Title

DDOT: A Swiss Army Knife for Investigating Data-Driven Biological Ontologies



Journal

Cell Systems, 8(3)

ISSN

2405-4712

Authors

Yu, Michael Ku

Ma, Jianzhu

Ono, Keiichiro

et al.

Publication Date

2019-03-01

DOI

10.1016/j.cels.2019.02.003

Peer reviewed

HHS Public Access

Author manuscript

*Cell Syst*. Author manuscript; available in PMC 2020 February 26.

Published in final edited form as:

*Cell Syst*. 2019 March 27; 8(3): 267–273.e3. doi:10.1016/j.cels.2019.02.003.

# DDOT: A Swiss Army Knife for Investigating Data-Driven Biological Ontologies

**Michael Ku Yu**[1,2,4,5], **Jianzhu Ma**[1,5], **Keiichiro Ono**[1,5], **Fan Zheng**[1,5], **Samson H. Fong**[1,3], **Aaron Gary**[1], **Jing Chen**[1], **Barry Demchak**[1], **Dexter Pratt**[1], **Trey Ideker**[1,2,3,6,*]

[1]Department of Medicine, University of California, San Diego, La Jolla, CA 92093, USA

[2]Graduate Program in Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA 92093, USA

[3]Department of Bioengineering, University of California, San Diego, La Jolla, CA 92093, USA

[4]Toyota Technological Institute at Chicago, Chicago, IL 60637, USA

[5]These authors contributed equally

[6]Lead Contact

## SUMMARY

Systems biology requires not only genome-scale data but also methods to integrate these data into interpretable models. Previously, we developed approaches that organize omics data into a structured hierarchy of cellular components and pathways, called a "data-driven ontology." Such hierarchies recapitulate known cellular subsystems and discover new ones. To broadly facilitate this type of modeling, we report the development of a software library called the Data-Driven Ontology Toolkit (DDOT), consisting of a Python package (https://github.com/idekerlab/ddot) to assemble and analyze ontologies and a web application (http://hiview.ucsd.edu) to visualize them. Using DDOT, we programmatically assemble a compendium of ontologies for 652 diseases by integrating gene-disease mappings with a gene similarity network derived from omics data. For example, the ontology for Fanconi anemia describes known and novel disease mechanisms in its hierarchy of 194 genes and 74 subsystems. DDOT provides an easy interface to share ontologies online at the Network Data Exchange.

## In Brief

*Correspondence: tideker@ucsd.edu.

AUTHOR CONTRIBUTIONS

Conceptualization, M.K.Y., J.M., and T.I.; Writing, M.K.Y., J.M., and T.I.; Python Package Design and Implementation, M.K.Y., F.Z., J.M., and S.F.; Gene Similarity Network, F.Z.; HiView Design, K.O., M.K.Y., F.Z., J.M., B.D., T.I.; HiView Implementation, K.O.; NDEx Integration, A.G., J.C., and D.P.

WEB RESOURCES

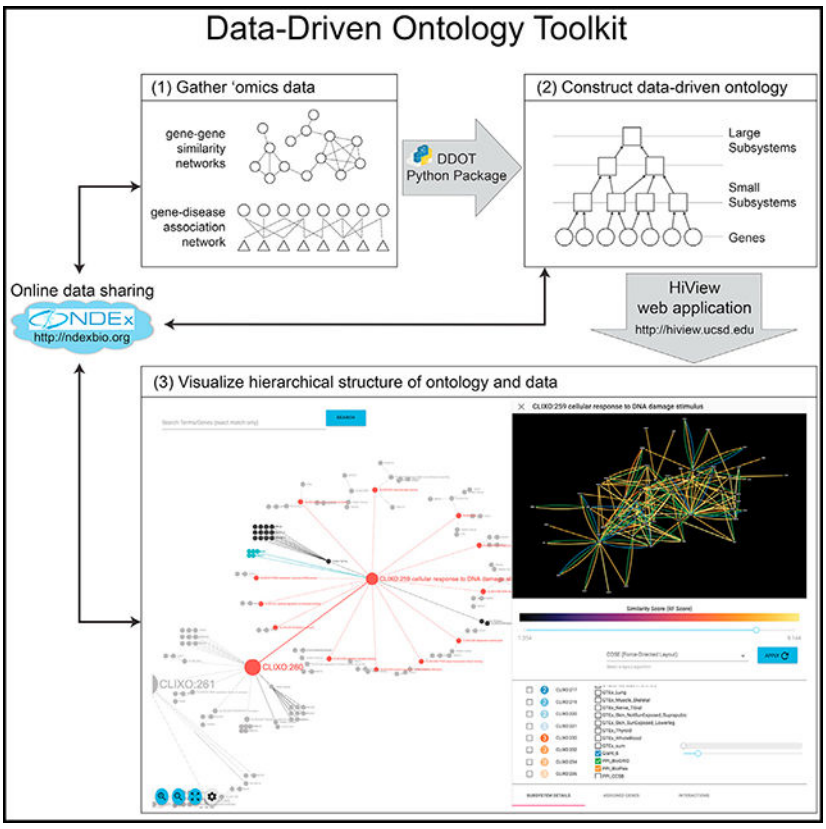
Fanconi Anemia Mutation Database, http://www2.rockefeller.edu/fanconi/
GitHub, https://github.com/idekerlab/ddot
HiView, http://hiview.ucsd.edu
NDEx, http://ndexbio.org

SUPPLEMENTAL INFORMATION

Supplemental Information can be found with this article online at https://doi.org/10.1016/j.cels.2019.02.003.

Recent computational techniques have used omics datasets to infer the multiscale structure of the cell and its biological systems. To facilitate construction and visualization of these models, Yu et al. present a software library called the Data-Driven Ontology Toolkit (DDOT) and demonstrate its utility in studying many different diseases.

