

# Finding cross-species phenomic similarity through integration of heterogeneous functional genomic data

Elissa J. Chesler, PhD

Associate Professor

The Jackson Laboratory

Supported by  
AA18776 jointly  
funded by NIDA and  
NIAAA.



# Addressing the challenge of diversity in models for complex disease

- Genetic polymorphisms can cause multiple diseases (pleiotropy).
- Named diseases may be caused by diverse mechanisms (heterogeneity).
- Nosology defined by external manifestations of disease may poorly align with the underlying biology.
- Face validity of animal models does not always indicate underlying biological construct validity
- Any effort to align model organisms to disease must simultaneously consider both the disease and model biology.



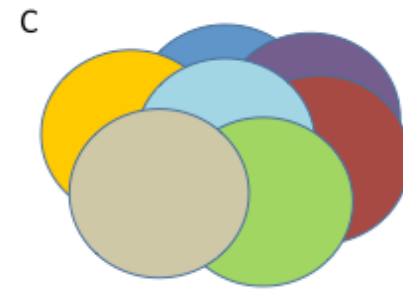
# Data-driven classification of traits and models based on underlying biology



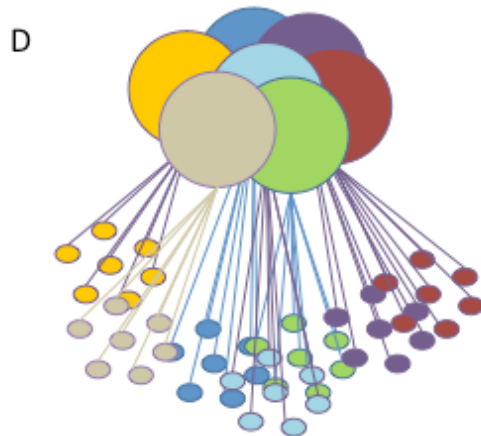
Abnormal behavior



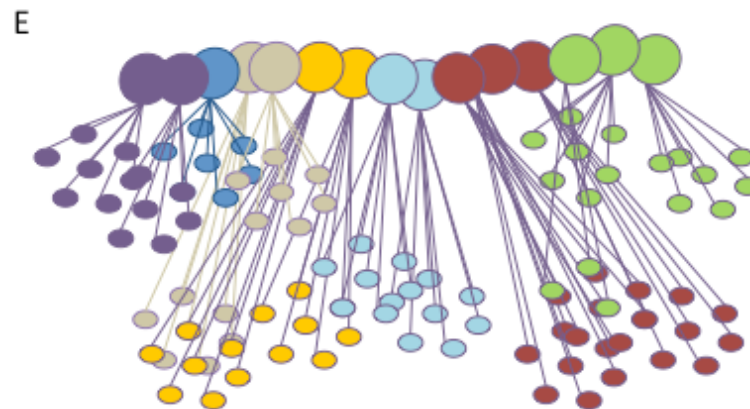
Psychoanalytic concepts



Formal diagnostic categories



Biological characterization of disorders



Refinement of diagnostic categories through biological investigation

# Modeling behavior in the laboratory mouse

“A mouse staring pensively into a flask reflecting on the direction his life has taken.”



# Toward alignment of disease and model through objective phenotypes

“So, how does that make you feel?”

Rodent assays based on Face Validity and Pharmacological Validity

ARRIVE guidelines document experimental conditions to ensure reproducibility

In psychiatry, objective Research Diagnostic Criteria (Rdocs) are being developed



