Published online 12 November 2011 Nucleic Acids Research, 2012, Vol. 40, Database issue D1067–D1076
doi:10.1093/nar/gkr968

GeneWeaver: a web-based system for integrative functional genomics

Erich J. Baker[1], Jeremy J. Jay[2], Jason A. Bubier[2], Michael A. Langston[3] and Elissa J. Chesler[2,*]

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

Received August 15, 2011; Revised October 4, 2011; Accepted October 14, 2011

ABSTRACT

High-throughput genome technologies have produced a wealth of data on the association of genes and gene products to biological functions. Investigators have discovered value in combining their experimental results with published genome-wide association studies, quantitative trait locus, microarray, RNA-sequencing and mutant phenotyping studies to identify gene-function associations across diverse experiments, species, conditions, behaviors or biological processes. These experimental results are typically derived from disparate data repositories, publication supplements or reconstructions from primary data stores. This leaves bench biologists with the complex and unscalable task of integrating data by identifying and gathering relevant studies, reanalyzing primary data, unifying gene identifiers and applying *ad hoc* computational analysis to the integrated set. The freely available GeneWeaver (http://www.GeneWeaver.org) powered by the Ontological Discovery Environment is a curated repository of genomic experimental results with an accompanying tool set for dynamic integration of these data sets, enabling users to interactively address questions about sets of biological functions and their relations to sets of genes. Thus, large numbers of independently published genomic results can be organized into new conceptual frameworks driven by the underlying, inferred biological relationships rather than a pre-existing semantic framework. An empirical 'ontology' is discovered from the aggregate of experimental knowledge around user-defined areas of biological inquiry.

INTRODUCTION

An increasing number of investigators have integrated genome-wide experimental data across studies (1–5), but have lacked the data resources, algorithms and tools for widespread deployment of this approach. The data comes largely from gene expression microarray and now RNA-sequencing experiments, quantitative trait locus (QTL) mapping and genome-wide association studies (GWAS). Additional functional genomic data come from mutation and perturbation screen analyses (6–8) and from the vast array of individually curated biological relationships that have been associated with an array of biological ontology terms (e.g. (9–11). There are many powerful applications of this strategy, including meta-analysis or convergent analysis across microarrays (4), the refinement of QTL positional candidates through gene-centered evidence (2,12), cross-species translation of gene expression (13), mapping model organism findings onto human disease (14), using similarity of functional associations to identify the function of poorly characterized genes and using the similarity of biological associations to identify relations among biological processes.

GENEWEAVER: A DATA REPOSITORY AND ANALYSIS SYSTEM FOR FUNCTIONAL GENOMIC DATA INTEGRATION

GeneWeaver, powered by the Ontological Discovery Environment (15), is a freely available web-based software system that enables biologists to perform integrative functional genomics on their own collections of experimental data in combination with the system's incorporated database. The system enables users to integrate genome-wide experimental studies across multiple protocols and species (1–5), using convergent evidence to find consensus among the noise in reported associations of genes and biological functions and to classify functions

*To whom correspondence should be addressed. Tel: +1 207 288 6000; Fax: +1 207 288 6847; Email: Elissa.chesler@jax.org
The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

based on their underlying biological entities. The system addresses a major hurdle to the widespread implementation of integrative functional genomics by integrating a curated data repository with accompanying analysis tools (16).

The GeneWeaver data repository currently contains over 48 000 gene sets consisting of over 80 000 genes from seven species, derived largely from gene expression microarrays, RNA-sequencing experiments, QTL mapping and GWAS, mutation and perturbation screen (6–8), and curated biological relationships (e.g. (9–11)). In addition, GeneWeaver integrates other large community data sources, such as the drug related gene database of the Neuroscience Information Framework (NIF) (17), GeneNetwork (18) and the Comparative Toxicogenomics Database (CTD) (19). Inclusion of these diverse data resources, often with incompatible formats, demonstrates how GeneWeaver can promote the identification of common biological processes of poorly annotated genes through data integration without the overhead of extensive reanalysis.

Unlike ontology over-representation analysis systems [e.g. DAVID (20)] that match a single gene set result to many curated associations of genes to ontology terms or pathways, GeneWeaver infers relations among genes and processes by set–set matching across all gene sets in a query. This enables users to examine similar and distinct results among a large number of related experiments (including those compiled into ontology annotations), and to compare the role of genes to each of the functions assessed in the various experiments.

At its core, GeneWeaver achieves this data convergence through use of discrete gene-function associations, where gene lists are converted into an edge list of gene-to-biological function relationships where gene set descriptors are loosely defined as a ‘biological function.’ The resulting pair-wise relationships between genes and functions are queried using multivariate statistical methods and novel algorithms based on graph data mining including biclique, clique and paraclique analysis, clustering of the similarity matrix, Jaccard and hypergeometric tests, graph clustering algorithms such as dominating set and a novel algorithm for inference of an empirically derived ontology of gene functional relations, the Phenome Map (15).

APPLICATIONS AND USE OF GENEWEAVER

A visitor to GeneWeaver lands on a public page with links to data management, advanced search and analysis tools. All users have access to numerous tutorials, including videos, How To guides and our standards documentation (see Supplemental Data, http://geneweaver.org/index.php?action = help). While anonymous users have access to the GeneWeaver’s curated data repository, an account is required to store uploaded data, view data shared directly among registered users and their groups and to create and share individual projects. The system consists of a database containing curated gene set data and community metadata, both of which interact with a variety of discovery and data management tools (Figure 1). A user of the site typically begins their analysis with a search for a term or gene of interest. When a user enters a free text term, the default is a full-text search of meta-data fields including descriptions, publication information and NCBO Annotator (21) derived MeSH and Disease Ontology (22) terms. When a user searches for a gene symbol or other gene identifier, any gene set containing that gene symbol, its homologs or any identifier mapped onto that identifier by major model organism databases is retrieved. Users have tremendous flexibility in specifying searches through field-restricted queries and Boolean search.

For example, a search for *‘Drd2’* retrieves gene sets including expression in the lateral septal complex from the Allen Mouse Brain Atlas (GS127928), positional candidates for mechanical sensitivity and for methamphetamine response in a mouse QTL analyses (GS84202), QTL analysis for alcohol consumption (GS84212, GS84216), mutant phenotypic alleles for abnormal alcohol consumption from Mouse Genome Informatics (GS87659), Gene Ontology annotations for post synaptic density (GS97454), paraquat interacting genes from CTD (GS121514) and 712 other gene sets. A query for ethanol retrieves 540 Gene Sets including genetic studies of alcoholism in humans (e.g. GS46979, GS46980, GS46982, GS46984, GS46985), mice (GS37188) and rats (GS37196, GS37197). Clicking on a Gene Set name opens a detail page that enables users to view the metadata or contents of a gene set, translate the identifiers, execute overrepresentation analyses via GAGGLE (23) and find similar gene sets. Users can select appropriate gene sets using check boxes and add them to a ‘Project’ using ‘Add Selected to Project.’ The user should then select ‘Analyze Gene Sets’ to perform integrative analyses using tools described below (Figure 2). The forms under ‘Manage Gene Sets’ allow users to upload their own gene sets and add them to new or existing projects for seamless integrative analysis with the data in the repository.

GENEWEAVER ANALYSIS TOOLS AND FUNCTIONS

User selected gene sets are combined through identifier cross-mapping to enable analysis of discrete relationships between genes and functions using a wide variety of tools. Each of the embedded tools functions independently to explore and prioritize genes and gene sets of interest. The inputs and outputs are described as genes or gene sets (Table 1). The modular strategy allows users to arrange tools in numerous distinct workflows. Most of these tools operate on the representation of a group of gene lists as a bipartite graph consisting of genes as one set of vertices and gene sets as a second set of vertices. The edges between these vertices are weighted by scores or unweighted discrete values (15). To execute an analysis from the ‘Analyze Gene Sets’ page, first select a project using the check box next to its name, or expand the project to select individual gene sets within it.