## Graphical Abstract




## INTRODUCTION

Biological systems are organized hierarchically across multiple scales, from genes and proteins to protein complexes and pathways to cells, tissues, and individuals. The proliferation of omics datasets creates the potential to reveal this organizational complexity in an unprecedented and unbiased manner, whether through generation of proteomic data (protein-protein interactions and co-localization) (Chong et al., 2015; Huttlin et al., 2017), transcriptomic data (RNA co-expression across conditions and time points) (Saha et al., 2017; Sefer et al., 2016), or genetic data (epistasis and synthetic lethality) (Costanzo et al., 2016).

In the interest of building models that are both hierarchical and data driven, we previously introduced an approach for organizing genes into a hierarchy of cellular subsystems based on their gene-gene pairwise similarities in omics data (Dutkowski et al., 2013; Kramer et al., 2014). Unlike other hierarchical clustering algorithms that produce binary trees (a.k.a.

dendrograms), the core idea is to produce hierarchies with the flexibility to capture the structure of a cell, recognizing, for instance, that a subsystem may factor into many subcomponents (not just two) and participate in several higher-order processes (pleiotropy). In this way, the resulting model, called a data-driven hierarchy or data-driven ontology, has the potential to complement the knowledge in literature-curated ontologies in an unbiased and scalable manner. Through a procedure known as ontology alignment, we showed that these hierarchies not only recapitulate subsystems in the Gene Ontology (GO), including 60% of known cellular components in *S. cerevisiae,* but also discover new subsystems (Dutkowski et al., 2013).

Motivated by the initial success of this approach (Carvunisand Ideker, 2014; Dolinski and Botstein, 2013), additional methods have since been developed for inferring data-driven hierarchies (Gligorijević et al., 2014; Li and Yip, 2016; Peng et al., 2016). The data-driven nature of these methods also enables the *de novo* modeling of diseases and biological processes (Ames, 2017; Kim et al., 2016; Kramer et al., 2017). For instance, previously, we used data from experimental models of autophagy to infer a hierarchy of autophagy-related processes (Kramer et al., 2017). This hierarchy suggested many mechanistic hypotheses, a number of which we experimentally confirmed, including a revised understanding of known processes such as selective autophagy, the discovery of new processes such as the transport of Atg19-receptor cargos, and the discovery of new functions of genes such as *Gyp1* and *Atg26.* Similarly, the co-expression of genes in the fungal pathogen *Magnaporthe oryzae* was used to infer a hierarchy of cellular subsystems that are invoked during infection (Ames, 2017). Beyond genes and cellular subsystems, this form of modeling has also been used to organize other biomedical concepts. For instance, Park et al. hierarchically organized diseases based on their shared molecular mechanisms and showed correspondences with MeSH and the Disease Ontology (Park et al., 2017). We have also organized text phrases into a hierarchy of higher-order semantic concepts based on their co-usage in the abstracts of biomedical papers (Wang et al., 2018).

To broadly facilitate these approaches in the biomedical research community, we now report the development of a software framework, the Data-Driven Ontology Toolkit (DDOT), to enable the construction and analysis of hierarchical models in a Python package and their visualization in a web application. In contrast to existing tools for studying hierarchical models and literature-curated ontologies, DDOT supports general hierarchies rather than trees and focuses on the analysis of data-driven structure instead of semantic relations.

## RESULTS

### Introduction with Application to Fanconi Anemia

DDOT implements four major functions.

- Build data-driven ontology: Given a set of genes and a gene similarity network, hierarchically cluster the genes to infer cellular subsystems using the CliXO algorithm (Kramer et al., 2014). The resulting hierarchy of subsystems defines a data-driven ontology.

- Visualize hierarchical structure: Browse the full hierarchical structure of a data-driven ontology, including the network of gene similarities used to infer it, in a web application called the Hierarchical Viewer (HiView, http://hiview.ucsd.edu).
- Align ontologies: Annotate a data-driven ontology by aligning it to a curated ontology such as the GO. For instance, if a data-driven subsystem contains a similar set of genes as the GO term for DNA repair, then annotate this subsystem as being involved in DNA repair. Data-driven subsystems with no such matches represent new molecular mechanisms.
- Expand gene set: Given a set of genes as a "seed set" and a gene similarity network, identify an expanded set of genes that are highly similar to the seed set. This function can broaden the scope of a data-driven ontology beyond genes that are already well known.

We illustrate the above functions in an example study of Fanconi anemia (FA), a rare genetic disorder that is associated with bone marrow failure, myeloid dysplasia, and increased cancer risk (Ceccaldi et al., 2016). A total of 20 genes have been classified as FA genes because their germline mutations in patients have been associated with FA clinical phenotypes (Fanconi Anemia Mutation Database, http://www2.rockefeller.edu/fanconi/). All of these genes have known functions in the repair of DNA damage due to interstrand cross-links. However, beyond these DNA repair functions, the full spectrum of genes and pathways underlying FA remains unclear. For example, recently, 7 of the 20 FA genes were linked to new functions in autophagy, separate from their classical roles in DNA repair (Sumpter et al., 2016). Moreover, 127 other genes have been co-cited with FA in at least one study (STAR Methods).

Based on our previous procedure for studying autophagy (Kramer et al., 2017), we applied DDOT in a five-step pipeline to construct a FA gene ontology (FanGO) as follows (Figure 1A). First, we gathered input data, consisting of the 20 known FA genes as a seed set of genes for modeling and a gene similarity network derived by integrating several types of molecular evidence including protein-protein interactions, co-expression, co-localization, and epistasis (STAR Methods; Figure S1A). Second, we scored every gene for its involvement in FA by calculating its average functional similarity to the seed genes. The minimum score among the seed genes was used as a threshold to identify an additional set of 174 candidate genes (Figure 2A). Third, we organized all genes in a hierarchy of 74 cellular subsystems to construct FanGO. Fourth, we aligned it with GO. Finally, we uploaded FanGO to an online database, the Network Data Exchange (NDEx, http://ndexbio.org) (Pratt et al., 2015), and visualized the results in HiView (Figure 1B).

