Published online 9 December 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D555–D559***
doi: 10.1093/nar/gkv1329

GeneWeaver: data driven alignment of cross-species genomics in biology and disease

Erich Baker[1,2], Jason A. Bubier[3], Timothy Reynolds[2], Michael A. Langston[4] and Elissa J. Chesler[3,*]

[1]School of Engineering & Computer Science, Baylor University, Waco, TX 76798, USA, [2]Institute for Biomedical Studies, Baylor University, Waco, TX 76798, USA, [3]The Jackson Laboratory, Bar Harbor, ME 04609, USA and [4]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA

Received September 15, 2015; Revised November 11, 2015; Accepted November 13, 2015

ABSTRACT

The GeneWeaver data and analytics website (www.geneweaver.org) is a publically available resource for storing, curating and analyzing sets of genes from heterogeneous data sources. The system enables discovery of relationships among genes, variants, traits, drugs, environments, anatomical structures and diseases implicitly found through gene set intersections. Since the previous review in the 2012 Nucleic Acids Research Database issue, GeneWeaver's underlying analytics platform has been enhanced, its number and variety of publically available gene set data sources has been increased, and its advanced search mechanisms have been expanded. In addition, its interface has been redesigned to take advantage of flexible web services, programmatic data access, and a refined data model for handling gene network data in addition to its original emphasis on gene set data. By enumerating the common and distinct biological molecules associated with all subsets of curated or user submitted groups of gene sets and gene networks, GeneWeaver empowers users with the ability to construct data driven descriptions of shared and unique biological processes, diseases and traits within and across species.

