be represented as a graph $DAG_a = (a, T_a, E_a)$, where T_a is the set of all GO terms in DAG_a, including term a itself and all of its ancestor terms in the GO graph, and E_a is the set of corresponding edges that connect the GO terms in DAG_a. To encode the semantic of a GO term in a measurable format to enable a quantitative comparison, Wang firstly defined the semantic value of term a as the combined contribution of all terms in DAG_a to the semantics of term a (Wang et al., 2007). Terms closer to term a in DAG_a contribute more to its semantics (Wang et al., 2010). Thus, the contribution of a GO term t in DAG_a is defined to the semantics of GO term a as the S value of the term t related to term a, $S_a(t)$, which can be calculated as:

$$S_a(t) = \begin{cases} S_a(a) = 1 \; if \; t = a \\ S_a(t) = max\{w_e * S_a(t') | t' \in \; children \; of \; (t)\} \; if \; t \neq a \end{cases} \quad (8)$$

where w_e is the semantic contribution factor for edge e ($e \in E_a$) linking term t with its child term $t^{'}$. It is assigned between 0 and 1 according to the types of associations. Term a contributes to its own is defined as one. Then the semantic value of GO term a, $SV(a)$ and the semantic value of GO term b, $SV(b)$ are calculated as:

$$SV(a) = \sum_{t \in T_a} S_a(t), \quad SV(b) = \sum_{t \in T_b} S_b(t) \quad (9)$$

Thus, for the given two GO terms a and b, the semantic similarity between these two terms is defined as:

$$S_{sim}(a, b) = \sum_{t \in T_a \cap T_b} \frac{S_a(t) + S_b(t)}{SV(a) + SV(b)} \quad (10)$$

where $S_a(t)$ is the semantic value of term t related to GO term a and $S_b(t)$ is the semantic value of GO term t associated to GO term b. The semantic similarity between two sets of GO terms A and B is calculated as

$$Sim(A, B) = \frac{1}{|A| + |B|}\left(\sum_{a \in A} Sim(a, B) + \sum_{b \in B} Sim(b, A)\right) \quad (11)$$

where $|A|$ and $|B|$ represent the numbers of terms in sets A and B respectively.

To obtain more insight into the shared risk factors mechanism of associated human genetic diseases, mapping was implemented from disease phenotype to gene based on the disease-gene association. With the integration of huge numbers and diverse set of experimental data, prediction of gene-phenotype interactions has emerged as a very productive subfield with great importance for the understanding of human disease. Given a specific set of human phenotype D, a set of human genes G and evidence E, these approach attempt to find whether gene $g \in G$ is associated with phenotype $d \in D$. It is notable that E could be gene-disease associations obtained through genetic studies. To quantitatively explore the phenotypic similarity between different phenotype records P_i and P_j, according to Zhang et al. (2010) we defined the association measure as cosine of the angle between their corresponding phenotype feature vectors using the following formula:

$$Sim(P_i, P_j) = \frac{\sum_{k=1}^{N} w_{k,i} * w_{k,j}}{\sqrt{\sum_{k=1}^{N}(w_{k,i})^2} * \sqrt{\sum_{k=1}^{N}(w_{k,j})^2}} \quad (12)$$

where N is the total mapping concepts, $w_{k,i}$ and $w_{k,j}$ were the k-th term, weight in phenotype record P_i and P_j, respectively.

For each of the phenotype clusters, mapping was implemented from disease phenotypes to their associated disease genes based on the disease-gene association list in the GAD and OMIM databases. Therefore, we can get the corresponding gene subsets mapped to different phenotype clusters. OMIM disease ids were mapped to the hierarchy of HPO to retrieve the matched HPO terms. Then, a new HPO similarity is calculated for each pair of phenotypes by Jaccard similarity Index

$$Sim_{HPO} = \frac{|P1 \cap P2|}{|P1 \cup P2|} \quad (13)$$

where $P1$ and $P2$ are the set of the matched HPO terms of the two phenotypes, respectively.

The way to assign terms to objects is to add annotations. In our case, the entities represent genes and terms corresponding to phenotypes (HPO terms) or biological processes (GO terms). The specificity of the terms associated with genes allows us to calculate the most significant relationships between them, which use to be related to its proximity to the root.