Since the time of constructing FanGO, one of the candidate genes, *RFWD3* (a.k.a. *FANCW),* was independently confirmed as a FA gene (Knies et al., 2017). Among the other candidate genes, 54 have been co-cited with FA (Figure 2B). Using the co-cited genes as a benchmark, the recall of our set of candidate genes is 54/127 = 0.43, and the precision is 54/174 = 0.31. Moreover, the co-cited genes tended to have stronger functional similarities to the seed set than other candidate genes (Figures S1B–S1D). An ontology alignment between FanGO and GO revealed that 43 of FanGO subsystems (58%) had significant

overlap with GO terms (Figure 2C). Consistent with prior knowledge that FA is marked by sensitivity to DNA damage, many of these overlapping GO terms are cellular complexes or pathways involved in the recognition of DNA lesions, including the GINS and MutSalpha complexes, or the repair of the DNA helix. Canonical FA subsystems, such as the "FA nuclear complex" and the "FANCM-MHF" complex, were also found (Figure 2D).

The recovery of these known connections suggests that the other 120 genes and 31 subsystems in FanGO are attractive hypotheses for further study in laboratory models or patients with FA phenotypes. In particular the genes *RFC4* and *RMI,* although not currently known to be involved in FA, have higher average similarity scores to the seed set than observed among the seed genes themselves. Several FanGO subsystems involve cellular functions that are not immediately recognizable as related to DNA damage repair, such as mRNA splicing (Figure 2E), the condensin complex, and telomere maintenance.

## Assembly of Data-Driven GOs for 652 Diseases

To demonstrate the accessibility and ease of computational modeling enabled by DDOT, we repeated the modeling procedure used for FA to programmatically construct data-driven ontologies for numerous other diseases, totaling 652 in all. These ontologies were based on two types of input data: a set of known gene associations for each disease, curated in the Monarch Initiative database (Mungall et al., 2017), and the same gene similarity network used to construct FanGO. By calling DDOT functions, the pipeline for constructing these ontologies was very concise, consisting of 16 lines of code for loading input data and setting parameters and 8 lines for modeling in a single Python script. The ontologies are available on NDEx and can be visualized through HiView (Table S1).

## A Suite of Functions Organized in a Python Package

Beyond the major functions described above, DDOT provides many other utility functions to analyze an ontology using the Python package (STAR Methods). At the core of the package is an "Ontology" class through which most analyses can be executed. This object-oriented design enables more intuitive software development, as conceptual manipulations to an ontology's structure can be reflected by programmatic changes to an Ontology object's attributes. DDOT's functions have all been designed to work together in concise pipelines that involve minimal boilerplate code. To facilitate shareable and reproducible software pipelines, DDOT has been designed such that both the input data and output ontologies can be stored and retrieved online at NDEx (Figure 1A) (Pratt et al., 2015). This built-in connection enables the use of a common data portal and sharing of results through URLs referencing data on NDEx.

## Ontology Visualization with HiView

The HiView web application provides an interactive visualization of the two major features of a data-driven ontology: (1) the hierarchical structure relating genes and subsystems and (2) the data supporting the inference of each subsystem (Figure 1B). Visualizing the hierarchical structure is challenging because it is a directed acyclic graph (DAG), in which each node may have multiple parents and multiple children. Drawing a DAG on a two-dimensional canvas often requires that many edges cross, inducing an inscrutable "hairball"

effect. Current tools for visualizing hierarchies are typically limited to simplified planar structures, such as a tree, where each node has at most one parent, or a small subgraph of the DAG based on local context. For instance, the QuickGO browser (Binns et al., 2009) for the GO shows the subgraph containing the ancestors of a selected term but excludes other close relations, such as sibling terms. In HiView, a user can graphically transform a DAG into a tree using three different methods, each providing a tradeoff between the size of the tree and the information captured. In one transformation (Figures S2A and S2B), edges in the DAG are pruned to leave behind a spanning tree–a smaller structure that loses some hierarchical information. In the other two transformations, a larger tree preserving all information is created by duplicating nodes, either to represent a top-down traversal of the DAG (Figure S2C) or to recover information lost in the spanning tree (Figure S2D). Given any of these transformations, a user can choose to render the tree as either a “node-link” diagram (Figure S2E), where hierarchical relations are represented by directed edges and drawn compactly with a layout algorithm, or a “circle-packing” diagram (Figure S2F), where relations are intuitively represented by drawing small circles nested within larger ones (like physical compartments of the cell). In addition, HiView allows the user to interactively zoom between more expansive views of the entire hierarchy and more focused views of particular subsystems. Finally, genes and subsystems can be searched based on their names and metadata.

Whereas a subsystem in a manually curated ontology is explained by a table of literature citations and evidence codes, a subsystem in a data-driven ontology is explained by a densely connected community of nodes in a biological network, which themselves require special graphical visualization. In HiView, these networks are displayed in a side panel when the user selects a subsystem (Figure 1B). Distinct interaction types, such as protein-protein versus co-expression, are distinguished by edge color, and the interaction strength is represented by edge thickness. To filter the amount of data shown, a user can select edges by their interaction type and strength. Furthermore, to understand why a subsystem was factorized into children subsystems, the user can also highlight the genes belonging to a particular child, making it easier to inspect the density of interactions within that child versus the density across all other interactions. HiView has been designed to be programmatic and capable of visualizing any ontology that has been pre-formatted with the Python package and hosted online at NDEx. Additional metadata about genes or subsystems can be viewed as node attributes, such as color and size.