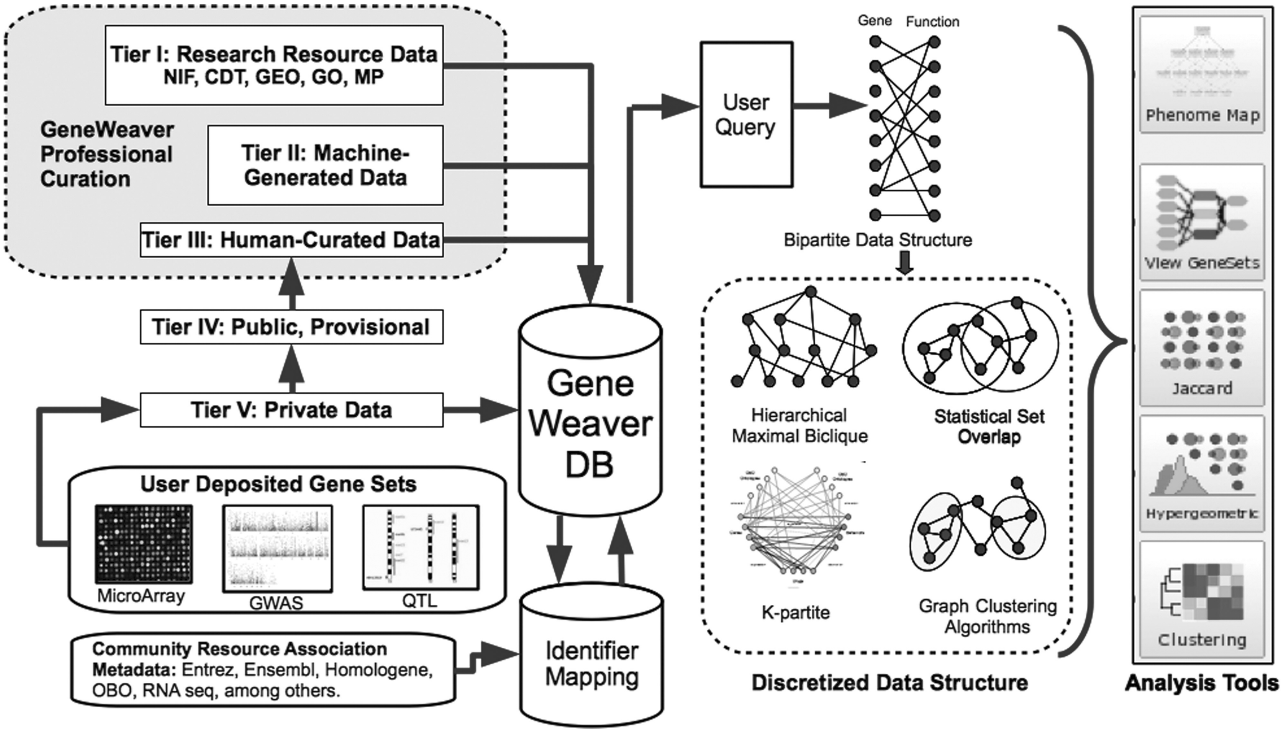

Tier I: Research Resource Data
NIF, CDT, GEO, GO, MP
GeneWeaver Professional Curation
Tier II: Machine-Generated Data
Tier III: Human-Curated Data
Tier IV: Public, Provisional
Tier V: Private Data
User Deposited Gene Sets
MicroArray
GWAS
QTL
Community Resource Association Metadata: Entrez, Ensembl, Homologene, OBO, RNA seq, among others.
Gene Weaver DB
Identifier Mapping
User Query
Gene
Function
Bipartite Data Structure
Hierarchical Maximal Biclique
Statistical Set Overlap
K-partite
Graph Clustering Algorithms
Discretized Data Structure
Phenome Map
View GeneSets
Jaccard
Hypergeometric
Clustering
Analysis Tools

Figure 1. Curation and integrative analysis of secondary data in the Ontological Discovery Environment. The overall system architecture consists of a centralized database that collects a variety of curated data and metadata and serves a suite of analysis tools. It uses a data from community resources to create clusters of gene homology across supported species, enabling ODE to rapidly translate gene sets.

Then select a tool using the icons to the left. Clicking a tool icon immediately executes the default analysis, but tool options may be set by first expanding the options by clicking on the plus sign next to the icon. Most tools have options for the inclusion of homologous genes and for setting thresholds and analysis parameters. Several tools also allow users to highlight nodes containing ‘emphasis genes,’ a user specified list of genes of particular interest.

The gene set graph

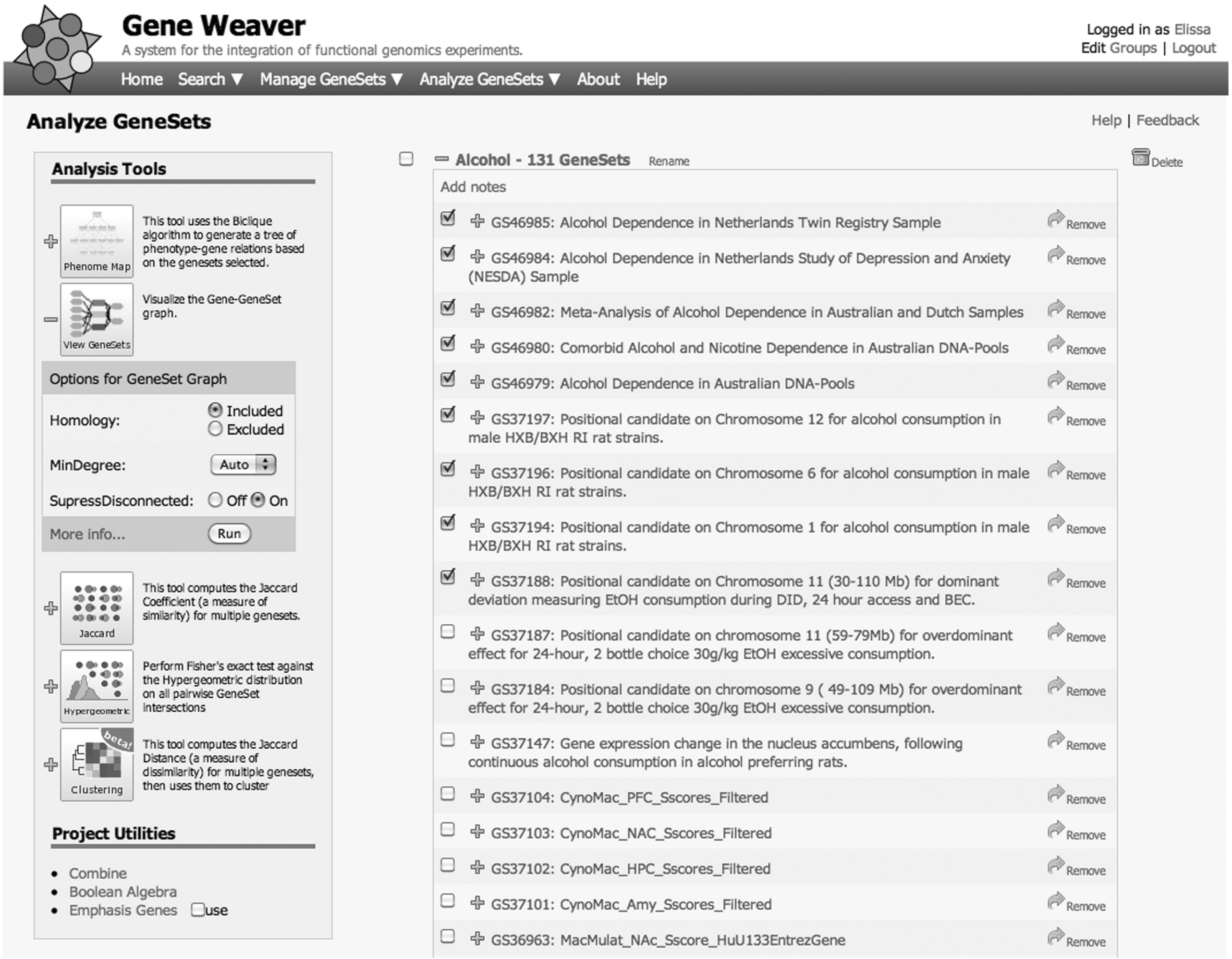
The gene set graph tool is a module that enables users to identify genes and gene sets that are highly connected among an input group of gene sets. Gene Sets are represented as a central column of vertices, with high degree (highly connected) gene vertices plotted to the right and lower degree gene vertices plotted to the left (Figure 3, Supplementary Figure S1). Users control the minimum degree threshold and can incorporate the ‘emphasis genes’ feature.