# Phenotype Ontologies

- Balance competing priorities

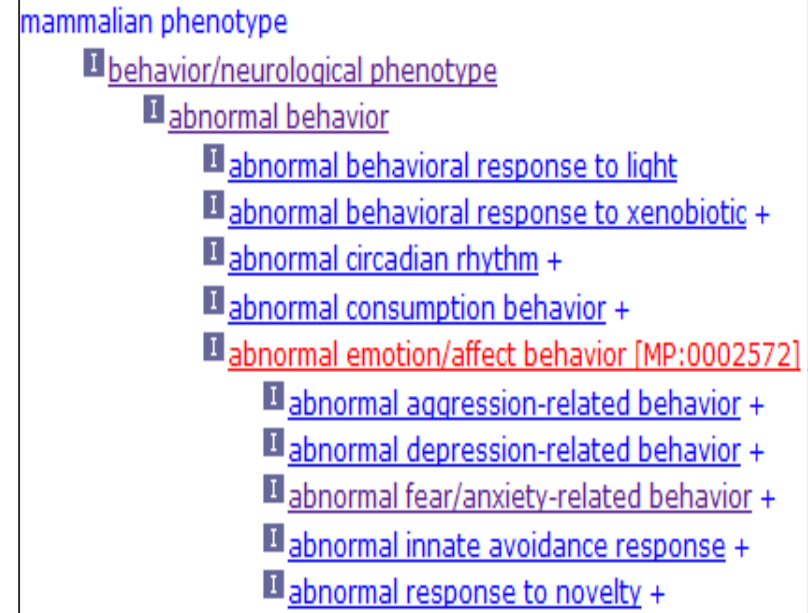
- avoid anthropomorphism
- allow cross species alignment
- retain objectivity
- retain behavioral meaning

- Enable harmonization

- Assays
- Contexts
- Interpretations

- Several approaches and resources

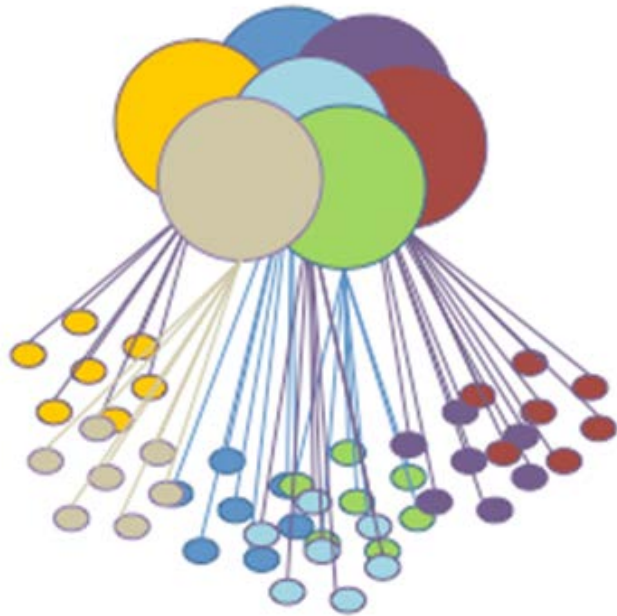
- Mammalian Phenotype, Vertebrate Trait Ontology, Neuro Behavioral Ontology, Animal Behavior Ontology



Smith CL, Eppig JT. Mamm Genome. 2012  
Park CA, et al. J Biomed Semantics. 2013  
Gkoutos GV Int Rev Neurobiol. 2012  
Midford PE. Bioinformatics. 2004