## DISCUSSION

Bioinformatics analysis often involves complex maneuvers among heterogeneous formatting, model construction, and interpretation. To facilitate these steps in creating hierarchical models of biological systems, DDOT has been engineered with several key design choices worthy of mention. First, we have implemented an Ontology class as a central data structure through which major functions are executed. Both low-level and high-level functions have been implemented to enable flexible and concise software pipelines. Second, we support several types of formats for importing and exporting Ontology objects, including text files (tabular, CX, and OBO formats), in-memory Python objects (Pandas data-frames, iGraph objects, and NetworkX objects), and online files stored on NDEx.

Third, model construction by the Python package has been seamlessly tied to model interpretation by HiView, enabling faster prototyping and iteration of ideas. Finally, we have provided in-depth documentation of every function and a tutorial of the Python package to minimize the learning curve for using it and to encourage software extensions by others.

Hierarchical models of the cell and other biological systems have long been curated in the form of biomedical ontologies, but their construction and visualization in a data-driven manner is a more recent endeavor for which no unified software framework yet exists. DDOT enables rapid exploration of hierarchies of cellular subsystems for numerous diseases and biological contexts. We have taken a first step in this exploration by creating hierarchies for 652 diseases.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

### CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Trey Ideker (tideker@ucsd.edu).

### METHOD DETAILS

#### Other Functions Implemented by DDOT

- Examine ontology structure. For each subsystem, retrieve its hierarchical connections (genes, child and descendant subsystems, parent and ancestral subsystems) and the subnetwork of gene similarities that supports the subsystem's existence. For each gene, retrieve its set of subsystems.
- Modify ontology structure. Reduce the size of an ontology by removing a set of subsystems or genes. Randomize connections between genes and subsystems to create new ontologies representing a null model for statistical tests.
- Flatten ontology structure. Instead of inferring an ontology from a gene similarity network, perform the reverse process of inferring a gene similarity network from an ontology. In particular, the similarity between two genes is calculated as the size of the smallest common subsystem, known as the Resnik semantic similarity score (Resnik, 1999).
- Map genotypes to the ontology. Given a set of mutations comprising a genotype, propagate the impact of these mutations to the subsystems containing these genes in the ontology. In particular, the impact on a subsystem is estimated by the number of its genes that have been mutated. These subsystem activities, which we have called an "ontotype", enables more accurate and interpretable predictions of phenotype from genotype (Yu et al., 2016).
- Load curated ontologies. Parse Open Biomedical Ontologies (OBO) and gene-association file (GAF) formats that are typically used to describe curated ontologies like GO.

**The Hierarchical Viewer—**In designing HiView, the rendering library was chosen based on a tradeoff between rendering speed and visual styling capabilities. The main panel for viewing a hierarchy's structure is rendered with sigma.js with WebGL enabled. The side panel for viewing the supporting interaction networks, on the other hand, was implemented using cytoscape.js (Franz et al., 2016), which renders more slowly than sigma.js but offers more styling capabilities out-of-the-box, such as node shapes and color gradients. The spanning tree used for the transformations in Figures S2B and S2D was calculated by connecting each subsystem to its smallest parent subsystem; all other connections were hidden. For the node-link diagram, the layout of nodes in a tree is calculated using the Bubble Tree Layout algorithm (Grivet et al., 2006) implemented in the Tulip Python library (Auber et al., 2017). The circle-packing diagram is implemented using a customized version of the D3.js library (https://d3js.org/).

## QUANTIFICATION AND STATISTICAL ANALYSIS

**Construction of Disease Gene Ontologies—**The 20 seed FA genes are *FANCA, FANCB, FANCC, BRCA2, FANCD2, FANCE, FANCF, FANCG, FANCI, BRIP1, FANCL, FANCM, PALB2, RAD51C, SLX4, ERCC4, RAD51*, *BRCA1*, *UBE2T,* and *XRCC2.* The set of genes that have been co-cited with Fanconi Anemia was gathered using a text mining procedure described in (Wang et al., 2018). Briefly, we gathered all Pubmed abstracts published up to date on April 30th, 2017 and identified all words that appear in an abstract with the phrase "Fanconi Anemia". The words that are either an official or alias symbol of a human gene, according to the HGNC consortium, were counted (disregarding letter casing).

Gene-disease associations for 8,590 diseases from the Monarch Database was downloaded on May 13, 2017 with the help of Kent Shefchek and Chris Mungall (personal communication). For most diseases, the known and candidate genes together totaled more than 500 genes, suggesting the lack of a coherent molecular signature. Ontologies were constructed and studied for the 651 diseases that had no more than 500 known and candidate genes.

The gene similarity network was based on more than 1000 gene-gene interaction networks from publicly available sources. Briefly, we compiled 16 tissue-specific coexpression networks from the GTEx project (GTEx Consortium, 2013; Saha et al., 2017), 10 PCA components of the 980 GEO co-expression networks (Edgar et al., 2002) curated by the GIANT study (Greene et al., 2015), 1 coexpression network from the Cancer Cell Line Encyclopedia (Barretina et al., 2012), 2 protein domain similarity networks (InterPro (Hunter et al., 2009), PFAM (Bateman et al., 2004)), 1 genetic interaction network from (Lin et al., 2010), 2 curated protein-protein interaction networks from databases (BioGRID (Stark et al., 2006), InBioMap (Li et al., 2017)), and 4 high-throughput protein-protein interaction networks ((Huttlin et al., 2017), (Havugimana et al., 2012), (Rolland et al., 2014), (Hein et al., 2015)), and 1 computationally predicted human interactome (Zhang et al., 2013). Following the procedure in (Kramer et al., 2017), we integrated these networks by using them as features in a supervised learning (random forests) of the Resnik semantic similarity (Resnik, 1999) between genes in the Gene Ontology. This integration was done in 5-fold cross validation, i.e., for each fold we only used the predictions made on the test set of gene

pairs. To assess the performance of this integration procedure, we found that our network has a substantial correlation (Pearson $\rho = 0.39$) with the Resnik similarity (Figure S1A). Agreements between our network and the Resnik similarity reflect consistency with known biological connections in the Gene Ontology. On the other hand, disagreements reflect biological connections in our network that is supported by data but not captured in the Gene Ontology.