The phenome graph

The phenome graph is a hierarchical network of multiway gene set intersections (Figure 4, Supplementary Figure S2). The result is a graph of gene set intersections of very high order, enabling users to find genes connected to all populated subsets of an input set of gene lists. In algorithmic terms, these intersections are created from the overlap of maximal bicliques in discrete bipartite structures (15). This tool creates an ‘empirical ontology’ of function based on the underlying genes within curated and user-defined gene sets in that similar gene sets, regardless of semantic annotation, are joined through similarity of their contents. Our fast implementation of bipartite algorithms (24) allows GeneWeaver to create very large hierarchies from over 100 typical inputs and our integration of homology and identifier translation tables enables integration across species and experimental systems. To ensure robustness of the result and reduce the size of the output graph, users may prune resulting hierarchies using bootstrapping procedures and stopping rules based on leaf contents (such as the minimum number of genes or gene sets in each node) or graph depth.

Anchored bicliques of biomolecular associates

The ABBA tool is an analysis module used to find genes that have similar functional associations to the members of a query set of genes. When a list of genes is input into ABBA, the tool generates a list of all gene sets from within the GeneWeaver database that contain a user determined number of the input genes. ABBA then returns a ranked list of similar genes that are enriched among the same Gene Sets as the input genes (see Supplementary Figures S3 and S4 for sample use cases). The user may designate a connectivity threshold for the number of gene sets that contain the predicted genes. This tool, for example, has been used to successfully identify genes that may be



Gene Weaver
A system for the integration of functional genomics experiments.

Logged in as Elissa
Edit Groups | Logout

Home | Search ▼ | Manage GeneSets ▼ | Analyze GeneSets ▼ | About | Help

Analyze GeneSets

Help | Feedback

Analysis Tools

Phenome Map: This tool uses the Biclique algorithm to generate a tree of phenotype-gene relations based on the genesets selected.

View GeneSets: Visualize the Gene-GeneSet graph.

Options for GeneSet Graph

Homology: ● Included ○ Excluded

MinDegree: Auto

SupressDisconnected: ○ Off ● On

More info... Run

Jaccard: This tool computes the Jaccard Coefficient (a measure of similarity) for multiple genesets.

Hypergeometric: Perform Fisher's exact test against the Hypergeometric distribution on all pairwise GeneSet intersections

Clustering (beta!): This tool computes the Jaccard Distance (a measure of dissimilarity) for multiple genesets, then uses them to cluster

Project Utilities

- Combine
- Boolean Algebra
- Emphasis Genes □use

□ Alcohol - 131 GeneSets Rename | Delete

Add notes

☑ GS46985: Alcohol Dependence in Netherlands Twin Registry Sample — Remove
☑ GS46984: Alcohol Dependence in Netherlands Study of Depression and Anxiety (NESDA) Sample — Remove
☑ GS46982: Meta-Analysis of Alcohol Dependence in Australian and Dutch Samples — Remove
☑ GS46980: Comorbid Alcohol and Nicotine Dependence in Australian DNA-Pools — Remove
☑ GS46979: Alcohol Dependence in Australian DNA-Pools — Remove
☑ GS37197: Positional candidate on Chromosome 12 for alcohol consumption in male HXB/BXH RI rat strains. — Remove
☑ GS37196: Positional candidate on Chromosome 6 for alcohol consumption in male HXB/BXH RI rat strains. — Remove
☑ GS37194: Positional candidate on Chromosome 1 for alcohol consumption in male HXB/BXH RI rat strains. — Remove
☑ GS37188: Positional candidate on Chromosome 11 (30-110 Mb) for dominant deviation measuring EtOH consumption during DID, 24 hour access and BEC. — Remove
□ GS37187: Positional candidate on chromosome 11 (59-79Mb) for overdominant effect for 24-hour, 2 bottle choice 30g/kg EtOH excessive consumption. — Remove
□ GS37184: Positional candidate on chromosome 9 (49-109 Mb) for overdominant effect for 24-hour, 2 bottle choice 30g/kg EtOH excessive consumption. — Remove
□ GS37147: Gene expression change in the nucleus accumbens, following continuous alcohol consumption in alcohol preferring rats. — Remove
□ GS37104: CynoMac_PFC_Sscores_Filtered — Remove
□ GS37103: CynoMac_NAC_Sscores_Filtered — Remove
□ GS37102: CynoMac_HPC_Sscores_Filtered — Remove
□ GS37101: CynoMac_Amy_Sscores_Filtered — Remove
□ GS36963: MacMulat_NAc_Sscore_HuU133EntrezGene — Remove

Figure 2. The analyze gene sets page. GeneWeaver's analysis functions are accessed from this page. Gene sets must first be collected and stored into one or more projects by the user. In this case, a project called 'Alcohol' contains 121 gene sets, nine of which are selected for analysis using the tools on the right. Options can be selected from this tool bar prior to executing the tool.

functionally similar to genes associated with autism within the MGI database (25).

Gene set similarity

Gene Set similarity is estimated for a single set against the entire database, executed from an individual gene set page, or it can be analyzed using pair-wise similarity tools, Jaccard Similarity (Supplementary Figure S5) and Hypergeometric test. These two tools operate on a user selected set of gene sets and produce both a matrix of similarity statistics and a matrix of pair-wise Venn diagrams for all Gene Sets in the analysis. Each operates on a reference gene set consisting only of those genes which can be mapped across identifiers from one gene set to the other.

Boolean gene set functions

Although our bipartite graph algorithms are both complete and scalable, the ability to handle large query results, construct inputs, interpret and visualize large graphical results is still a challenge for human users. A set of Boolean Gene Set Logic features enable users to reduce large query results. For example, a user may identify a set of overlapping quantitative trait loci, each for a related biological phenotype. A common positional candidate is most likely, so using the gene set intersection feature, users can reduce the group of QTL candidate gene lists to a single gene set containing only those genes in a high order intersection of the QTLs. Likewise, users may want to distill large sets of genes for comparison of related functions. By taking the union of all genes that are represented among groups of user selected gene sets, comparisons of smaller numbers of gene sets can be performed. For example, to identify the similarities and differences among genes associated with broad classes of psychological disorders, a user could take the union of all genes associated with major classes of conditions: mood disorders, alcoholism, anxiety disorders and chronic stress and then use the Phenome Graph function to examine high-order intersections among them (Supplementary Figure S6).