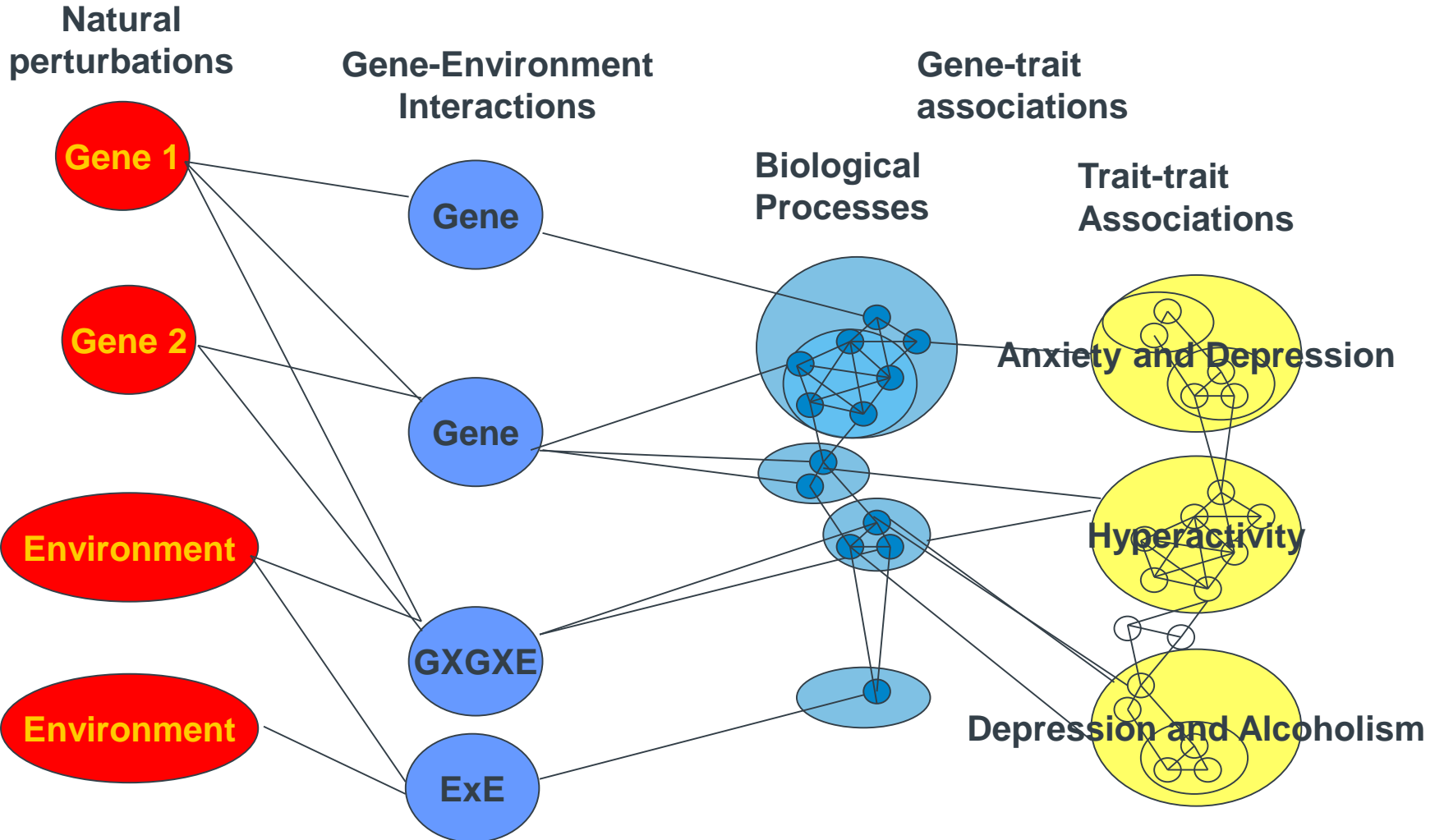
# Many mouse genetic strategies associate genes to traits and phenotype terms



Biological characterization of disorders


- Mutant characterization, e.g. IMPC screen of knockout mice ([mousephenotypes.org](http://mousephenotypes.org); MGI Phenotypic Alleles)
- Genetic loci containing variants that influence phenotype (QTLs from MGI)
- Differential Expression (GEO, publication gene lists)

# Systems genetic analysis holistically connects traits to sets of genes and variants





# Systems Genetics and the 'dark web': gene-trait associations are available via web services



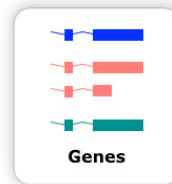
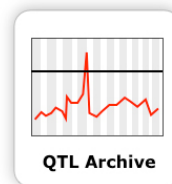
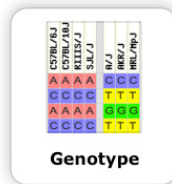
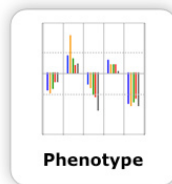
Mouse Phenome Database  
at The Jackson Laboratory

Search:

Enter a keyword, investigator, strain, gene, ontology term

Welcome data previewer [JaxKOMPheno1 home](#) [Logout](#)

- About MPD
- Approaches
- What's new
- Contributing data
- Investigators
- Larger initiatives
- Publications
- Pheno tools demo
- Tutorial videos ▶
- Your collection 🗂
- Download data
- Also at JAX
- Suggestion box
- Help desk




Chesler EJ, et al Nat Neurosci. 2004  
Wang J, et al Neuroinformatics. 2003.