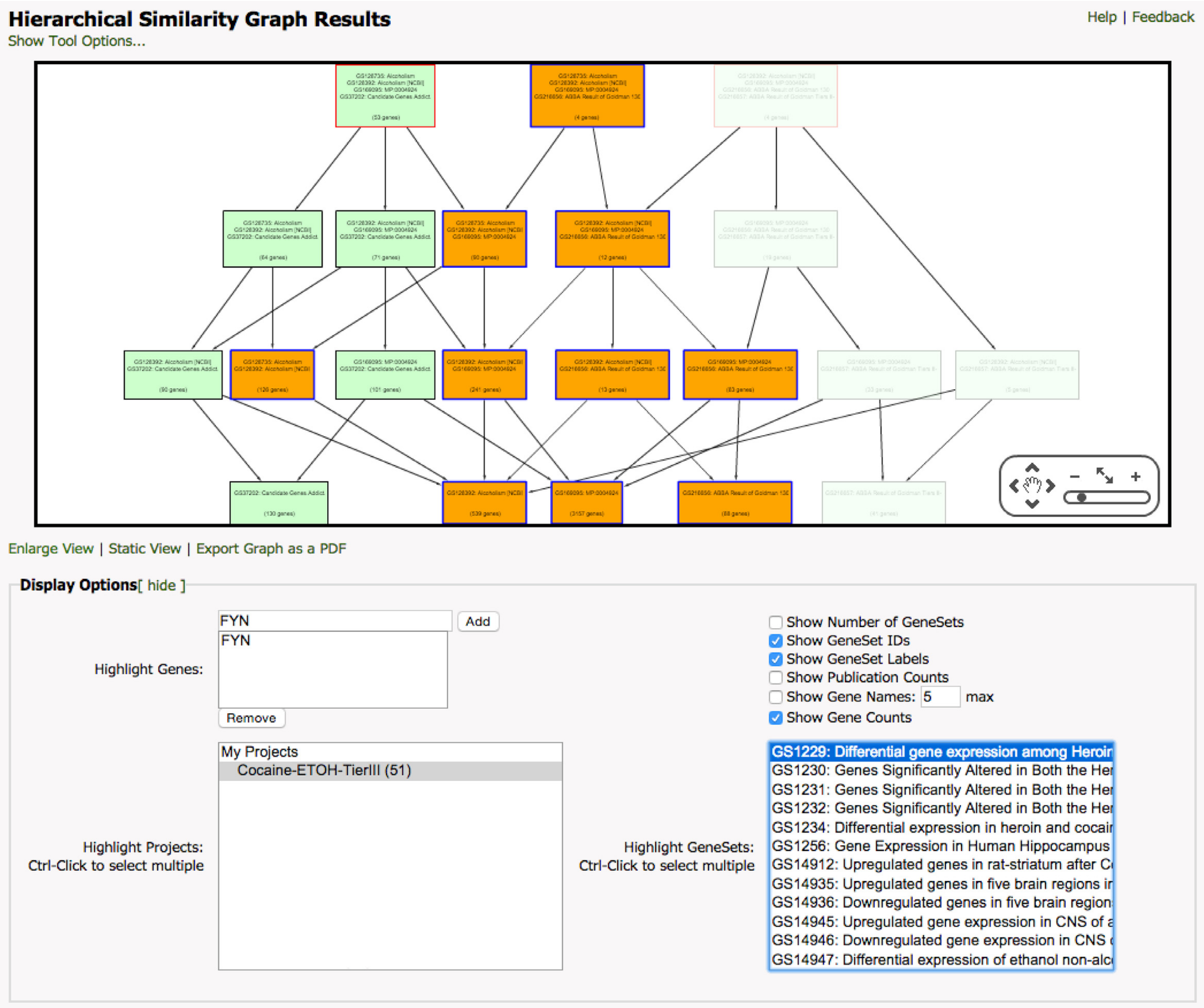
INTRODUCTION

There are many circumstances that benefit from the rapid and detailed one-to-many or many-to-many comparison of sets of genes and variants. These types of analytics arise in personal genomics, experimental functional genomics, genetic mapping and other analyses in which collections of diverse associations of genes and genomes to biological concepts, patients, diseases or samples must be compared and interpreted. GeneWeaver.org is a data repository and analytics platform that meets these needs through the storage, curation and analysis of publicly sourced and user-defined sets of genes across species (1,2).

Initially referred to as the Ontological Discovery Environment, this system enables users to apply biclique centric analyses to infer the relations among any biological concept that can be represented by a set of associated genes. A computationally tractable bipartite analysis tool (3) makes it possible for GeneWeaver users to analyze collections of gene-centric data to describe the emergent gene-to-disease, -ontology or -phenotype relationships hidden among biological concepts and molecular components by making use of labeled gene sets that retain the contextual information about the conditions under which co-occurring genes are present. This approach is akin to more recent work in data driven ontology using clique enumeration and intersection to identify relations among co-occurring genes to find the underlying biological components observed across systems (4). By taking advantage of a bipartite data structure, our suite of tools dynamically enumerates maximal bicliques which can be arranged into hierarchical associations for the identification of common and unique components shared between biological processes (HiSim Graph) or highly connected genes (GeneSet Graph) (1,2). In addition to graph-based approaches, GeneWeaver allows statistical interrogation of data, including quantitative assessment of gene set overlap through Jaccard similarity analysis and clustering. A suite of Boolean tools permits users to perform set combination or enumeration, allowing the creation of novel gene sets based on shared components, and thus refining their derived ontological structure over time.

Importantly, GeneWeaver's ability to integrate sets of genes and gene identifiers across species through identifier mapping enables heterogeneous data integration. In addition, genes may be associated with biological processes, disease states or semantic descriptions through Gene Ontology (GO) (5), Human Phenotype (HP) Ontology (6), Mammalian Phenotype (MP) Ontology (7) or the Ontology for Biomedical Investigations (OBI) (8). Aggregating sparse sets of genes by including multiple model organisms across

[*]To whom correspondence should be addressed. Tel: +1 207 288 6000; Fax: +1 207 288 6847; Email: Elissa.chesler@jax.org