Each disease is generally mapped to multiple phenotypic features. In order to compute associations between two diseases, $d1$ and $d2$, we adapt a method previously developed for estimating protein similarity with GO (Pesquita et al., 2008),

where each feature of $d1$ is matched with the most similar feature of $d2$ and the average is taken over all such pairs of features:

$$sim(d1 \rightarrow d2) = avg\left[\sum_{s \in d1} \max_{t \in d2} sim(s, t)\right] \quad (14)$$

Equation (14) is not symmetric with respect to $d1$ and $d2$, the final similarity metric is defined as the mean of Equation (14) taken in both orientations:

$$sim(d1, d2) = \frac{1}{2} * sim(d1 \rightarrow d2) + \frac{1}{2} * sim(d2 \rightarrow d1) \quad (15)$$

This metric is used to indicate the similarity between two disorders, each of which is mapped to multiple HPO terms.

Multiplex Network Model for Data Integration

We developed multiplex network model to integrate diverse set of omics and clinical data to predict disease comorbidities. It is a special type of multilayered network which is called the multiplex network, in which the same nodes are present in all layers, i.e., $V_1 = V_2 = ... = V_M = V$ and where nodes can only have interlayer connections to their counterpart nodes, i.e., $E_{\alpha\beta} = (v, v); v \in V$ for all $\alpha, \beta \in 1, ..., M, \alpha \neq \beta$ (Boccaletti et al., 2014).

Let's consider that we have a set of associated diseases. Each pair of diseases has different types of associated data describing them in some way. In each data type, diseases have some level of association to each other and each data type has a level of dependency or interaction. Each layer in the multiplex represents a particular type of data with each node representing a disease in each layer of the multiplex. The edges between nodes in each layer represent a measure of association between diseases in corresponding to the level of similarity between diseases for the particular data type which the layer represents. The strength of interaction between each data type can be modeled by a weight connecting each layer in the multiplex. **Figure 10** shows an example with three layers (data types) and four diseases. In this case we can model the association among diseases in a multiplex network that can be represented in a matrix as follows:

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A}_1 & \omega_{12}\boldsymbol{I} & \dots & \omega_{1h}\boldsymbol{I} \\ \omega_{21}\boldsymbol{I} & \boldsymbol{A}_2 & \dots & \omega_{2h}\boldsymbol{I} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{h1}\boldsymbol{I} & \omega_{h2}\boldsymbol{I} & \dots & \boldsymbol{A}_h \end{pmatrix}, \quad (16)$$

where h is the number of layers, A_i is the adjacency matrix of layer i, ω_{ij} is the interlayer interaction strength from layer i to j and I is the corresponding identity matrix. The strength between layers in the multiplex, ω, represents a measure of dependency or strength of interaction between the layers. The edge weights between nodes represent a measure of similarity between nodes in the same layer, normalized between zero and one. Therefore, it is natural for the values of ω to represent a measure of dependence between zero and one, where zero and one indicate independence and total dependence between the layers respectively. In our case the strength of interaction is undirected and symmetric, i.e., $\omega_{i,j} = \omega_{j,i}$.

To compute an overall disease similarity between patients given all sets of data, we can find the disease similarity by aggregating the descriptive layers in some way, taking into account the properties of the multiplex. Estrada and Gómez-Gardeñes (2014) defined the aggregate network, $\hat{G}$, of a multiplex network as follows. Let $G_1 = (V_1, E_1)$, $G_2 = (V_1, E_2)$, ..., $G_h = (V_1, E_h)$ be the set of layers in the multiplex. Then $\hat{G} = (\hat{V}, \hat{E})$ where $\hat{V} = V_1$ and $\hat{E} = \cup_{i=1}^{h} E_i$. In other words, the aggregate is defined as the union of all edges across all layers of the multiplex. In the literature, the aggregate of a multiplex is often defined in this way. This method can aggregate layers of a multiplex in which the layers are unweighted graphs. However, it is not sufficient for a weighted graph, particularly a complete weighted graph. In addition, the strengths between layers are not accounted for.

Let's consider that the edge weights between nodes provides a normalized measure of similarity between zero and one. We can define the weight of a path between two nodes in the multiplex to be the product of the edges between each node in each step of the path. Since the weight between nodes is a measure of similarity or information shared between the nodes, it follows that the weight of the path provides a measure of information flowing through the path.