Grubb SC, et al Nucleic Acids Res. 2014

## GeneNetwork

University of Tennessee: [www.genenetwork.org](http://www.genenetwork.org)

[Use GeneNetwork 2](#)



Home | Search | Help | News | References | Policies | Links
Welcome! [Login](#)

### Select and Search

Species:

Group:  [Info](#)

Type:

Data Set:  [Info](#)

Databases marked with \*\* suffix are not public yet. Access requires [user login](#).

Get Any:

Enter terms, genes, ID numbers in the **Get Any** field.  
Use \* or ? wildcards (Cyp\*a?, synap\*<sup>o</sup>).  
Use **Combined** for terms such as tyrosine kinase.

Combined:

[Search](#)   [Make Default](#)   [Advanced Search](#)

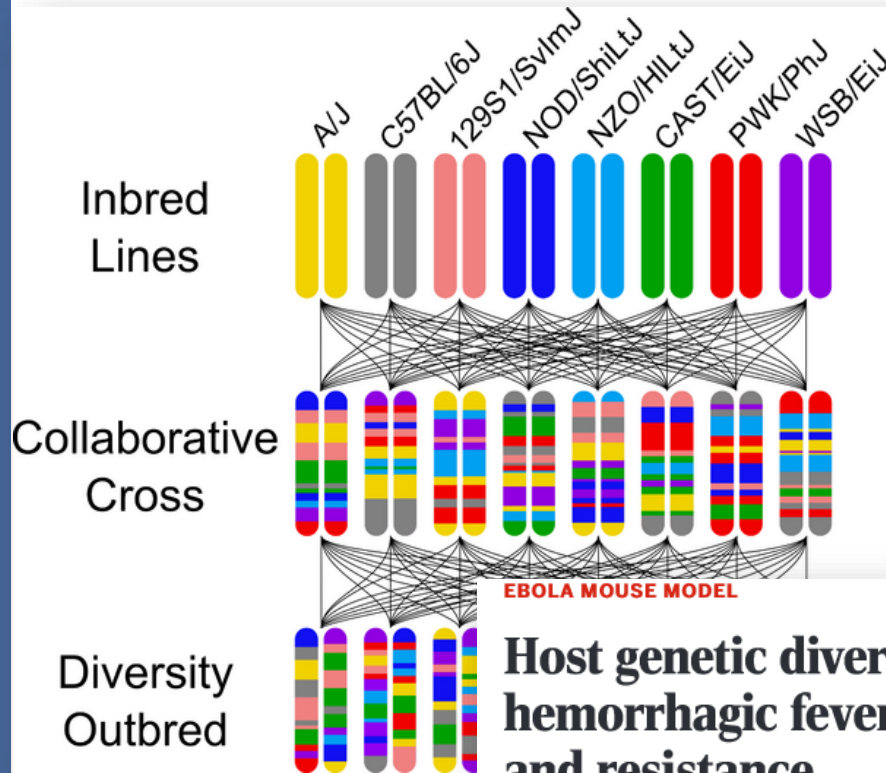
### Websites Affiliated with GeneNetwork

- UTHSC Genome Browser [Classic](#) and [Newest](#)
- UTHSC [Galaxy](#) Service
- UTHSC [Bayesian Network Web Server](#)
- GeneNetwork Classic on [Amazon Cloud](#)
- GeneNetwork Classic Code on [GitHub](#)
- GeneNetwork 2.0 Development Code on [GitHub](#)
- [GeneNetwork 2.0](#) Development

### Getting Started

1. Select **Species** (or select All)
2. Select **Group** (a specific sample)
3. Select **Type** of data:
  - Phenotype (traits)
  - Genotype (markers)
  - Expression (mRNAs)
4. Select a **Database**
5. Enter search terms in the **Get Any** or **Combined** field: words, genes, ID numbers, probes, advanced search commands
6. Click on the **Search** button