Table 1. Analysis tools and basic functions available in GeneWeaver

Analysis tools	Gene sets		Genes		Input	Output	Description	Settings
	Explore	Prioritize	Explore	Prioritize				
Anchored bicliques of biomolecular associates	X		X	X	Genes	Highly similar genes, highly connected phenotypes	Find genes that are connected to similar genes sets as your input genes.	Degree threshold
Similar gene sets	X				1 Gene set	Ranked list of similar gene sets (top 500)	Rank all gene sets by similarity to a given gene set.	
Phenome map		X	X	X	Gene sets	Graph of hierarchical intersections of gene sets	Generate directed acyclic graph of intersecting gene sets.	Bootstraps Permutations Stopping rules
Gene set graph		X	X	X	Gene sets	High degree connectivity of genes to your gene sets	Plot the bipartite graph of gene sets.	Degree threshold
Jaccard similarity		X		X	Gene sets	Matrix of pair-wise Venn diagrams and Jaccard similarity scores	Pair-wise similarity of gene sets.	
Hypergeometric tests		X		X	Gene sets	Matrix of pair-wise Venn diagrams and Jaccard similarity scores	Pair-wise similarity of gene sets.	
Jaccard clustering		X			Gene sets	Dendogram and cluster heat map revealing gene set similarity.	Similarity clustering of gene sets.	Clustering method
Combine	X				Gene sets	Adjaceny matrix	Create a matrix of genes and gene lists	
Emphasis genes				X	Genes		Highlight genes of interest on GeneSet graph and Phenome map output	
Boolean gene set logic	X				Gene sets	Gene set consisting of union, intersection or highly connected genes from a group of gene sets.	Condense a large number of gene sets into a single gene set based on connectivity, e.g. to find the intersection of multiple QTL positional candidates, or the union of all genes annotated to a set of related biological functions.	Connectivity threshold

All functions are available from the 'Analyze GeneSets' page, except for ABBA and 'Find similar gene sets,' which are accessed from the 'Search for Genes' menu and 'GeneSet Pages,' respectively. Emphasis genes may be accessed from either 'Analyze GeneSets' or any 'Gene Set Page'.

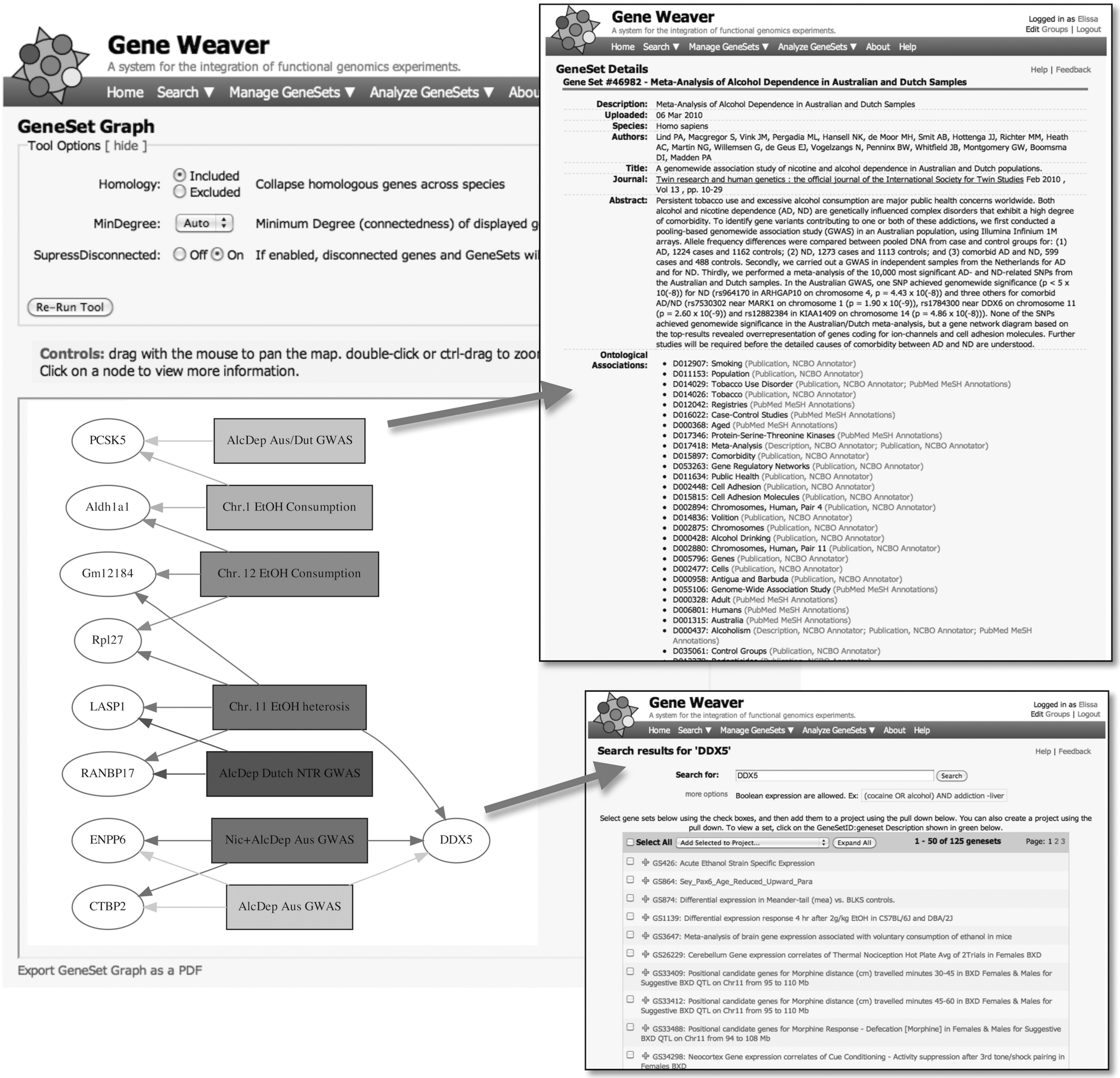
Gene set similarity clustering

GeneWeaver employs multivariate clustering of the Jaccard gene set similarity matrix to rapidly collapse large gene sets and arrange them into hierarchical clusters. A number of common clustering algorithms are included (Supplementary Figure S7). This tool enables users to rapidly identify groups of similar and dissimilar inputs that can be used to perform important quality control for redundant inputs among a set of user-submitted data.

DATA CURATION

It is critical for any bioinformatics resource to contain traceable data, ample documentation and a mechanism for reproducible results. The GeneWeaver database consists of curated data from a variety of sources (Supplementary Figures S9–S14, Supplementary Tables S4 and S5), and all data shared publicly on the site must pass curatorial review. As a final check on the accuracy of data in the database, users are given sufficient information to be able to retrieve or regenerate the underlying Gene Set and may submit their corrected version to our curator.

Each gene set in GeneWeaver is stored with descriptive metadata at several levels, including publication information, reference gene identifiers, free text descriptions and structured term annotations for disease, biological process and experimental context. Controlled description of the gene sets are user submitted by tree-based term navigation and selection from the Open Biomedical Ontologies [Mouse Anatomy, Gene Ontology, Mammalian Phenotype Ontology (26), MeSH]; additional terms are suggested using the NCBO Annotator tool. GeneWeaver's curator reviews these terms for accuracy. These levels provide human and machine interpretable information about each functional genomics experiment to describe the analysis process by which a result was originally generated. Free-text 'Gene Set names' succinctly describe the contents of a set. A fuller 'description' field provides a detailed description of the data generation and rules for inclusion on a gene set list. An 'abbreviation' field



GeneSet Graph

Tool Options [hide]

Homology: ● Included ○ Excluded Collapse homologous genes across species

MinDegree: Auto Minimum Degree (connectedness) of displayed g

SupressDisconnected: ○ Off ● On If enabled, disconnected genes and GeneSets wi

Re-Run Tool

Controls: drag with the mouse to pan the map. double-click or ctrl-drag to zoo
Click on a node to view more information.