There are a number of ways we can provide a new measure of similarity between two nodes given the properties of the

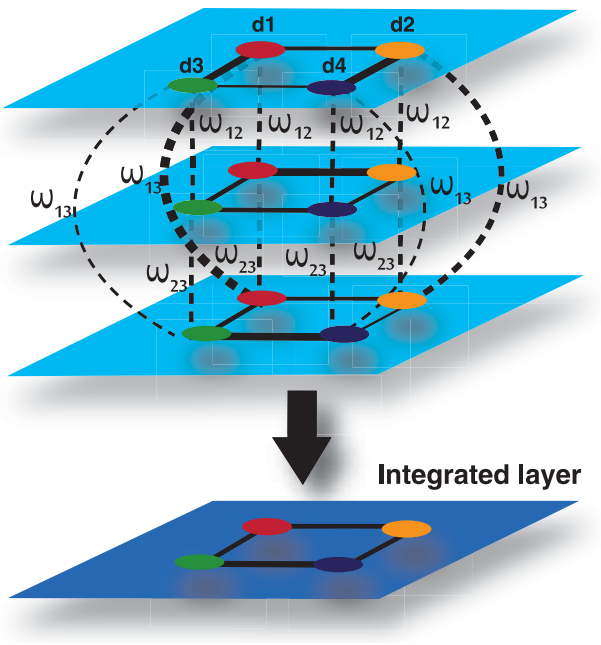


FIGURE 10 | Multiplex formed by three input layers, each representing a data type, and four nodes, each representing a disease. The 4th layer is an output layer, which is an integrated layer of the 3 input layers.

multiplex network. One way would be to take the mean of the direct paths connecting each patient to and from another patient in each and every layer. We defined this mathematically as follows:

$$R_{direct} = \frac{\sum_{i=1}^{h}(\boldsymbol{M}|_{p_iq_i} + \sum_{j=1,j\neq i}^{h}\boldsymbol{M}^2|_{p_iq_j})}{h^2}, \quad (17)$$

Where h is the number of layers in the multiplex, $M|_{p_iq_i}$ is the element in the multiplex matrix representing the weight between node p and q in layer i and $M^2|_{p_iq_j}$ is the element in the square of the multiplex network, representing the weight of the path from node p in layer i to node q in layer j. Another way would be to take the maximum or minimum information shared directly between two nodes.

$$R_{direct_{min}} = \min_{i=1}^{h}\left(\boldsymbol{M}|_{p_iq_i} + \sum_{j=1,j\neq i}^{h}\boldsymbol{M}^2|_{p_iq_j}\right) \quad (18)$$

$$R_{direct_{max}} = \max_{i=1}^{h}\left(\boldsymbol{M}|_{p_iq_i} + \sum_{j=1,j\neq i}^{h}\boldsymbol{M}^2|_{p_iq_j}\right) \quad (19)$$

In many situations, a pair of nodes in a network does not communicate only through the shortest-path routes connecting both nodes, but also through all possible routes connecting both nodes. The number of these possible routes can be enormous. Moreover, the information can also go back and forth before connecting the pair of nodes. Network communicability, which was introduced by Estrada and Gómez-Gardeñes (2014), attempts to quantify such correlation effects in the communication between nodes in complex networks. Estrada and Gomez-Gardenes defined communicability as a measure that "quantifies the number of possible routes that two nodes have to communicate with each other." In multiplex networks, the communicability, C, between two nodes p and q, is a weighted sum of all walks from p to q.

$$C_{pq} = \boldsymbol{I} + \boldsymbol{M} + \frac{\boldsymbol{M}^2}{2!} + \ldots = \sum_{k=0}^{k}\frac{\boldsymbol{M}^k}{k!}\Big|_{pq}. \quad (20)$$

Hence, the communicability between nodes p and q is given by:

$$C_{pq} = [e^{(\mathbf{A}_L + \mathbf{v}_{LL})}]_{pq} = [e^{\mathbf{M}}]_{pq}, \quad (21)$$

where the p,q-th entry in the minor, $\boldsymbol{C}$, defines the communicability broadcasted from node p in layer i to node q in layer j. Therefore, the communicability broadcasted and received by the nodes in the multiplex is given by:

$$\boldsymbol{C} = e^{(\mathbf{A}_L + \mathbf{v}_{LL})} = \begin{pmatrix} C_{11} & C_{12} & \ldots & C_{1h} \\ C_{21} & C_{22} & \ldots & C_{2h} \\ \vdots & \vdots & \ddots & \ddots \\ C_{h1} & C_{h2} & \ldots & C_{hh} \end{pmatrix} \quad (22)$$