# Identifying extremes from advanced mouse populations as disease models



Promising for qualitative traits

Very challenging for complex traits

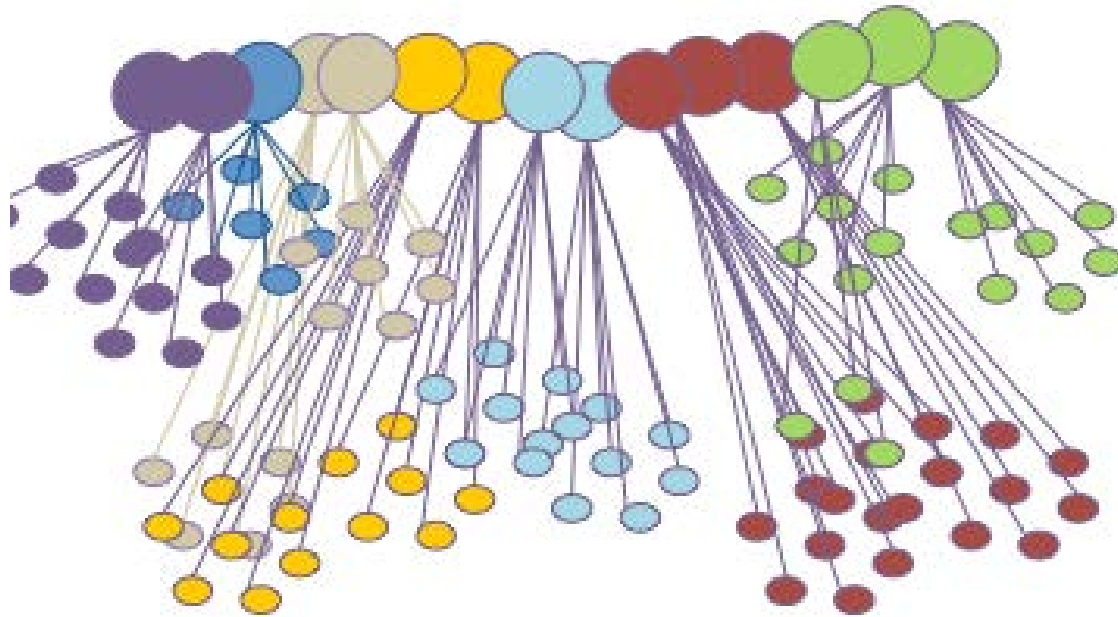
Extremes are defined statistically  
Based on study population

## Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance

Angela L. Rasmussen,<sup>\*1</sup> Atsushi Okumura,<sup>\*1,4</sup> Martin T. Ferris,<sup>2</sup> Richard Green,<sup>1</sup> Friederike Feldmann,<sup>2</sup> Sara M. Kelly,<sup>1</sup> Dana P. Scott,<sup>3</sup> David Safronetz,<sup>4</sup> Elaine Haddock,<sup>4</sup> Rachel LaCasse,<sup>2</sup> Matthew J. Thomas,<sup>1</sup> Pavel Sova,<sup>1</sup> Victoria S. Carter,<sup>1</sup> Jeffrey M. Weiss,<sup>1</sup> Darla R. Miller,<sup>2</sup> Ginger D. Shaw,<sup>2</sup> Marcus J. Korth,<sup>1</sup> Mark T. Heise,<sup>2,5</sup> Ralph S. Baric,<sup>5</sup> Fernando Pardo-Manuel de Villena,<sup>2</sup> Heinz Feldmann,<sup>4</sup> Michael G. Katze<sup>1,6†</sup>