Export GeneSet Graph as a PDF

GeneSet Details

Gene Set #46982 - Meta-Analysis of Alcohol Dependence in Australian and Dutch Samples

Description: Meta-Analysis of Alcohol Dependence in Australian and Dutch Samples
Uploaded: 06 Mar 2010
Species: Homo sapiens
Authors: Lind PA, Macgregor S, Vink JM, Pergadia ML, Hansell NK, de Moor MH, Smit AB, Hottenga JJ, Richter MM, Heath AC, Martin NG, Willemsen G, de Geus EJ, Vogelzangs N, Penninx BW, Whitfield JB, Montgomery GW, Boomsma DI, Madden PA
Title: A genomewide association study of nicotine and alcohol dependence in Australian and Dutch populations.
Journal: Twin research and human genetics : the official journal of the International Society for Twin Studies Feb 2010 , Vol 13 , pp. 10-29
Abstract: Persistent tobacco use and excessive alcohol consumption are major public health concerns worldwide. Both alcohol and nicotine dependence (AD, ND) are genetically influenced complex disorders that exhibit a high degree of comorbidity. To identify gene variants contributing to one or both of these addictions, we first conducted a pooling-based genomewide association study (GWAS) in an Australian population, using Illumina Infinium 1M arrays. Allele frequency differences were compared between pooled DNA from case and control groups for: (1) AD, 1224 cases and 1162 controls; (2) ND, 1273 cases and 1113 controls; and (3) comorbid AD and ND, 599 cases and 488 controls. Secondly, we carried out a GWAS in independent samples from the Netherlands for AD and for ND. Thirdly, we performed a meta-analysis of the 10,000 most significant AD- and ND-related SNPs from the Australian and Dutch samples. In the Australian GWAS, one SNP achieved genomewide significance (p < 5 x 10(-8)) for ND (rs964170 in ARHGAP10 on chromosome 4, p = 4.43 x 10(-8)) and three others for comorbid AD/ND (rs7530302 near MARK1 on chromosome 1 (p = 1.90 x 10(-9)), rs1784300 near DDX6 on chromosome 11 (p = 2.60 x 10(-9)) and rs12882384 in KIAA1409 on chromosome 14 (p = 4.86 x 10(-8))). None of the SNPs achieved genomewide significance in the Australian/Dutch meta-analysis, but a gene network diagram based on the top-results revealed overrepresentation of genes coding for ion-channels and cell adhesion molecules. Further studies will be required before the detailed causes of comorbidity between AD and ND are understood.
Ontological Associations:

- D012907: Smoking (Publication, NCBO Annotator)
- D011153: Population (Publication, NCBO Annotator)
- D014029: Tobacco Use Disorder (Publication, NCBO Annotator; PubMed MeSH Annotations)
- D014026: Tobacco (Publication, NCBO Annotator)
- D012042: Registries (PubMed MeSH Annotations)
- D016022: Case-Control Studies (PubMed MeSH Annotations)
- D000368: Aged (PubMed MeSH Annotations)
- D017346: Protein-Serine-Threonine Kinases (PubMed MeSH Annotations)
- D017418: Meta-Analysis (Description, NCBO Annotator; Publication, NCBO Annotator)
- D015897: Comorbidity (Publication, NCBO Annotator)
- D053263: Gene Regulatory Networks (Publication, NCBO Annotator)
- D011634: Public Health (Publication, NCBO Annotator)
- D002448: Cell Adhesion (Publication, NCBO Annotator)
- D015815: Cell Adhesion Molecules (Publication, NCBO Annotator)
- D002894: Chromosomes, Human, Pair 4 (Publication, NCBO Annotator)
- D014836: Volition (Publication, NCBO Annotator)
- D002875: Chromosomes (Publication, NCBO Annotator)
- D000428: Alcohol Drinking (Publication, NCBO Annotator)
- D002880: Chromosomes, Human, Pair 11 (Publication, NCBO Annotator)
- D005796: Genes (Publication, NCBO Annotator)
- D002477: Cells (Publication, NCBO Annotator)
- D000958: Antigua and Barbuda (Publication, NCBO Annotator)
- D055106: Genome-Wide Association Study (PubMed MeSH Annotations)
- D000328: Adult (PubMed MeSH Annotations)
- D006801: Humans (PubMed MeSH Annotations)
- D001315: Australia (PubMed MeSH Annotations)
- D000437: Alcoholism (Description, NCBO Annotator; Publication, NCBO Annotator; PubMed MeSH Annotations)
- D035061: Control Groups (Publication, NCBO Annotator)

Search results for 'DDX5'

Search for: DDX5 Search

more options Boolean expression are allowed. Ex: (cocaine OR alcohol) AND addiction -liver

Select gene sets below using the check boxes, and then add them to a project using the pull down below. You can also create a project using the pull down. To view a set, click on the GeneSetID:geneset Description shown in green below.

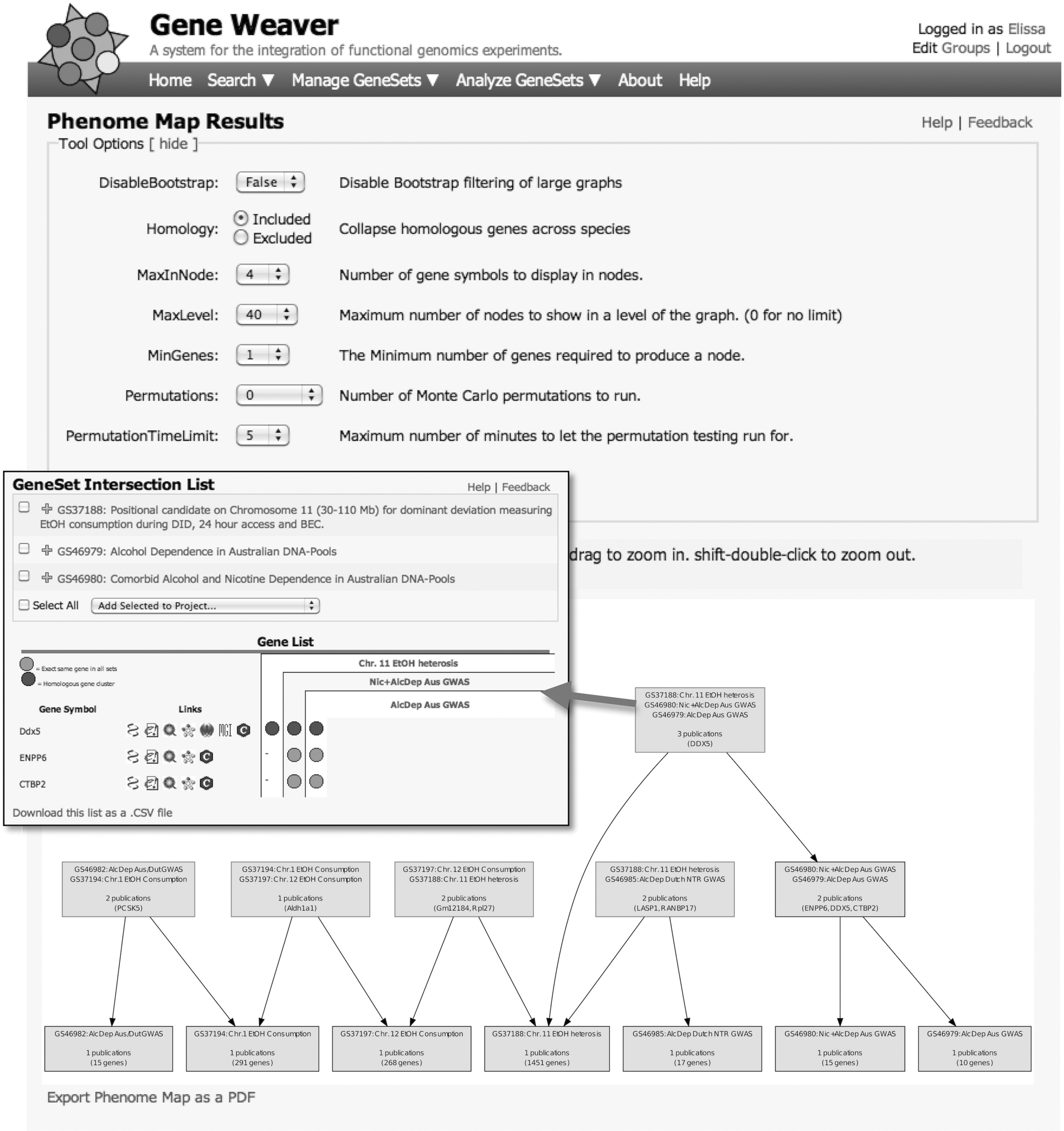
□ **Select All** Add Selected to Project... Expand All **1 - 50 of 125 genesets** Page: 1 2 3