Help | Feedback
Show Tool Options...
GS126735: Alcoholism
GS128392: Alcoholism [NCBI]
GS169095: MP:0004924
GS37202: Candidate Genes Addict
(53 genes)
GS126735: Alcoholism
GS128392: Alcoholism [NCBI]
GS169095: MP:0004924
GS216856: ABBA Result of Goldman 13E
(4 genes)
GS128392: Alcoholism [NCBI]
GS169095: MP:0004924
GS216856: ABBA Result of Goldman 13E
GS216857: ABBA Result of Goldman Tiers II-
(4 genes)
GS126735: Alcoholism
GS128392: Alcoholism [NCBI]
GS37202: Candidate Genes Addict
(64 genes)
GS128392: Alcoholism [NCBI]
GS169095: MP:0004924
GS37202: Candidate Genes Addict
(71 genes)
GS126735: Alcoholism
GS128392: Alcoholism [NCBI]
GS169095: MP:0004924
(90 genes)
GS128392: Alcoholism [NCBI]
GS169095: MP:0004924
GS216856: ABBA Result of Goldman 13E
(12 genes)
GS169095: MP:0004924
GS216856: ABBA Result of Goldman 13E
GS216857: ABBA Result of Goldman Tiers II-
(19 genes)
GS128392: Alcoholism [NCBI]
GS37202: Candidate Genes Addict
(90 genes)
GS126735: Alcoholism
GS128392: Alcoholism [NCBI]
(126 genes)
GS169095: MP:0004924
GS37202: Candidate Genes Addict
(101 genes)
GS128392: Alcoholism [NCBI]
GS169095: MP:0004924
(241 genes)
GS128392: Alcoholism [NCBI]
GS216856: ABBA Result of Goldman 13E
(13 genes)
GS169095: MP:0004924
GS216856: ABBA Result of Goldman 13E
(83 genes)
GS169095: MP:0004924
GS216857: ABBA Result of Goldman Tiers II-
(33 genes)
GS128392: Alcoholism [NCBI]
GS216857: ABBA Result of Goldman Tiers II-
(5 genes)
GS37202: Candidate Genes Addict
(130 genes)
GS128392: Alcoholism [NCBI]
(539 genes)
GS169095: MP:0004924
(3157 genes)
GS216856: ABBA Result of Goldman 13E
(88 genes)
GS216857: ABBA Result of Goldman Tiers II-
(41 genes)
Enlarge View | Static View | Export Graph as a PDF
Add
Remove

Hierarchical Similarity Graph Results

Display Options[hide]

Highlight Genes: FYN

FYN

☐ Show Number of GeneSets
☑ Show GeneSet IDs
☑ Show GeneSet Labels
☐ Show Publication Counts
☐ Show Gene Names: 5 max
☑ Show Gene Counts

Highlight Projects:
Ctrl-Click to select multiple

My Projects
Cocaine-ETOH-TierIII (51)

Highlight GeneSets:
Ctrl-Click to select multiple

GS1229: Differential gene expression among Heroin
GS1230: Genes Significantly Altered in Both the Her
GS1231: Genes Significantly Altered in Both the Her
GS1232: Genes Significantly Altered in Both the Her
GS1234: Differential expression in heroin and cocair
GS1256: Gene Expression in Human Hippocampus
GS14912: Upregulated genes in rat-striatum after C
GS14935: Upregulated genes in five brain regions ir
GS14936: Downregulated genes in five brain region:
GS14945: Upregulated gene expression in CNS of a
GS14946: Downregulated gene expression in CNS (
GS14947: Differential expression of ethanol non-alc

Figure 1. The GeneWeaver HiSim graph represents hierarchical intersections of maximal bicliques based on shared genes. This structure relates a data driven ontology. Here, intersecting sets of genes are colored by inclusion of a gene of interest (orange), or by sets of user-defined genes (green opacity).

multiple disease spaces unleashes the potential promised by convergent functional genomics. Effectively, this allows users to align disease phenotypes across model organisms, isolate sets of genes shared by biological processes or discover shared biological substrates within data driven disease hierarchies, as described in a recent review using GeneWeaver to find consilience among diverse data sets (9).