To discover candidate genes for every disease, we applied the same strategy as for Fanconi Anemia. In particular, we scored every gene by their average similarity to the seed set, and we used the minimum score among seed genes as a threshold to identify candidate genes. Although many methods exist for discovering candidate genes using biological networks (reviewed in (Leiserson et al., 2013) and (Chimusa et al., 2018)), we chose this method in order to demonstrate a simple and clear analytical pipeline. Users of the Python package can input their own set of genes to analyze and can also substitute their own method for identifying candidate genes.

To infer gene ontologies from the similarity network, we invoked the CliXO algorithm (Kramer et al., 2014) using the DDOT function Ontology.infer_ontology(…) with parameters alpha = 0.05 and beta = 0.5. Briefly, CliXO searches for cliques (sets of genes in which all gene pairs have high similarity) or dense subnetworks (sets of genes in which many gene pairs do) above a specified similarity (scale) threshold. By progressively loosening this threshold, CliXO identifies progressively larger cliques, which subsume the smaller cliques found at earlier thresholds. Each clique defines a cellular subsystem, and all cliques are arranged to form a “data-driven hierarchy” (or ontology) of subsystems.

**Alignment of Data-Driven Ontologies to GO—**Each data-driven ontology was aligned to the Gene Ontology (GO), downloaded on October 3, 2017. Because GO is curated as a general structure to represent all species, we created a human-focused GO by removing terms that do not contain any human genes or contain the same genes as its parents terms (DDOT function *Ontology.precollapse*(…)).

The ontology alignment was performed using the algorithm described in (Dutkowski et al., 2013) with FDR cutoff of 0.05, calculated from 100 randomized iterations (DDOT function *Ontology.align*(…)). Briefly, the alignment attempts to find an optimal matching of subsystems in one ontology to subsystems in the other ontology. The similarity of two subsystems is defined by two notions: an “intrinsic similarity” of the set of genes in the subsystems, and a “relational similarity” of the parents and children of one subsystem with those of the other. The alignment is constrained so that each subsystem is matched to at most one other subsystem. To respect the hierarchical structure of the ontologies, the alignment also avoids “parent-child criss crosses”: if subsystem A is matched with subsystem B, then no ancestor of A can be matched with a descendant B, or vice versa. To calculate an FDR of each match, we simulated a null model by re-running the alignment algorithm on randomizations of the ontologies.

## DATA AND SOFTWARE AVAILABILITY

Source code and installation of the Python package is available at https://github.com/idekerlab/ddot under a MIT open source license. Further documentation of each Python function is at https://ddot.readthedocs.io. Jupyter notebooks for loading datasets and constructing data-driven ontologies of the 652 diseases are at https://github.com/idekerlab/ddot/tree/master/examples. NDEx and HiView URLs of these ontologies are at https://github.com/idekerlab/ddot/bob/master/examples/disease_gene_ontologies.txt (also Table S1). The HiView web application can be accessed at http://hiview.ucsd.edu/, and its source code is available at https://github.com/idekerlab/hiview under a MIT open source license.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health to T.I. (TR002026, GM103504, and CA184427) and the California Institute for Regenerative Medicine Center of Excellence in Stem Cell Genomics. We wish to thank Kent Shefchek and Chris Mungall for helping us obtain data from the Monarch Initiative. We also wish to thank Anton Kratz and Jisoo Park for testing software. Finally, we wish to thank Tamara Munzner for excellent discussions on the visualization architecture of HiView.

DECLARATION OF INTERESTS

Trey Ideker is the co-founder of Data4Cure, Inc. and has an equity interest. Trey Ideker has an equity interest in Ideaya BioSciences, Inc. and is funded by a sponsored research agreement from Ideaya. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

## REFERENCES

Ames RM (2017). Using network extracted ontologies to identify novel genes with roles in appressorium development in the rice blast fungus Magnaporthe oryzae. Microorganisms 5, E3. [PubMed: 28106722]

Auber D, Archambault D, Bourqui R, Delest M, DuBois J, Lambert A, Mary P, Mathiault M, Melançon G, Pinaud B, et al. (2017). TULIP 5 In Encyclopedia of Social Network Analysis and Mining (Springer), pp. 1–28.

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D, et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607. [PubMed: 22460905]

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, et al. (2004). The Pfam protein families database. Nucleic Acids Res. 32, D138–D141. [PubMed: 14681378]

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. [PubMed: 19744993]

Carvunis AR, and Ideker T (2014). Siri of the cell: what biology could learn from the iPhone. Cell 157, 534–538. [PubMed: 24766803]

Ceccaldi R, Sarangi P, and D'Andrea AD (2016). The Fanconi anaemia pathway: new players and new functions. Nat. Rev. Mol. Cell Biol 17, 337–349. [PubMed: 27145721]

Chimusa ER, Dalvie S, Dandara C, Wonkam A, and Mazandu GK (2018). Post genome-wide association analysis: dissecting computational pathway/network-based approaches. Brief. Bioinform bby035.