- □ GS426: Acute Ethanol Strain Specific Expression
- □ GS864: Sey_Pax6_Age_Reduced_Upward_Para
- □ GS874: Differential expression in Meander-tail (mea) vs. BLKS controls.
- □ GS1139: Differential expression response 4 hr after 2g/kg EtOH in C57BL/6J and DBA/2J
- □ GS3647: Meta-analysis of brain gene expression associated with voluntary consumption of ethanol in mice
- □ GS26229: Cerebellum Gene expression correlates of Thermal Nociception Hot Plate Avg of 2Trials in Females BXD
- □ GS33409: Positional candidate genes for Morphine distance (cm) travelled minutes 30-45 in BXD Females & Males for Suggestive BXD QTL on Chr11 from 95 to 110 Mb
- □ GS33412: Positional candidate genes for Morphine distance (cm) travelled minutes 45-60 in BXD Females & Males for Suggestive BXD QTL on Chr11 from 95 to 110 Mb
- □ GS33488: Positional candidate genes for Morphine Response - Defecation [Morphine] in Females & Males for Suggestive BXD QTL on Chr11 from 94 to 108 Mb
- □ GS34298: Neocortex Gene expression correlates of Cue Conditioning - Activity suppression after 3rd tone/shock pairing in Females BXD

Figure 3. The gene set graph. The gene set graph reveals the highly connected genes among the nine gene sets selected in Figure 2. This analysis reveals DDX5 as the most highly connected gene, connected to both human and mouse alcohol-related measures. Inset: clicking on a gene node executes a search for gene sets containing the featured gene or its homologs. Clicking on a gene set node reveals the contents and metadata for that gene set.

provides a compact recognizable label for graphics. Finally, publication information is obtained to enable users to readily retrieve complete information about the underlying experiment.

GeneWeaver has a systematic curation process to supply validated secondary data for analysis tools. Internal curation is driven by directed research, largely related to alcoholism, addiction and behavioral disorders, but the workflow may be applied to multiple interest groups. From literature, data curation starts with term identification, which is subsequently crossed with genome experimentation key words such as ‘GWAS,’ ‘QTL,’ ‘microarray’ and ‘RNAseq.’ Combined terms are used in MEDLINE searches to identify publications which may contain secondary functional genomics results. Each publication is checked against the existing database and manually scanned for secondary data, including supplemental files. Appropriate data sets are uploaded along with descriptive metadata. Descriptions are written to contain sufficient information to enable users to readily retrieve the original source data and to access documentation for the individual studies.



Phenome Map Results

Help | Feedback

Tool Options [hide]

DisableBootstrap:	False	Disable Bootstrap filtering of large graphs
Homology:	● Included ○ Excluded	Collapse homologous genes across species
MaxInNode:	4	Number of gene symbols to display in nodes.
MaxLevel:	40	Maximum number of nodes to show in a level of the graph. (0 for no limit)
MinGenes:	1	The Minimum number of genes required to produce a node.
Permutations:	0	Number of Monte Carlo permutations to run.
PermutationTimeLimit:	5	Maximum number of minutes to let the permutation testing run for.

drag to zoom in. shift-double-click to zoom out.

GeneSet Intersection List

Help | Feedback

- □ GS37188: Positional candidate on Chromosome 11 (30-110 Mb) for dominant deviation measuring EtOH consumption during DID, 24 hour access and BEC.
- □ GS46979: Alcohol Dependence in Australian DNA-Pools
- □ GS46980: Comorbid Alcohol and Nicotine Dependence in Australian DNA-Pools

□ Select All   Add Selected to Project...

Gene List

○ = Exact same gene in all sets
● = Homologous gene cluster

Gene Symbol	Links	Chr. 11 EtOH heterosis	Nic+AlcDep Aus GWAS	AlcDep Aus GWAS
Ddx5	MGI	●	●	●
ENPP6		-	○	○
CTBP2		-	○	○

Download this list as a .CSV file

GS37188:Chr.11 EtOH heterosis
GS46980:Nic+AlcDep Aus GWAS
GS46979:AlcDep Aus GWAS
3 publications
(DDX5)
GS46982:AlcDep Aus/DutGWAS
GS37194:Chr.1 EtOH Consumption
2 publications
(PCSK5)
GS37194:Chr.1 EtOH Consumption
GS37197:Chr.12 EtOH Consumption
1 publications
(Aldh1a1)
GS37197:Chr.12 EtOH Consumption
GS37188:Chr.11 EtOH heterosis
2 publications
(Gm12184,Rpl27)
GS37188:Chr.11 EtOH heterosis
GS46985:AlcDep Dutch NTR GWAS
2 publications
(LASP1,RANBP17)
GS46980:Nic+AlcDep Aus GWAS
GS46979:AlcDep Aus GWAS
2 publications
(ENPP6,DDX5,CTBP2)
GS46982:AlcDep Aus/DutGWAS
1 publications
(15 genes)
GS37194:Chr.1 EtOH Consumption
1 publications
(291 genes)
GS37197:Chr.12 EtOH Consumption
1 publications
(268 genes)
GS37188:Chr.11 EtOH heterosis
1 publications
(1451 genes)
GS46985:AlcDep Dutch NTR GWAS
1 publications
(17 genes)
GS46980:Nic+AlcDep Aus GWAS
1 publications
(15 genes)
GS46979:AlcDep Aus GWAS
1 publications
(10 genes)

Export Phenome Map as a PDF

Figure 4. The phenome graph. The phenome graph drawn from nine inputs selected in Figure 2. The phenome graph is a directed acyclic graph of the intersections of gene sets. Each node represents gene sets and the genes they share. Higher order intersections are represented in the root nodes at the top, and individual gene sets in the leaves at the bottom. Inset: clicking a node opens a page showing the intersections among gene sets in list form. Results from this page can be sent to other tools for annotation, including GAGGLE.

The NCBO Annotator is then applied to suggest MeSH terms, Disease Ontology, Gene Ontology, Mammalian Phenotype Ontology and other structured vocabulary terms that may be appropriate to the data set. These terms are reviewed by the curator to comprise the final entry.

Gene sets may also be submitted by users for public access via the GeneWeaver database. Each of these is marked provisional until the curator reviews them by first searching for similar sets from the same publication, and searching for similar sets by contents to identify duplicate entries. The curator then determines whether the description is compliant with our standards document (see Supplementary Data), and finally assesses whether the gene set conforms to the literature and/or description provided. Finally, the NCBO Annotator is applied to identify structured annotations to the gene set. Ultimately, the user may wish to reanalyze data from the original source and evaluate the appropriateness of particular source data for their analyses. Sufficient details and links to external sites are provided for this purpose.