DATA CONTENT

Since its initial description in the Nucleic Acids Research Database issue in 2012, GeneWeaver has undergone more than a two fold growth of publically available and private gene sets (1). A primary objective of GeneWeaver is to allow users to load, store and curate *ad hoc* sets of genes, derived based on user defined metrics, such a microarray results, genetic mapping and association studies, and semantic or publication associations, among others. A historical problem with genomic approaches is that in isolation, a set of genes, variants or other molecular entities may contain many false positive or false negative associations to a disease state or biological concept. The same set may reveal highly relevant information when analyzed in the context of aggregate data consisting of thousands of gene sets derived from multiple species, tissue types, physiological states or genetic background. This approach effectively reduces noise through high degree convergence among multiple experimentally derived data types. To this end, we have curated a large collection of gene sets, each assigned to a tier based on source and status. Publically available sets of genes annotated to structured vocabularies and ontologies are assigned Tier I, or public resource data. In the case of semantic ontologies, each GeneWeaver gene set represents the closure of genes associated with a particular term. Other sets of genes, such as MeSH term-to-gene annotations, are derived from the processing of public sources and attributed to Tier II. In the case of MeSH, we take advantage of NCBI's gene-to-Pubmed and Pubmed-to-mesh files to produce sets of genes annotated through their transitive associations. Tiers

III (reviewed) and IV (user submitted and pending review) data are manually curated and publically available, while Tier V designates private user submitted data that has not been subject to curatorial review or released to the public. To date, GeneWeaver houses 100 069 active sets of genes, including 64 639 Tier I, 17 482 Tier II, 1070 Tier III and 14 386 Tier V gene sets. These numbers represent a 225% increase during the last three years and include the addition of gene sets associated with the Kyoto Encyclopedia of Genes and Genomes (KEGG) (10), MeSH, Molecular Signatures Database (MSigDB) (11), Mammalian Genome Institute (MGI), Online Mendelian Inheritance in Man (OMIM) (12), Pathway Commons (13), and rat QTLs from the Rat Genome Database (RGD) (14). Gene sets included in the initial 2012 publication have been augmented or refined based on changes in the data sources. A complete list of included data sources is available in Supplement Table S1.

Species are added based on criteria that include user requests, position of the organism as a model disease platform, existence of adequate functional genomic data sources and availability of a stable genome build, with preference given to species represented by an active annotation consortium. GeneWeaver currently houses data on nine species: *Macaca mulatta*, *Canis familiaris*, *Drosophila melanogaster*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Caenorhabditis elegans*, *Gallus gallus*, *Saccharomyces cerevisiae* and *Homo sapiens*. Genes associated with these species are derived from a variety of sources, including primary model organism databases (Rat Genome Database (RGD) (14), FlyBase (15), WormBase (16), ZFIN (17), Saccharomyces Genome Database (SGD) (18)), NCBI, Ensembl, Mammalian Genome Institute (MGI), and Uniprot. Collectively, GeneWeaver contains a total of more than four million external reference identifiers, which translates to over three million unique GeneWeaver identifiers mapped onto 29 266 homolog clusters based on homologene-based alignments (19).

NEW DEVELOPMENTS

Analysis tools

Tools for the identification of potentially informative biological entities among sets of intersecting heterogeneous data are continually evaluated, upgraded and made more efficient, with preference given to scalable but exhaustive solutions over mere heuristic approaches, and the ability of new tools to provide interpretable results through an intuitive user interface. Notably, complete bipartite Hierarchical Similarity (HiSim) graphs can be constructed of 100s or 1000s of sets of genes, which produce enormous graphs. These graphs now include bootstrap algorithms that sample edges within result sets to remove underrepresented edges and nodes, greatly reducing complexity. Visualizations have been enhanced to color nodes based on pre-selected emphasis genes, dynamically selected genes, or set similarity to existing user-defined sets, Figure 1.