Since all nodes are present in each layer of the multiplex, we can calculate the integrated communicability from node p and q in all layers in the multiplex by taking the harmonic mean of the communicability between them in each minor in the matrix $\boldsymbol{C}$.

$$\hat{C}_{pq} = \frac{h}{\sum_{i=1}^{h}\frac{1}{[C_{i,i}]_{pq}} + \sum_{j,k=1,j\neq k}^{h}\frac{1}{[C_{jk}]_{pq}}}. \quad (23)$$

Hence, the integrated communicability matrix is formed by:

$$\hat{\mathbf{C}} = \begin{pmatrix} 0 & \hat{C}_{12} & \ldots & \hat{C}_{1h} \\ \hat{C}_{21} & 0 & \ldots & \hat{C}_{2h} \\ \vdots & \vdots & \ddots & \ddots \\ \hat{C}_{h1} & \hat{C}_{h2} & \ldots & 0 \end{pmatrix}, \quad (24)$$

where $\hat{C}_{ij}$ represents the interaction of layer i with layer j. Therefore, this multiplex network model is applicable to integrate omics and clinical information of a number of diseases or patients in an efficient way.

Evaluation

We incorporated verified data from different data source with our software. Data integration reduces noise associated with each experimental limitation, thus increases sensitivity and specificity to detect true association relationships which results in less number of false positives. By integrating different types of omics and clinical data can produce more reliable predictions with increased sensitivity and specificity for detecting true functional disease comorbidity associations. This can help in finding the hidden connections between complex diseases. Such connections between complex diseases reflect common biological pathways and biological functions that may become manifest in the form of comorbidity. For an example, we show a comparative representation of dysregulated genes and lifestyle impact on the disease comorbidity in **Figure 11**. Here, panel A (see **Figure 11A**) represents an example of only dysregulated genetic influences on diseases with good lifestyle. Panel B (see **Figure 11B**) shows an example of only bad lifestyle influences on diseases with no genetic variation. Panel C (see **Figure 11C**) represents the combined impacts of lifestyle and dysregulated genes on diseases. Here, we observed that the combined impact of both lifestyle and dysregulated genes influences more and multiway on the diseases and disease comorbidities. It is conceivable that by integrating the data ranging from genotype to multiple levels of phenotypes, more precise and robust stratification of the patients with clinical outcome difference can be achieved.

Discussion

Development of methods combining omics, ontology and clinical information could assist clinical decision making and