Chong YT, Koh JL, Friesen H, Duffy SK, Cox MJ, Moses A, Moffat J, Boone C, and Andrews BJ (2015). Yeast proteome dynamics from single cell imaging and automated analysis. Cell 161, 1413–1424. [PubMed: 26046442]

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, et al. (2016). A global genetic interaction network maps a wiring diagram of cellular function. Science 353, aaf1420. [PubMed: 27708008]

Dolinski K, and Botstein D (2013). Automating the construction of gene ontologies. Nat. Biotechnol 31, 34–35. [PubMed: 23302932]

Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, and Ideker T (2013). A gene ontology inferred from molecular networks. Nat. Biotechnol 31, 38–45. [PubMed: 23242164]

Edgar R, Domrachev M, and Lash AE (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210. [PubMed: 11752295]

Franz M, Lopes CT, Huck G, Dong Y, Sumer O, and Bader GD (2016). Cytoscape.js: a graph theory library for visualisation and analysis. Bioinformatics 32, 309–311. [PubMed: 26415722]

Gligorijević V, Janjić V, and Pržulj N (2014). Integration of molecular network data reconstructs Gene Ontology. Bioinformatics 30, i594–i600. [PubMed: 25161252]

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet 47, 569–576. [PubMed: 25915600]

Grivet S, Auber D, Domenger JP, and Melancon G (2006). Bubble tree drawing algorithm In Computer Vision and Graphics (Springer), pp. 633–641.

GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. Nat. Genet 45, 580–585. [PubMed: 23715323]

Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, Wang PI, Boutz DR, Fong V, Phanse S, et al. (2012). A census of human soluble protein complexes. Cell 150, 1068–1081. [PubMed: 22939629]

Hein MY, Hubner NC, Poser I, Cox J, Nagaraj N, Toyoda Y, Gak IA, Weisswange I, Mansfeld J, Buchholz F, et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell 163, 712–723. [PubMed: 26496610]

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. (2009). InterPro: the integrative protein signature database. Nucleic Acids Res. 37, D211–D215. [PubMed: 18940856]

Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, Colby G, Gebreab F, Gygi MP, Parzen H, et al. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature 545, 505–509. [PubMed: 28514442]

Kim M, Rai N, Zorraquino V, and Tagkopoulos I (2016). Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli. Nat. Commun 7, 13090. [PubMed: 27713404]

Knies K, Inano S, Ramírez MJ, Ishiai M, Surrallés J, Takata M, and Schindler D (2017). Biallelic mutations in the ubiquitin ligase RFWD3 cause Fanconi anemia. J. Clin. Invest 127, 3013–3027. [PubMed: 28691929]

Kramer M, Dutkowski J, Yu M, Bafna V, and Ideker T (2014). Inferring gene ontologies from pairwise similarity data. Bioinformatics 30, i34–i42. [PubMed: 24932003]

Kramer MH, Farré JC, Mitra K, Yu MK, Ono K, Demchak B, Licon K, Flagg M, Balakrishnan R, Cherry JM, et al. (2017). Active interaction mapping reveals the hierarchical organization of autophagy. Mol. Cell 65, 761–774.e5. [PubMed: 28132844]

Leiserson MDM, Eldridge JV, Ramachandran S, and Raphael BJ (2013). Network analysis of GWAS data. Curr. Opin. Genet. Dev 23, 602–610. [PubMed: 24287332]

Li L, and Yip KY (2016). Integrating information in biological ontologies and molecular networks to infer novel terms. Sci. Rep 6, 39237. [PubMed: 27976738]

Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, Workman CT, Rigina O, Rapacki K, Stærfeldt HH, et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. Nat. Methods 14, 61–64. [PubMed: 27892958]

Lin A, Wang RT, Ahn S, Park CC, and Smith DJ (2010). A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. Genome Res. 20, 1122–1132. [PubMed: 20508145]

Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, Carbon S, Conlin T, Dunn N, Engelstad M, et al. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res. 45, D712–D722. [PubMed: 27899636]

Park J, Hescott BJ, and Slonim DK (2017). Towards a more molecular taxonomy of disease. J. Biomed. Semantics 8, 25. [PubMed: 28750648]

Peng J, Wang T, Wang J, Wang Y, and Chen J (2016). Extending gene ontology with gene association networks. Bioinformatics 32, 1185–1194. [PubMed: 26644414]

Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, Ono K, Miello C, Hicks L, Szalma S, et al. (2015). NDEx, the Network Data Exchange. Cell Syst. 1, 302–305. [PubMed: 26594663]

Resnik P (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res 11, 95–130.

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, et al. (2014). A proteome-scale map of the human interactome network. Cell 159, 1212–1226. [PubMed: 25416956]

Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, GTEx Consortium, Engelhardt BE, and Battle A (2017). Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res. 27, 1843–1858. [PubMed: 29021288]

Sefer E, Kleyman M, and Bar-Joseph Z (2016). Tradeoffs between dense and replicate sampling strategies for high-throughput time series experiments. Cell Syst. 3, 35–42. [PubMed: 27453445]

Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, and Tyers M (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34, D535–D539. [PubMed: 16381927]

Sumpter R Jr., Sirasanagandla S, Fernández ÁF, Wei Y, Dong X, Franco L, Zou Z, Marchal C, Lee MY, Clapp DW, et al. (2016). Fanconi anemia proteins function in mitophagy and immunity. Cell 165, 867–881. [PubMed: 27133164]

Wang S, Ma J, Yu MK, Zheng F, Huang EW, Han J, Peng J, and Ideker T (2018). Annotating gene sets by mining large literature collections with protein networks. In Proceedings of the Pacific Symposium on Biocomputing 2018, Altman RB, Dunker AK, Hunter L, Ritchie MD, Murray TA, and Klein TE, eds. (World Scientific), pp. 602–613.