Multi-partite relationships

As the number and variety of data sources continue to expand, it is evident that relevant biological associations may only be apparent through the intersection of multiple partitions. For example, one may wish to identify genes with a maximal association to MeSH derived gene sets and gene sets annotated to empirical cocaine experiments. In order to account for high order partition associations, GeneWeaver has recently adopted the bipartite gene association graph to include multi-partite sets of genes (20). Users are able to select each partite set based on sets of genes associated with a project. Edges between partite sets can be created by setting a minimum threshold of Jaccard overlap between individual sets of genes contained with each set. Alternatively, edges between sets can be created based on shared genes within each. Results highlight common and unique genes and gene sets associated with the underlying partite sets. This provides a means through which a prospective analysis of multi-way set intersections can be performed, potentially aligning semantic and data driven ontologies through mediating sets of genes.

User interface

To take advantage of the benefits of increased interoperability, stability, flexibility and modularity in modern web-based platforms, the GeneWeaver interface has been wholly redesigned based on python and the flask microframework, leveraging its native jinja2 template agent, RESTful request dispatching, Web Service Gateway Interface (WSGI) 1.0 compliance and secure session settings. The overall look and feel has been intentionally streamlined to highlight analysis functionality and user operations around the *Gene Set* metaphor. Thus, each operation functions on a set of genes.

Data sharing

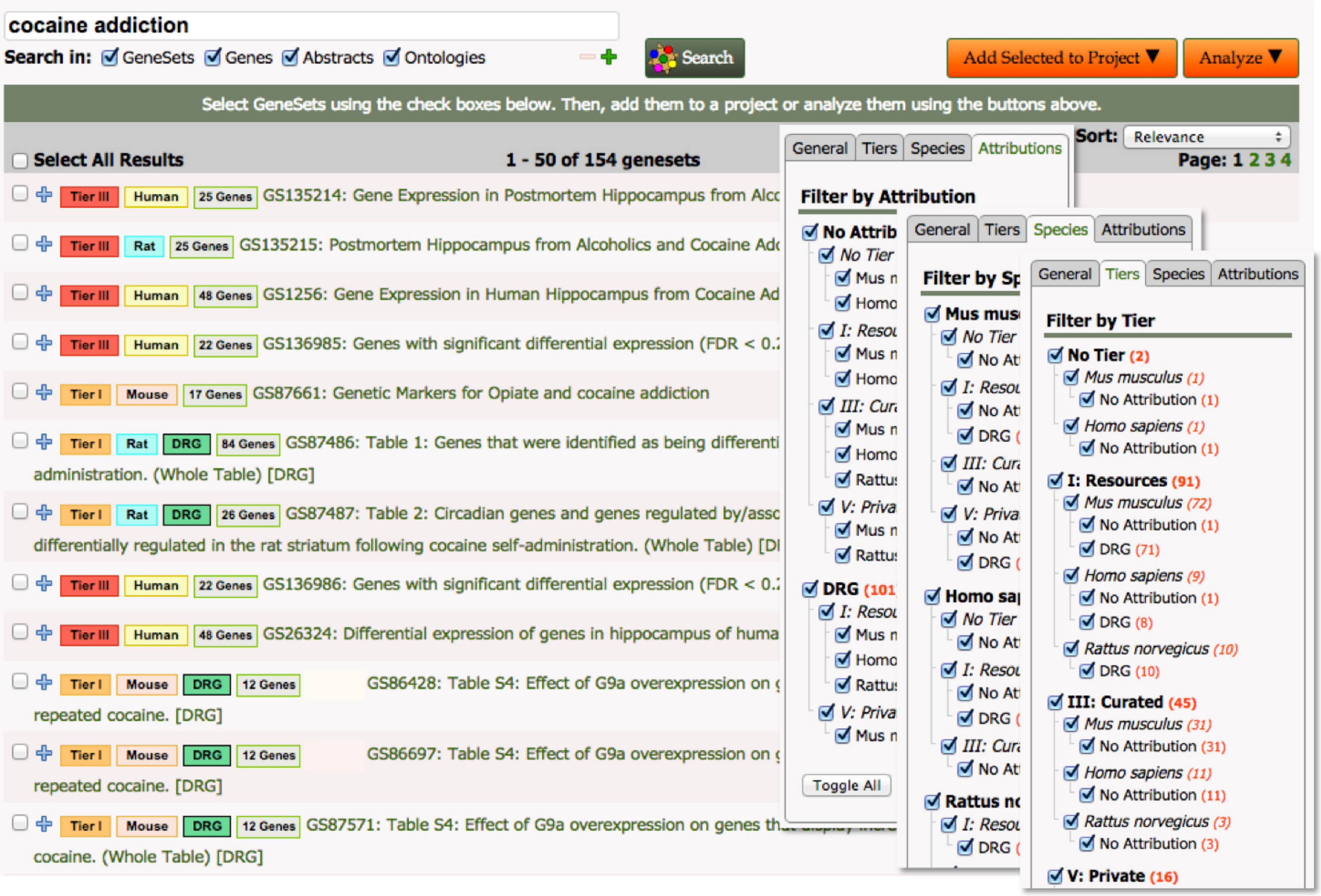
Expanded capabilities for user sharing, user-driven group administration and project sharing allow users to share access to data and to analysis results, so that a team of investigators can collaborative on pre-publication data, ultimately releasing the data for public access. Improved flexibility in this system allows users to work with specific collaborators within a session, and to transfer work to another group member as the project and team evolve.