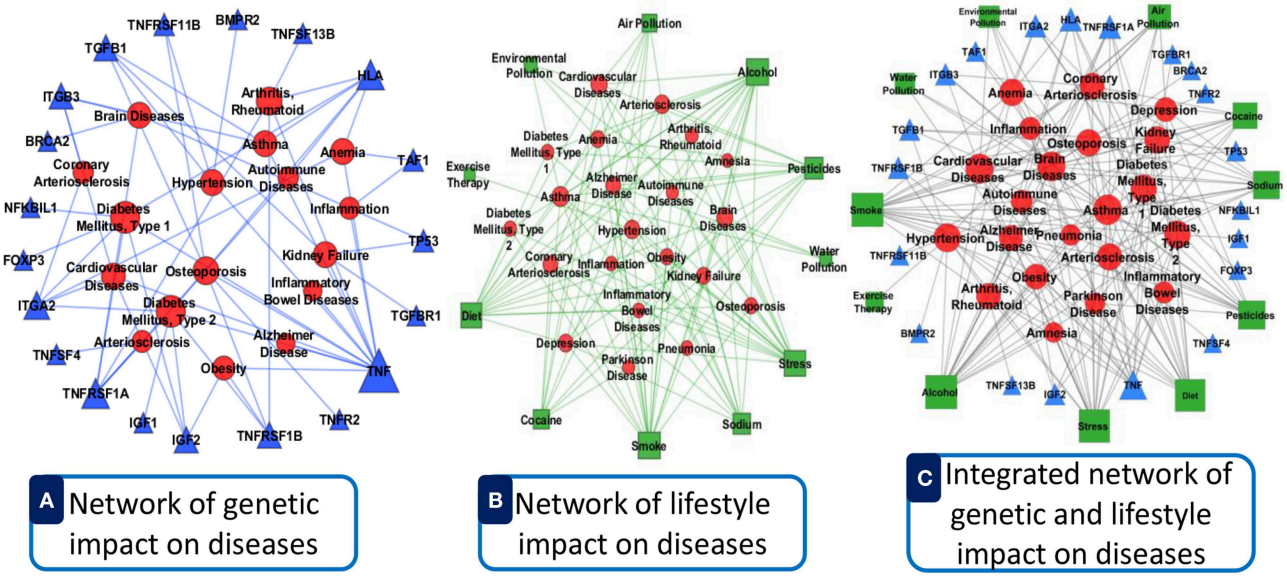


FIGURE 11 | Comparative representation of dysregulated genes and lifestyle impact on the disease comorbidity. (A) Represents an example of only dysregulated genetic influences on diseases with good lifestyle. **(B)** Shows an example of only bad lifestyle influences on diseases with no genetic variation. **(C)** Represents the combined impacts of lifestyle and dysregulated genes on diseases. Here, the circular red nodes nodes represent diseases name, triangular blue nodes represent genes symbols and square green nodes represent lifestyle factors. The size of the nodes represents the degree of associations.

represent a large step toward personalized medicine. Proactive and personalized medicine will bring fundamental changes to health care, taking carefully targeted preventative or therapeutic action at the earliest indications of risk or disease. In order to facilitate the necessary changes, better tool is needed for assessing risk and optimizing treatments, which in turn require better understanding of disease interdependencies, genetic influence, and translation into a patient's future. However, most software is designed to make a prediction about a single disease or a class of some specific diseases based on the single omics or clinical information. Phenomizer is a web-based system that produces a ranked list of hereditary diseases, taking a set of clinical features (Köhler et al., 2009). This system only considers the phenotypic annotation to diseases, and semantic similarity metrics to measure phenotypic similarity between query phenotypes and disease phenotypes with the use of the HPO (Robinson and Mundlos, 2010). Another software DGFinder which is used to assess candidate genes in interested chromosome regions for their possibility relating to a given disease (Yuan et al., 2010). It integrated a dataset containing 1045 genes related to 305 diseases. Hidalgo et al. analyzed comorbidity associations using the medical records (Hidalgo et al., 2009). There are some online information retrieval tools, such as AmiGO[4] and QuickGO[5], to collect gene annotation data from various databases and manually discover the correlations or similarities of gene products by their biological functions (Binns et al., 2009). FindZebra (Dragusin et al., 2013) is a vertical search engine for rare diseases. This system does not consider the genetic effects on disease or phenotypic effects on genes rather it presents a list of disease documents for a given query of symptoms. CARE uses collaborative filtering methods to predict each patient's disease risks based only on their own medical history and that of similar patient's information (Davis et al., 2010). Recently, a tool KnIT has been developed for the complete medical literature knowledge integration (Spangler et al., 2014). DisGeNET is a coherent tool that analyses and interprets human gene network to disease network (Bauer-Mehren et al., 2010). It is able to display gene-disease association networks as bipartite graphs and provides gene centric and disease centric views of the data.

An R package "comorbidities" is able to categorize ICD-9-CM codes based on published 30 comorbidity indices using Deyo adaptation of Charlson index and the Elixhauser index (Deyo et al., 1992; Elixhauser et al., 1998). Our previous R package comoR that provides relative risk, ϕ-correlation, associated genes and pathway between the comorbidity diseases (Moni and Lio, 2014). It is limited to gene expression and pathway molecular data. To our knowledge, there is no available complete software tool for the prediction of disease comorbidities maps based on the multiple omics, gene ontology, phenotype and environmental influences. So, we developed POGO, another R package that implements different statistical approach for the prediction of disease comorbidity maps by integrating diverse set of data. This software could provide comorbidity mapping among all diseases using ontology, miRNA, SNPs, CNVs, phenotypic and environmental information. This software also incorporated a prediction model that explores the past medical patient

[4]http://www.godatabase.org.
[5]http://www.ebi.ac.uk/ego/.

history to determine the risk of patients to develop future diseases.