Yu MK, Kramer M, Dutkowski J, Srivas R, Licon K, Kreisberg J, Ng CT, Krogan N, Sharan R, and Ideker T (2016). Translation of genotype to phenotype by a hierarchy of cell subsystems. Cell Syst. 2, 77–88. [PubMed: 26949740]

Zhang QC, Petrey D, Garzón JI, Deng L, and Honig B (2013). PrePPI: a structure-informed database of protein-protein interactions. Nucleic Acids Res. 41, D828–D833. [PubMed: 23193263]

## Highlights

- Python package to model hierarchical biological structure from omics data
- Web application to visualize hierarchical structure and its support in omics data
- Compendium of pathway hierarchies for 652 diseases
- Online sharing of input and output data through the Network Data Exchange

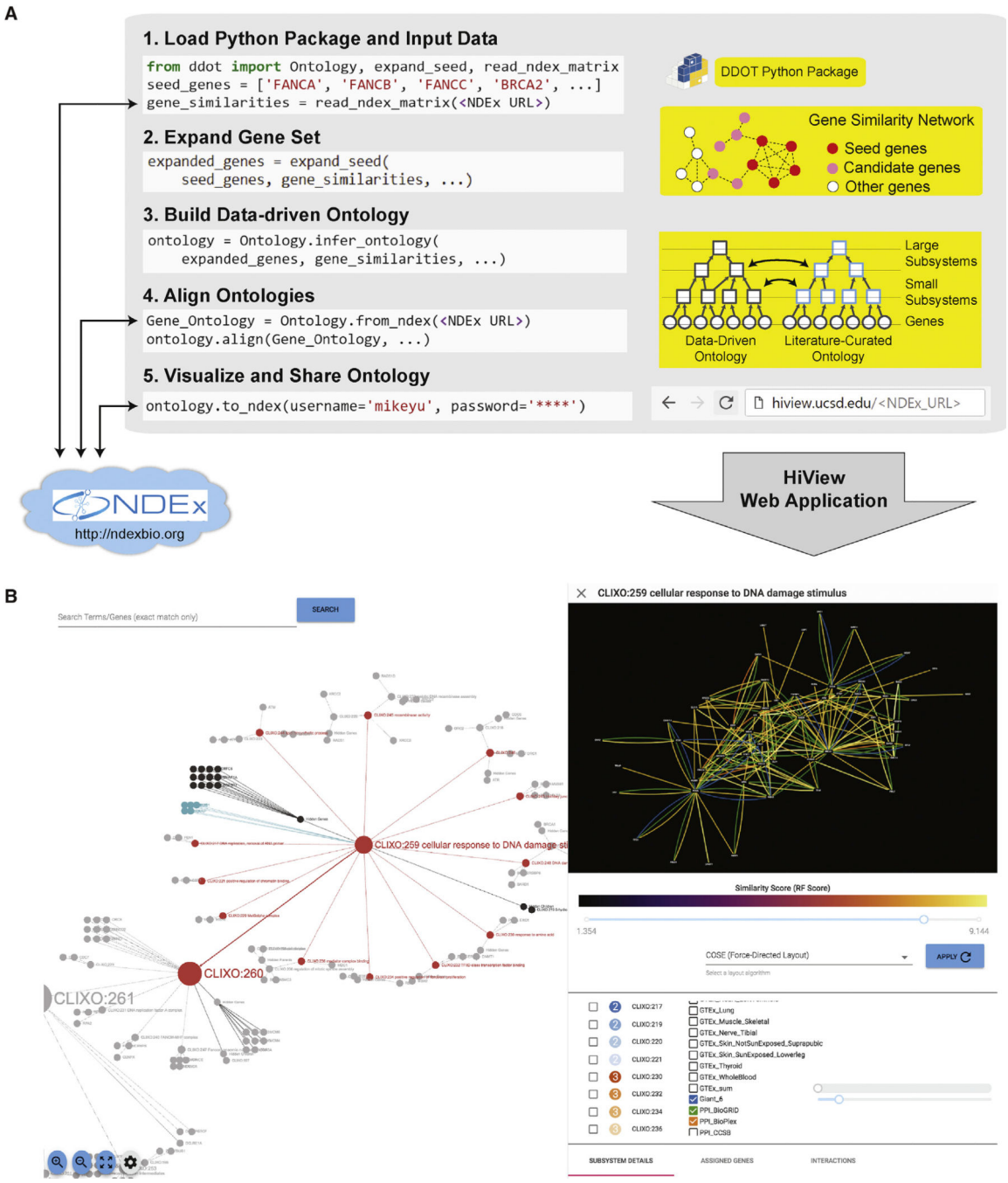



**Figure 1. Software Architecture of the Data-Driven Ontology Toolkit**

(A) An example workflow of using DDOT to construct, analyze, and visualize a data-driven gene ontology. This workflow is executed in an integrated software framework, consisting of a Python package and a web application called the Hierarchical Viewer (HiView, http://hiview.ucsd.edu). Input and output data can be stored online at the Network Data Exchange (NDEx, http://ndexbio.org), facilitating the sharing and reproducibility of results.

(B) HiView visualizes both the hierarchical structure of an ontology (left), as well as the omics data (right), in the form of gene interaction networks, which were used to infer a

subsystem and its hierarchical relations. Genes and subsystems can be searched by name or metadata (top-left).