Search

Because GeneWeaver is designed to present real time data query and analysis, the web interface has been optimize to search sets of genes rapidly, based on meta data, size, ontological associations, related publications and curatorial text. We have adopted Sphinx, a cross-file format indexer based on reStructuredText (rest) extensible parsing language (http://www.sphinxsearch.com). Search results can be organized by species, curation tier or attribution type, and filtered by set size, status or other attributes (Figure 2).

Documentation

GeneWeaver has reconfigured its documentation within a wiki (GeneWeaver.org/wiki). Here, users can find tutorials, a quick start guide (Supplement Figure S1) and details



cocaine addiction

Search in: ☑ GeneSets ☑ Genes ☑ Abstracts ☑ Ontologies

Select GeneSets using the check boxes below. Then, add them to a project or analyze them using the buttons above.

□ Select All Results 1 - 50 of 154 genesets

Sort: Relevance

Page: 1 2 3 4

- □ Tier III | Human | 25 Genes | GS135214: Gene Expression in Postmortem Hippocampus from Alc
- □ Tier III | Rat | 25 Genes | GS135215: Postmortem Hippocampus from Alcoholics and Cocaine Ad
- □ Tier III | Human | 48 Genes | GS1256: Gene Expression in Human Hippocampus from Cocaine Ad
- □ Tier III | Human | 22 Genes | GS136985: Genes with significant differential expression (FDR < 0.
- □ Tier I | Mouse | 17 Genes | GS87661: Genetic Markers for Opiate and cocaine addiction
- □ Tier I | Rat | DRG | 84 Genes | GS87486: Table 1: Genes that were identified as being differenti administration. (Whole Table) [DRG]
- □ Tier I | Rat | DRG | 26 Genes | GS87487: Table 2: Circadian genes and genes regulated by/assc differentially regulated in the rat striatum following cocaine self-administration. (Whole Table) [DI
- □ Tier III | Human | 22 Genes | GS136986: Genes with significant differential expression (FDR < 0.
- □ Tier III | Human | 48 Genes | GS26324: Differential expression of genes in hippocampus of huma
- □ Tier I | Mouse | DRG | 12 Genes | GS86428: Table S4: Effect of G9a overexpression on repeated cocaine. [DRG]
- □ Tier I | Mouse | DRG | 12 Genes | GS86697: Table S4: Effect of G9a overexpression on repeated cocaine. [DRG]
- □ Tier I | Mouse | DRG | 12 Genes | GS87571: Table S4: Effect of G9a overexpression on genes th cocaine. (Whole Table) [DRG]