A large amount of GeneWeaver data comes from major bioinformatics resources including NCBI, ENSEMBL and various model organism databases, including MGD (9), Rat Genome Database [RGD (27)], HUGO Gene Nomenclature Committee [HGNC (28)], Saccharomyces Genome Database [SGD (29)], FlyBase (30), WormBase (31) and the Zebrafish Model Organism Database [ZFIN (32)]. Some of these data are converted to gene sets, including GO and MP annotations, Comparative Toxicogenomics Database (33) associations and QTL positional candidates from RGD and MGI. These data sources are updated every 6 months. A large accessory data warehouse contains the gene identifier mappings obtained from Homologene and the model organism databases. A table of update dates is found at http://geneweaver.org/index.php?action = help&cmd = updates, and acquisition dates of individual gene sets are indicated at the level of the gene set. When a newer version of a gene set in a user project exists, the older version is flagged as 'deprecated,' and the user has the opportunity to update the stored set to the latest version.

Each time the data warehouse is updated, gene identifiers are mapped onto one another using the mapping provided by each species' model organism database, and assigned a unique ID. Users specify the source species of their upload, and the uploaded list is compared to the gene identifiers in the database for that species to identify the appropriate mapping to gene identifiers. Gene symbols are ambiguous and users are cautioned not to use this method if at all possible. When an ambiguous symbol, i.e. one for which there are multiple gene ids from the Generic Model Organism Database (GMOD), is uploaded, the user is notified and all of the corresponding gene IDs are assigned to the set. This inherent noise in the user input is filtered out through convergent data analysis, provided that multiple instances of the precise identifier occur, whereas spurious associations are unlikely to recur by chance.

A multitiered system denotes the level of curation applied to each data set, allowing users to limit the scope of their analysis as desired. All data obtained from established community resources such as MP and GO annotations to genes and Kyoto Encyclopedia of Genes and Genomes pathway members are classified as Tier I, Public Research Resource data and is updated on a 6-month cycle. Pathway data (e.g. in KEGG), is converted to gene lists representing the gene products in each pathway. Gene Ontology annotations are converted to gene lists through transitive closure such that the genes annotated to any child node are also annotated to parent nodes successively up each branch of the ontology. A similar process is used to convert phenotypic alleles annotated to the Mammalian Phenotype Ontology. Each allele is assigned to the gene symbol for which it is allelic, and that symbol is associated to gene lists corresponding to each MPO term. Transitive closure is again used to assign gene symbols to parents of any child term in the MPO. Data version and accession date are integrated into the metadata for these records.

Tier II consists of human-curated machine-generated data, such as phenotype to gene correlation data in reference populations and QTL positional candidates or GWAS positional candidates. GeneWeaver's curation team examines the fidelity of these large data sets to our standards documentation and assigns appropriate metadata to them. Tier III data consists of curated results of individual functional experimental studies, typically obtained from publications. Tier IV, or provisional data, is submitted data, advanced by users for public sharing but still pending review. It is exposed to search and analytic modules, but can be filtered or excluded readily. Curators examine provisional data for its adherence to our standards documentation for metadata quality, redundancy with existing data or unusual threshold artifacts, before moving it to Tier III, human curated. Tier V consists of data explicitly for private, personal or group use and is not shared with the larger community.

Users and groups

In order to tightly couple analysis to the collation and curation of user-submitted data, ODE allows guests to create individual accounts, user-defined groups and associated projects. Automated graded-access is built into analysis tools to control data sharing across users and groups. Any gene set may be maintained as private and excluded from our curation Tiers, although it is strongly suggested that users make every effort to meet requirements outlined in the standards documentation (see Supplementay Data) to facilitate interpretation of analysis results by users and group members. Projects provide an effective means to sort and store related data sets and query results over time. Version control of data in user accounts is also provided. Data curation is a dynamic process. It is possible that user selected genes or gene sets may change status over time with some data being deprecated. Legacy data are not removed from the repository, but, data sets containing deprecated genes or gene-function associations are clearly noted and may or may not be updated at the user's discretion to maintain continuity and reproducibility of results through time.

Community contribution to the GeneWeaver repository

Users are encouraged to augment the repository by uploading data specific to their research question. Data upload is in the form of a simple two-column tab delimited text file consisting of one header row. The left column is a list of the gene symbols or identifiers, and the second column is a score that can later be used to threshold genes of interest. These scores may be the results of statistical analyses such as *P*-values, *q*-values, correlation metrics and effect sizes or they may be binary values representing list membership. The user is prompted to provide a threshold for the scores. This flexibility enables the aggregation of highly diverse results using combinatorial algorithms, while providing sufficient information for future development of more formal meta-analysis modules. During the upload process, users are also prompted to assign metadata. This step is imperative as it allows discovery of relevant gene sets for integration with existing data. Required metadata includes ontology assignment, PubMed IDs, a textual description of the data and appropriate labeling. If the data set is put forward as 'Public' it is immediately assigned to Tier IV and marked 'Provisional' until a curator can verify that the data and metadata meet our standards protocol for upgrade to Tier III.

RESOURCE DISSEMINATION

The GeneWeaver analysis system is made available on the GitHub community code repository under the GPL license. The system includes HTML, AJAX, SVG pages, a PostgreSQL database, Python and C modules for analysis. Outputs include PDF graphics export, along with XML, GAGGLE micro-formats and tab delimited text for data interoperability. A standalone GeneWeaver package is currently unavailable because the data, tools and schema framework are highly system-dependent and intended as an Internet-based analysis framework. However, we are developing a set of APIs to allow users to operate on and contribute to our secondary data environment. Tools are documented in a variety of formats including a movie demo, tutorial exercise, quick start PDF, tool tips and interactive help. A web-based form and email links provide individualized user support.

FUTURE DIRECTIONS

The underlying framework of GeneWeaver is extensible on several fronts. Tool development is centered on providing users better approaches to managing the scope of output that the site provides. This will be accomplished through further development of simple text output that can be analyzed using other tools, and in the development of additional software for local processing including FGPU assisted processing of larger visual output. GeneWeaver may be augmented by our user community through the development of additional modules that operate on the gene-phenotype graph which can be piloted on machine readable output generated from the GeneWeaver site. New species are included through the incorporation of additional annotated genomes and homology tables, and new gene identifiers, particularly for microarray platforms, are incorporated through the use of the GEO IDs (34). GeneWeaver has a relational back-end designed to scale vastly beyond the current data and to accommodate biomolecular entities other than genes. We are working to expand the underlying data structure to accept discrete relationships between biological objects that are outside of the concept of gene. For example, ongoing development will allow users to submit RNA, SNP or methylation data in conjunction with assigned biological functions without having to generalize these data to gene symbols. Our intention is to generalize the secondary database to a solution that will incorporate most relevant current and future technologies for genome-wide function characterization.

Data sets from publications representing genome-wide studies in functional biology are continuously being added and curated. This includes the active and passive solicitation of user-submitted data. The goal is to continue to identify quality public research resource data for their potential incorporation into ODE. Rich documentation and active training of users will facilitate the best use of the analysis tools by biological researchers.

CONCLUSIONS

GeneWeaver provides a real-time, interactive and extensible environment for user-initiated integrative analysis of functional genomics data. It contains a significant repository of curated multispecies gene sets assigned to biological function and associated with community metadata. Access to GeneWeaver allows users to efficiently ask questions about the underlying functional genomics of complex biological systems by spanning secondary data covering multiple species and experimental protocols. By leveraging our large curated repository of discrete gene set relationships users can easily integrate their own functional genomics studies and apply robust analysis tools to discover new knowledge without the burden of having to collect, reanalyze and collate appropriate comparative studies.