Patient's omics data is becoming important for clinical decision making, including disease risk assessment, disease diagnosis and subtyping, drug therapy and dose selection (Ullman-Cullere and Mathew, 2011). In the near future, physician will have to consider omics implications to patient care throughout their clinical work flow, including electronic prescribing of medications. In the not-so-distant future, as we move in to an era of personalized and preventive medicine, healthy individuals may be tracked by multiple layers of omic and clinical data in an effort to track potential disease progression. Our software tool incorporated an integrated framework to establish the associations between genetic diseases and ontology information, which may help to uncover the molecular mechanisms of genetic diseases. The identified disease patterns from POGO could be useful for further investigations with regards to their diagnostic utility or help in the prediction of novel therapeutic targets. Therefore, POGO could be helpful for the personalized medicine system. They are able to detect many diseases at the earliest detectable phase, weeks, months, and maybe years before symptoms appear. POGO could easily be integrated into pipelines for high-throughput analysis, such as Galaxy, and other gene expression data mining, protein interactions validation, predicting causal relationships among phenotypes and miRNA-regulated network interpretation. The underlying hypothesis behind this line of research is that once we catalog all disease-disease relations through the omics, ontology, phenotypic and environmental influence, we will be able to predict the susceptibility of each individual to future diseases using various molecular biomarkers, ushering us into an era of predictive medicine.

Thus, a combination of genetic, ontology and population-level data and information could be analyzed by this software tool to establish and study novel hypotheses about unknown disease mechanisms and disease comorbidity. Understanding how different diseases relate to each other will not only provide us with a global view of disease associations, but also provide potentially new insights into the etiology, classification, and design of novel therapeutic interventions. This has led to the advent of stratified medicine, which translates advances in basic research by targeting etiological mechanisms underlying diseases. Method and tool for stratifying (classifying) patients in order to reliably predict prognosis or success of treatments are of critical importance in the field of medicine. However, with the identification of the new omics and clinical information, we need to update the integrated databases of the POGO. Using the temporal data explored by the time dimension approach, POGO could be extended to predict the time of expected disease diagnosis in addition to the likelihood of occurrence. The result is a patient stratification could be based on more complete profiles than the primary diagnosis. Therefore, POGO is useful for the stratified medicine.

Conclusion

Integration of multi-omics, ontology and phenotypic information is important for comorbidity prediction and patient stratification. Therefore, our methodological framework and software for integrating genetic and clinical data could be applicable in clinical decision making for personalized medicine. We expect that this combined approach may increase accuracy and decrease effort for disease comorbidity diagnosis. POGO software tool provides robust approaches to study disease comorbidity mappings by integrating omics, phenotype and ontology information, which can be easily integrated into pipelines for high-throughput and clinical data analysis, and to predict causal inference of a disease. This software tool will help to gain a better understanding of the complex pathogenesis of disease risk phenotypes and the heterogeneity of disease comorbidities. Moreover, the disease comorbidity patterns identified using this software tool could be useful for diagnostic utility or to help in the prediction of novel therapeutic targets. Thus, this software tool could be applicable in personalized medicine and clinical bioinformatics. So our software tool for comorbidity diagnosis and patient stratification could result in effective aids to the health practice. This will not only result in improving health outcomes of the patient, but also in reducing the health care costs.

Availability and Requirements

The software package POGO has been written in the platform independent R programming language. It requires R version 2.16 or newer to run. The software is freely available at www.cl.cam.ac.uk/~mam211/POGO/ and will appear in Comprehensive R Archive Network (CRAN) at (http://cran.r-project.org/).

Funding

This work is supported by the EU Mission T2D project.

Acknowledgment

PL thanks Claudio Franceschi, Paolo Garagnani and Filippo Castiglione for interesting suggestions.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fcell.2015.00028

Table S1 | Data of disease-disease associations based on the shared genes.

Table S2 | Data of disease-disease associations based on the shared pathways.

Table S3 | Data of disease-disease associations based on the clinical information.

References

Adamic, L. A., and Adar, E. (2003). Friends and neighbors on the web. *Soc. Netw.* 25, 211–230. doi: 10.1016/S0378-8733(03)00009-1

Astrup, A. (2001). Healthy lifestyles in europe: prevention of obesity and type ii diabetes by diet and physical activity. *Public Health Nutr.* 4, 499–515. doi: 10.1079/PHN2001136

Bae, J. S., Cheong, H. S., Kim, J.-H., Park, B. L., Kim, J.-H., Park, T. J., et al. (2011). The genetic effect of copy number variations on the risk of type 2 diabetes in a korean population. *PLoS ONE* 6:e19091. doi: 10.1371/journal.pone.0019091