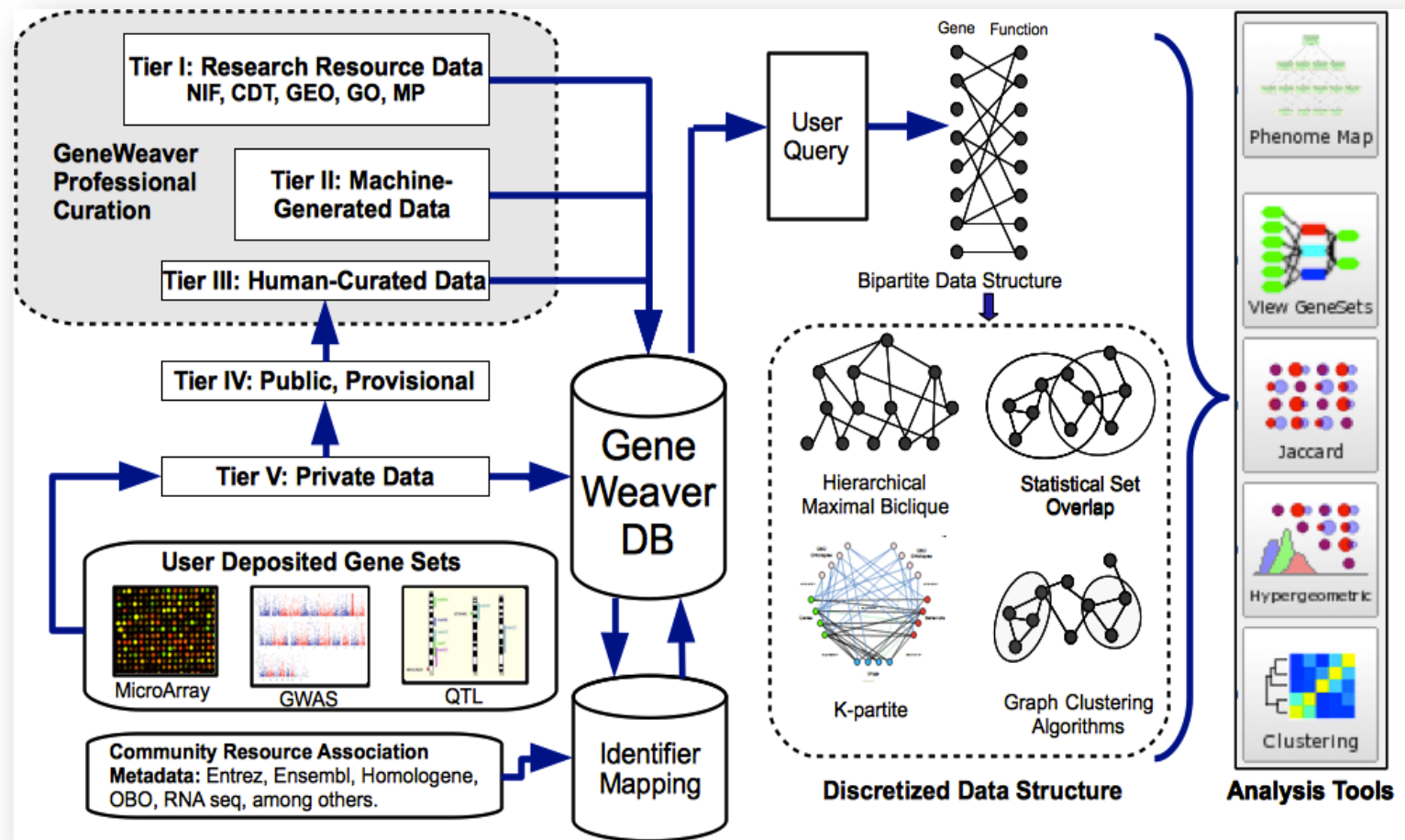
# These data resources enable an alternative data-driven classification of traits and models based on underlying biology



;

Refinement of diagnostic categories through biological investigation

# Cross-species and cross-population integration in GeneWeaver



Baker EJ, Jay JJ, Bubier JA, Langston MA, Chesler EJ. Nucleic Acids Res. 2012



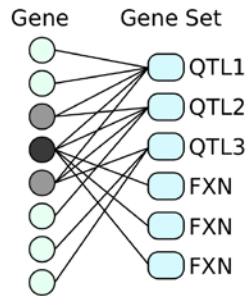
# Research questions for integrative functional genomics

- Which assays and conditions provide annotations that most resemble disease features and patient biomarkers?
- In diverse assays of the same underlying disease construct, what genes and gene products are consistently observed?
- Which animal models map onto the human disease based on genomic associations?