When systematic data curation is applied in conjunction with GeneWeaver's analysis tools, a consistent framework can be imposed on the perceived disorder of diverse secondary data in any area of functional genomics analysis, ensuring scalability, data quality and a common vocabulary. By rendering these data computable, we retrieve valuable biological information from functional genomic studies that would otherwise be relegated to the challenges of diverse, poorly standardized, yet publicly available data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–5, Supplementary Figures 1–14.

ACKNOWLEDGEMENTS

We would like to thank Dr. Judith A. Blake for helpful comments on this manuscript and the GeneWeaver software system.

FUNDING

National Institutes of Health (NIAAA U01 AA13499, RO1 AA18776). Early development of the Ontological Discovery Environment occurred as a pilot of the

Integrative Neuroscience Initiative on Alcoholism. Funding for open access charge: NIH RO1 AA18776.

Conflict of interest statement. None declared.

REFERENCES

1. Guo,A.Y., Webb,B.T., Miles,M.F., Zimmerman,M.P., Kendler,K.S. and Zhao,Z. (2009) ERGR: an ethanol-related gene resource. *Nucleic Acids Res.*, **37**, D840–D845.
2. Le-Niculescu,H., Patel,S.D. and Niculescu,A.B. (2010) Convergent integration of animal model and human studies of bipolar disorder (manic-depressive illness). *Curr. Opin. Pharmacol.*, **10**, 594–600.
3. Li,C.Y., Mao,X. and Wei,L. (2008) Genes and (common) pathways underlying drug addiction. *PLoS Comput. Biol.*, **4**, e2.
4. Mulligan,M.K., Ponomarev,I., Hitzemann,R.J., Belknap,J.K., Tabakoff,B., Harris,R.A., Crabbe,J.C., Blednov,Y.A., Grahame,N.J., Phillips,T.J. *et al.* (2006) Toward understanding the genetics of alcohol drinking through transcriptome meta-analysis. *Proc. Natl Acad. Sci. USA*, **103**, 6368–6373.
5. Nissenbaum,J., Devor,M., Seltzer,Z., Gebauer,M., Michaelis,M., Tal,M., Dorfman,R., Abitbul-Yarkoni,M., Lu,Y., Elahipanah,T. *et al.* (2010) Susceptibility to chronic pain following nerve injury is genetically affected by *CACNG2*. *Genome Res.*, **20**, 1180–1190.
6. Austin,C.P., Battey,J.F., Bradley,A., Bucan,M., Capecchi,M., Collins,F.S., Dove,W.F., Duyk,G., Dymecki,S., Eppig,J.T. *et al.* (2004) The knockout mouse project. *Nat. Genet.*, **36**, 921–924.
7. Li,Z., Vizeacoumar,F.J., Bahr,S., Li,J., Warringer,J., Vizeacoumar,F.S., Min,R., Vandersluis,B., Bellay,J., Devit,M. *et al.* (2011) Systematic exploration of essential yeast gene function with temperature-sensitive mutants. *Nat. Biotechnol.*, **29**, 361–367.
8. Mnaimneh,S., Davierwala,A.P., Haynes,J., Moffat,J., Peng,W.T., Zhang,W., Yang,X., Pootoolal,J., Chua,G., Lopez,A. *et al.* (2004) Exploration of essential gene functions via titratable promoter alleles. *Cell*, **118**, 31–44.
9. Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E. and Eppig,J.T. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.*, **39**, D842– D848.
10. Smith,C.L. and Eppig,J.T. (2009) The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Interdiscip. Rev. Syst. Biol. Med.*, **1**, 390–399.
11. The Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
12. Guan,Y., Ackert-Bicknell,C.L., Kell,B., Troyanskaya,O.G. and Hibbs,M.A. (2010) Functional genomics complements quantitative genetics in identifying disease-gene associations. *PLoS Comput. Biol.*, **6**, e1000991.
13. Neely,G.G., Hess,A., Costigan,M., Keene,A.C., Goulas,S., Langeslag,M., Griffin,R.S., Belfer,I., Dai,F., Smith,S.B. *et al.* (2010) A genome-wide Drosophila screen for heat nociception identifies *alpha2delta3* as an evolutionarily conserved pain gene. *Cell*, **143**, 628–638.
14. McGary,K.L., Park,T.J., Woods,J.O., Cha,H.J., Wallingford,J.B. and Marcotte,E.M. (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl Acad. Sci. USA*, **107**, 6544–6549.
15. 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.
16. Chesler,E.J. and Baker,E.J. (2010) The importance of open-source integrative genomics to drug discovery. *Curr. Opin. Drug Discov. Dev.*, **13**, 310–316.
17. Gardner,D., Akil,H., Ascoli,G.A., Bowden,D.M., Bug,W., Donohue,D.E., Goldberg,D.H., Grafstein,B., Grethe,J.S., Gupta,A. *et al.* (2008) The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics*, **6**, 149–160.
18. Wang,J., Williams,R.W. and Manly,K.F. (2003) WebQTL: web-based complex trait analysis. *Neuroinformatics*, **1**, 299–308.
19. Davis,A.P., King,B.L., Mockus,S., Murphy,C.G., Saraceni-Richards,C., Rosenstein,M., Wiegers,T. and Mattingly,C.J. (2011) The Comparative Toxicogenomics Database: update 2011. *Nucleic Acids Res.*, **39**, D1067–D1072.
20. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, P3.
21. Jonquet,C., Shah,N.H. and Musen,M.A. (2009) The open biomedical annotator. *Summit on Translat. Bioinformat*. AMIA, San Francisco, pp. 56–60.
22. Osborne,J.D., Flatow,J., Holko,M., Lin,S.M., Kibbe,W.A., Zhu,L.J., Danila,M.I., Feng,G. and Chisholm,R.L. (2009) Annotating the human genome with disease ontology. *BMC Genomics*, **10(Suppl. 1)**, S6.
23. Shannon,P.T., Reiss,D.J., Bonneau,R. and Baliga,N.S. (2006) The Gaggle: an open-source software system for integrating bioinformatics software and data sources. *BMC Bioinformatics*, **7**, 176.
24. Zhang,Y., Chesler,E.J. and Langston,M.A. (2007) On finding Bicliques in Bipartite graphs: a novel algorithm with application to the integration of diverse biological data types. *Hawaii International Conference on System Sciences*, Waikoloa, HI.
25. Meehan,T.F., Carr,C.J., Jay,J.J., Bult,C.J., Chesler,E.J. and Blake,J.A. (2011) Autism candidate genes via mouse phenomics. *J. Biomed. Inform.*, March 11 (doi:10.1016/j.jbi.2011.03.003; epub ahead of print).
26. Smith,C.L., Goldsmith,C.A. and Eppig,J.T. (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.
27. Twigger,S.N., Shimoyama,M., Bromberg,S., Kwitek,A.E. and Jacob,H.J. (2007) The Rat Genome Database, update 2007—Easing the path from disease to data and back again. *Nucleic Acids Res.*, **35**, D658–D662.
28. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: A resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.
29. Cherry,J.M., Adler,C., Ball,C., Chervitz,S.A., Dwight,S.S., Hester,E.T., Jia,Y., Juvik,G., Roe,T., Schroeder,M. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.
30. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al.* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.*, **37**, D555–D559.
31. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chan,J., Chen,W.J., De La Cruz,N., Davis,P., Duesbury,M., Fang,R. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.
32. Sprague,J., Bayraktaroglu,L., Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Haendel,M., Howe,D.G., Knight,J. *et al.* (2008) The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.*, **36**, D768–D772.
33. Mattingly,C.J., Colby,G.T., Forrest,J.N. and Boyer,J.L. (2003) The Comparative Toxicogenomics Database (CTD). *Environ. Health Perspect.*, **111**, 793–795.
34. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. and Edgar,R. (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.