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12, 56–68. doi: 10.1038/nrg2918

Bauer-Mehren, A., Rautschka, M., Sanz, F., and Furlong, L. I. (2010). Disgenet: a cytoscape plugin to visualize, integrate, search and analyze gene–disease networks. *Bioinformatics* 26, 2924–2926. doi: 10.1093/bioinformatics/btq538

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* 36, 431–432. doi: 10.1038/ng0504-431

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). Quickgo: a web-based tool for gene ontology searching. *Bioinformatics* 25, 3045–3046. doi: 10.1093/bioinformatics/btp536

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C., Gómez-Gardeñes, J., Romance, M., et al. (2014). The structure and dynamics of multilayer networks. *Phys. Rep.* 544, 1. doi: 10.1016/j.physrep.2014.07.001

Davis, D. A., Chawla, N. V., Christakis, N. A., and Barabási, A.-L. (2010). Time to care: a collaborative engine for practical disease prediction. *Data Mining Knowl. Discov.* 20, 388–415. doi: 10.1007/s10618-009-0156-z

Deyo, R. A., Cherkin, D. C., and Ciol, M. A. (1992). Adapting a clinical comorbidity index for use with icd-9-cm administrative databases. *J. Clin. Epidemiol.* 45, 613–619. doi: 10.1016/0895-4356(92)90133-8

Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jørgensen, H. L., Cox, I. J., et al. (2013). Findzebra: a search engine for rare diseases. *Int. J. Med. Inform.* 82, 528–538. doi: 10.1016/j.ijmedinf.2013.01.005

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Med. Care* 36, 8–27. doi: 10.1097/00005650-199801000-00004

Estrada, E., and Gómez-Gardeñes, J. (2014). Communicability reveals a transition to coordinated behavior in multiplex networks. *Phys. Rev. E* 89:042819. doi: 10.1103/PhysRevE.89.042819

Fredman, D., Siegfried, M., Yuan, Y. P., Bork, P., Lehväslaiho, H., and Brookes, A. J. (2002). Hgvbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.* 30, 387–391. doi: 10.1093/nar/30.1.387

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80. doi: 10.1186/gb-2004-5-10-r80

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8685–8690. doi: 10.1073/pnas.0701361104

Hidalgo, C. A., Blumm, N., Barabási, A.-L., and Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLoS Comput. Biol.* 5:e1000353. doi: 10.1371/journal.pcbi.1000353

Hirakawa, M., Tanaka, T., Hashimoto, Y., Kuroda, M., Takagi, T., and Nakamura, Y. (2002). Jsnp: a database of common gene variations in the japanese population. *Nucleic Acids Res.* 30, 158–162. doi: 10.1093/nar/30.1.158

Hu, G., and Agarwal, P. (2009). Human disease-drug network based on genomic expression profiles. *PLoS ONE* 4:e6536. doi: 10.1371/journal.pone.0006536

Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., et al. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* 5:4022. doi: 10.1038/ncomms5022

Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease micrornas through a human phenome-micrornaome network. *BMC Syst. Biol.* 4(Suppl. 1):S2. doi: 10.1186/1752-0509-4-S1-S2

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). mir2disease: a manually curated database for microrna deregulation in human disease. *Nucleic Acids Res.* 37(Suppl. 1), D98–D104. doi: 10.1093/nar/gkn714

Kahn, S. E., Hull, R. L., and Utzschneider, K. M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* 444, 840–846. doi: 10.1038/nature05482

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., et al. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85, 457–464. doi: 10.1016/j.ajhg.2009.09.003

Kozomara, A., and Griffiths-Jones, S. (2011). mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic Acids Res.* 39, D152–D157. doi: 10.1093/nar/gkq1027

Lee, D.-S., Park, J., Kay, K., Christakis, N., Oltvai, Z., and Barabási, A.-L. (2008). The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. U.S.A.* 105, 9880–9885. doi: 10.1073/pnas.0802208105

Lewis, S. N., Nsoesie, E., Weeks, C., Qiao, D., and Zhang, L. (2011). Prediction of disease and phenotype associations from genome-wide association studies. *PLoS ONE* 6:e27175. doi: 10.1371/journal.pone.0027175

Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). Hmdd v2. 0: a database for experimentally supported human microrna and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi: 10.1093/nar/gkt1023