Filter by Tier

- ☑ No Tier (2)
  - ☑ *Mus musculus* (1)
    - ☑ No Attribution (1)
  - ☑ *Homo sapiens* (1)
    - ☑ No Attribution (1)
- ☑ I: Resources (91)
  - ☑ *Mus musculus* (72)
    - ☑ No Attribution (1)
    - ☑ DRG (71)
  - ☑ *Homo sapiens* (9)
    - ☑ No Attribution (1)
    - ☑ DRG (8)
  - ☑ *Rattus norvegicus* (10)
    - ☑ DRG (10)
- ☑ III: Curated (45)
  - ☑ *Mus musculus* (31)
    - ☑ No Attribution (31)
  - ☑ *Homo sapiens* (11)
    - ☑ No Attribution (11)
  - ☑ *Rattus norvegicus* (3)
    - ☑ No Attribution (3)
- ☑ V: Private (16)

Figure 2. Search results can be organized by species, curation tiers and attribution metadata, and filtered by gene set size and status and group privileges. In this example, a search for 'cocaine addiction' returns 100 sets of genes, predominantly from rat, mouse and human. These are mostly Tier 1 and III data from published sources or experimentally derived, respectively. Numerous sets are also identified as being from the drug-related gene database (DRG).

about each tool and data set. The wiki also contains updated curation standards and instructions for connecting GeneWeaver tools to other community resources.

Data access and web services

GeneWeaver has adopted a REpresentational State Transfer (REST)-ful web services model that allows programmatic interaction with underlying data sets, such as data export (21). These dynamic facets are designed to support direct query of all analysis tools, including job status, stored results and other data. Data is returned as result image or JavaScript Object Notion (JSON).

Annotation to OBI and other widely used ontologies

GeneWeaver gene sets and networks are each annotated with appropriate ontologies. Data curated from individual published studies are among the hardest to annotate, but have been previously supported with extensive free text documentation. We have also recently initiated the formal annotation of these sets to terms in the OBI ontology (8).

CONCLUSIONS

With a greatly increased repository of background gene sets, GeneWeaver enables its users to perform a tremendous variety of applications directed at emerging questions in the comparison and prioritization of genes and variants and their role in disease (9). GeneWeaver is maintained under active development and continues to move towards web services, big data storage and computation paradigms, and intentionally maintained curatorial groups. The addition of updated user interfaces, new search features and analysis tools positions GeneWeaver for continued growth and use within the community. The foundational approach supported by the GeneWeaver model, namely, that of finding consilience among cross-species heterogeneous data, has produced numerous success stories, where data analysis explicitly informs hypothesis creation *in vitro* and *in vivo*.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

NIH [R01 AA18776] Jointly Funded by NIAAA and NIDA. Funding for open access charge: NIH [R01 AA18776].

Conflict of interest statement. None declared.