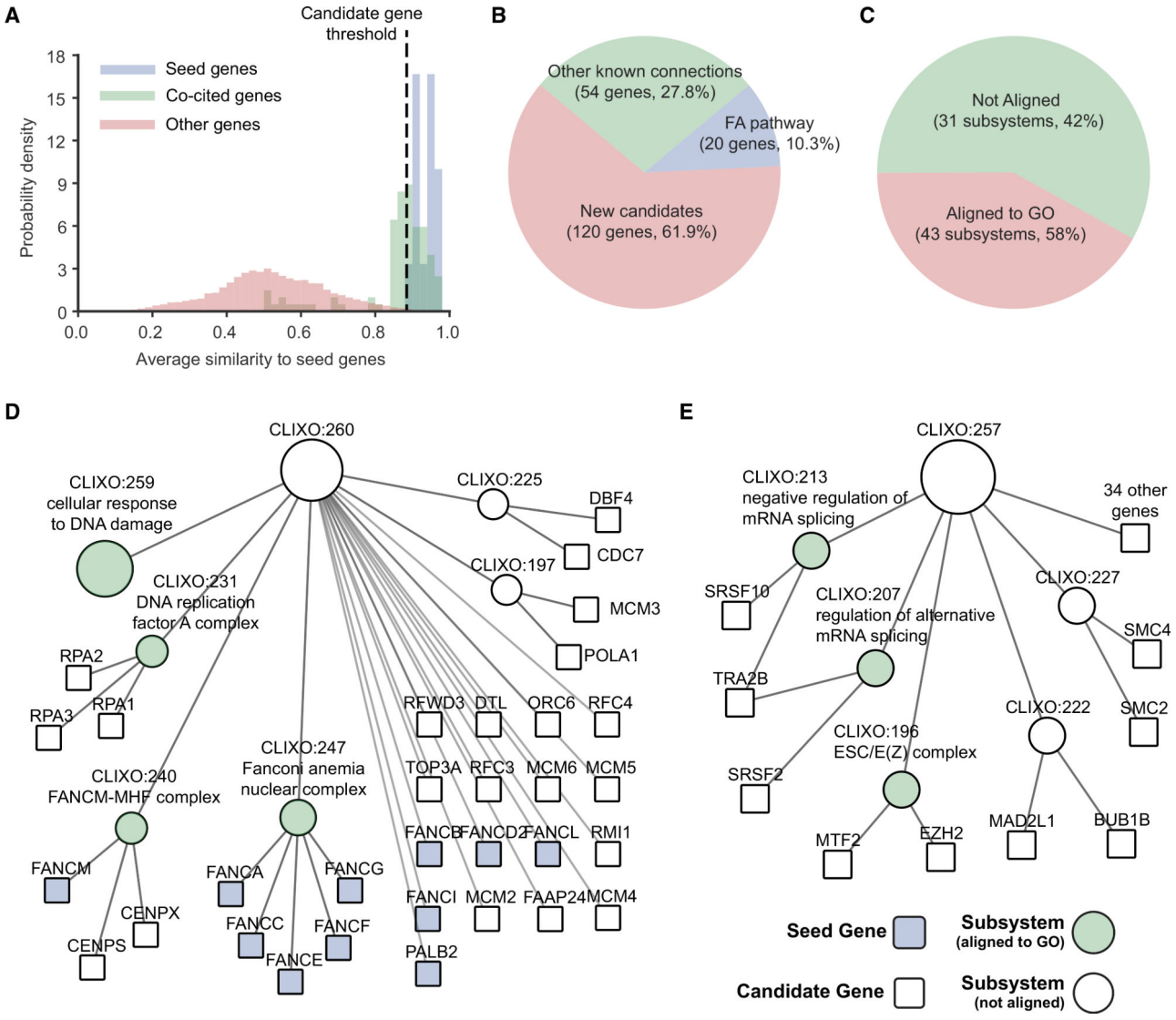



**Figure 2. Discovery of Genes and Cellular Subsystems Involved in 652 Diseases**

(A–C) A Fanconi anemia gene ontology (FanGO) assembled by combining prior knowledge of a seed set of 20 known FA genes with large-scale omics data. (A) Every gene was scored for its involvement in FA by calculating its average functional similarity to the seed set. A histogram of these scores is shown for genes in the seed set (blue), genes that are not in the seed set but have been co-cited with the phrase “Fanconi Anemia” (green), and all other genes (red). The minimum score among genes in the seed set (0.88) was used as a threshold to identify a candidate set of 174 genes. (B) Decomposition of the combined set of 194 genes.

(C) Decomposition of the 74 subsystems in FanGO based on an alignment to the Gene Ontology (GO).

(D and E) Focused view on FanGO subsystems related to the Fanconi anemia nuclear complex and FANCM-MHF complex (D) and mRNA splicing (E).

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Software and Algorithms | | |
| Python package source code | This paper | https://github.com/idekerlab/ddot |
| Python package documentation | This paper | https://ddot.readthedocs.io |
| Python package tutorial | This paper | https://github.com/idekerlab/ddot/blob/master/examples/Tutorial.ipynb |
| Jupyter notebook to make data-driven gene ontologies of 652 diseases | This paper | https://github.com/idekerlab/ddot/blob/master/examples/Make_disease_gene_ontologies.ipynb |
| Jupyter notebook to process the curated Gene Ontology (for ontology alignment) | This paper | https://github.com/idekerlab/ddot/blob/master/examples/Process_the_Gene_Ontology.ipynb |
| Jupyter notebook to load datasets (gene-disease associations, processed GO, gene-gene similarity network) | This paper | https://github.com/idekerlab/ddot/blob/master/examples/Load_example_datasets.ipynb |
| HiView web application | This paper | http://hiview.ucsd.edu |
| NDEx | (Pratt et al., 2015) | http://ndexbio.org |
| Visualize FanGO in HiView | This paper | http://hiview.ucsd.edu/0fb9fec3-f772-11e8-aaa6-0ac135e8bacf?type=public&serverhttp://public.ndexbio.org |
| Visualize other disease gene ontologies in HiView | This paper | https://github.com/idekerlab/ddot/blob/master/examples/disease_gene_ontologies.txt (also Table S1) |