Liu, Y. I., Wise, P. H., and Butte, A. J. (2009). The etiome: identification and clustering of human disease etiological factors. *BMC Bioinform.* 10(Suppl. 2):S14. doi: 10.1186/1471-2105-10-S2-S14

Lu, M., Zhang, Q., Deng, M., Miao, J., Guo, Y., Gao, W., et al. (2008). An analysis of human microrna and disease associations. *PLoS ONE* 3:e3420. doi: 10.1371/journal.pone.0003420

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., and Scherer, S. W. (2014). The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 42, D986–D992. doi: 10.1093/nar/gkt958

McCarroll, S. A., and Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nat. Genet.* 39, S37–S42. doi: 10.1038/ng2080

McKusick, V. A. (2007). Mendelian inheritance in man and its online version, omim. *Am. J. Hum. Genet.* 80, 588. doi: 10.1086/514346

Moni, M. A., and Lio, P. (2014). comor: a software for disease comorbidity risk assessment. *J. Clin. Bioinform.* 4:8. doi: 10.1186/2043-9113-4-8

Park, J., Lee, D.-S., Christakis, N. A., and Barabási, A.-L. (2009). The impact of cellular networks on disease comorbidity. *Mol. Syst. Biol.* 5, 262. doi: 10.1038/msb.2009.16

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E., Falcão, A. O., and Couto, F. M. (2008). Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinform.* 9(Suppl. 5):S4. doi: 10.1186/1471-2105-9-S5-S4

Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* 83, 610–615. doi: 10.1016/j.ajhg.2008.09.017

Robinson, P. N., and Mundlos, S. (2010). The human phenotype ontology. *Clin. Genet.* 77, 525–534.

Schadt, E. E. (2009). Molecular networks as sensors and drivers of common human diseases. *Nature* 461, 218–223. doi: 10.1111/j.1399-0004.2010.01436.x

Sherry, S. T., Ward, M.-H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbsnp: the ncbi database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Spangler, S., Wilkins, A. D., Bachman, B. J., Nagarajan, M., Dayaram, T., Haas, P., et al. (2014). "Automated hypothesis generation based on mining scientific literature," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM), 1877–1886.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Suthram, S., Dudley, J. T., Chiang, A. P., Chen, R., Hastie, T. J., and Butte, A. J. (2010). Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.* 6:e1000662. doi: 10.1371/journal.pcbi.1000662

Tong, B., and Stevenson, C. (2007). *Comorbidity of Cardiovascular Disease, Diabetes and Chronic Kidney Disease in Australia*. Australian Institute of Health and Welfare.

Ullman-Cullere, M. H., and Mathew, J. P. (2011). Emerging landscape of genomics in the electronic health record for personalized medicine. *Hum. Mutat.* 32, 512–516. doi: 10.1002/humu.21456

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241

Wang, J. Z., Du, Z., Payattakool, R., Philip, S. Y., and Chen, C.-F. (2007). A new method to measure the semantic similarity of go terms. *Bioinformatics* 23, 1274–1281. doi: 10.1093/bioinformatics/btm087

Yang, J. O., Hwang, S., Oh, J., Bhak, J., and Sohn, T.-K. (2008). An integrated database-pipeline system for studying single nucleotide polymorphisms and diseases. *BMC Bioinform.* 9(Suppl. 12):S19. doi: 10.1186/1471-2105-9-S12-S19

Yu, G., Wang, L. G., Yan, G. R., and He, Q. Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31, 608–609. doi: 10.1093/bioinformatics/btu684

Yuan, F., Wang, R., Guan, M., and He, G. (2010). "A novel computational method for predicting disease genes based on functional similarity," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence* (Springer), 42–51.

Zhang, F., Gu, W., Hurles, M. E., and Lupski, J. R. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi: 10.1146/annurev.genom.9.081307.164217

Zhang, S.-H., Wu, C., Li, X., Chen, X., Jiang, W., Gong, B.-S., et al. (2010). From phenotype to gene: detecting disease-specific gene functional modules via a text-based human disease phenotype network construction. *FEBS Lett.* 584, 3635–3643. doi: 10.1016/j.febslet.2010.07.038

Žitnik, M., Janjić, V., Larminie, C., Zupan, B., and Pržulj, N. (2013). Discovering disease-disease associations by fusing systems-level molecular data. *Sci. Rep.* 3:3202. doi: 10.1038/srep03202

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.