REFERENCES

1. Baker,E.J., Jay,J.J., Bubier,J.A., Langston,M.A. and Chesler,E.J. (2012) GeneWeaver: a web-based system for integrative functional genomics. *Nucleic Acids Res.*, **40**, D1067–D1076.
2. Baker,E.J., Jay,J.J., Philip,V.M., Zhang,Y., Li,Z., Kirova,R., Langston,M.A. and Chesler,E.J. (2009) Ontological Discovery Environment: a system for integrating gene-phenotype associations. *Genomics*, **94**, 377–387.
3. Zhang,Y., Phillips,C.A., Rogers,G.L., Baker,E.J., Chesler,E.J. and Langston,M.A. (2014) On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. *BMC Bioinformatics*, **15**, 110.
4. Kramer,M., Dutkowski,J., Yu,M., Bafna,V. and Ideker,T. (2014) Inferring gene ontologies from pairwise similarity data. *Bioinforma. Oxf. Engl.*, **30**, i34–i42.
5. Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
6. Groza,T., Köhler,S., Moldenhauer,D., Vasilevsky,N., Baynam,G., Zemojtel,T., Schriml,L.M., Kibbe,W.A., Schofield,P.N., Beck,T. *et al.* (2015) The human phenotype ontology: semantic unification of common and rare disease. *Am. J. Hum. Genet.*, **97**, 111–124.
7. Smith,C.L. and Eppig,J.T. (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **1**, 390–399.
8. Brinkman,R.R., Courtot,M., Derom,D., Fostel,J.M., He,Y., Lord,P., Malone,J., Parkinson,H., Peters,B., Rocca-Serra,P. *et al.* (2010) Modeling biomedical experimental processes with OBI. *J. Biomed. Semant.*, **1**(Suppl. 1), S7.
9. Bubier,J.A., Phillips,C.A., Langston,M.A., Baker,E.J. and Chesler,E.J. (2015) GeneWeaver: finding consilience in heterogeneous cross-species functional genomics data. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.*, doi:10.1007/s00335-015-9575-x.
10. Du,J., Yuan,Z., Ma,Z., Song,J., Xie,X. and Chen,Y. (2014) KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Mol. Biosyst.*, **10**, 2441–2447.
11. Liberzon,A. (2014) A description of the Molecular Signatures Database (MSigDB) Web site. *Methods Mol. Biol. Clifton NJ*, **1150**, 153–160.
12. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
13. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,O., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.
14. Shimoyama,M., De Pons,J., Hayman,G.T., Laulederkind,S.J.F., Liu,W., Nigam,R., Petri,V., Smith,J.R., Tutaj,M., Wang,S.-J. *et al.* (2015) The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res.*, **43**, D743–D750.
15. Dos Santos,G., Schroeder,A.J., Goodman,J.L., Strelets,V.B., Crosby,M.A., Thurmond,J., Emmert,D.B., Gelbart,W.M. and FlyBase Consortium (2015) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.*, **43**, D690–D697.
16. Harris,T.W., Baran,J., Bieri,T., Cabunoc,A., Chan,J., Chen,W.J., Davis,P., Done,J., Grove,C., Howe,K. *et al.* (2014) WormBase 2014: new views of curated biology. *Nucleic Acids Res.*, **42**, D789–D793.
17. Ruzicka,L., Bradford,Y.M., Frazer,K., Howe,D.G., Paddock,H., Ramachandran,S., Singer,A., Toro,S., Van Slyke,C.E., Eagle,A.E. *et al.* (2015) ZFIN, The zebrafish model organism database: Updates and new directions. *Genes. N. Y. N 2000*, **53**, 498–509.
18. Costanzo,M.C., Engel,S.R., Wong,E.D., Lloyd,P., Karra,K., Chan,E.T., Weng,S., Paskov,K.M., Roe,G.R., Binkley,G. *et al.* (2014) Saccharomyces genome database provides new regulation data. *Nucleic Acids Res.*, **42**, D717–D725.
19. NCBI Resource Coordinators (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
20. Phillips,C.A., Wang,K, Bubier,J., Baker,E.J., Chesler,E.J. and Langston,M.A. (2015) Scalable multipartite subgraph enumeration for integrative analysis of heterogeneous experimental functional genomics data. In: *ACM International Workshop on Big Data in Life Sciences*, Atlanta, GA.
21. Pautasso,C., Zimmermann,O. and Leymann,F. (2008) Restful web services vs. big'web services: making the right architectural decision. In: *Proceedings of the 17th International Conference on World Wide Web*. ACM, pp. 805–814.