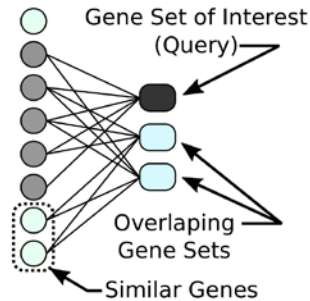


# Statistical and graph theoretical methods for integrative functional genomics in GeneWeaver

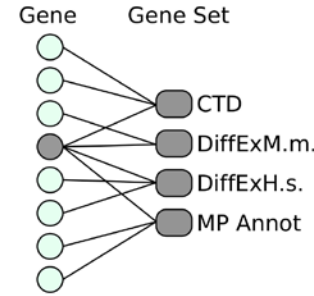
1. Refine Overlapping QTL



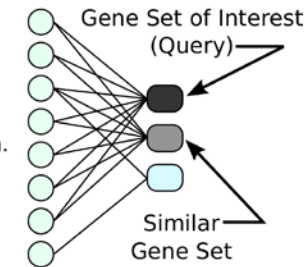
2. Find Highly Connected Genes



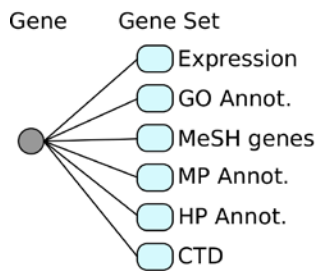
3. Find High Degree Genes



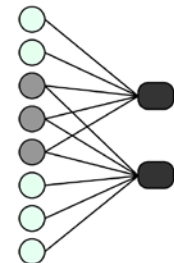
4. Find Similar Gene Sets



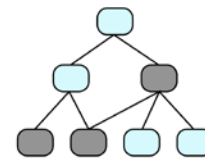
5. Search By Gene



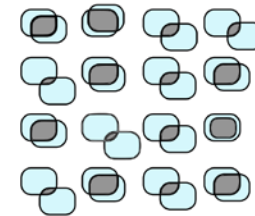
6. Enumerate Intersections



7. Hierarchical Gene Set Similarity



8. Pairwise Gene Set Intersection



Bubier et al, Mammalian Genome, 2015



# Identification of a new mouse model for alcohol preference



**GeneWeaver.org**

A system for the integration of functional genomics experiments.

Welcome Guest! To ensure future access to your data, please Register — or Login

Home Search Manage GeneSets Analyze GeneSets About Help

## GeneSets Similar to GS128735

Help | Feedback

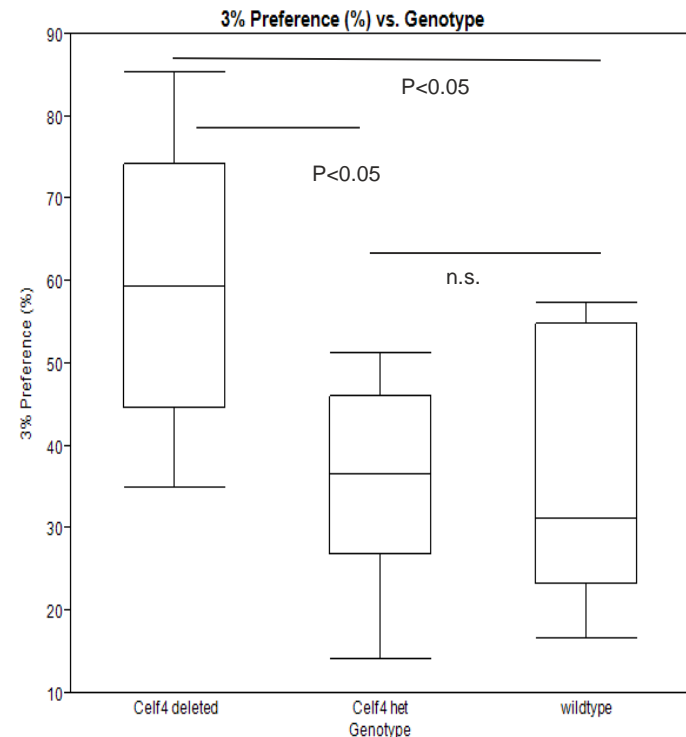
### Gene Set #128735 - Alcoholism MeSH associations in PubMed

**Description:** Genes associated to MeSH term 'Alcoholism' or a descendant in PubMed's curated annotations. METHOD: NCBI's gene2pubmed and e-utilities were used to associate MeSH PubMed annotations to Entrez Gene IDs. To control for outlier associations, curated annotations must have at least 2 occurrences to be retained. Retained annotations were then pushed to ancestor terms to ensure a complete tree. All data fetched 28 Feb 2012.

[« Back to GeneSet details](#)

#### Similar GeneSets:

	Select All	Add Selected to Project...	Expand All	Gene Similarity (Jaccard)
<input type="checkbox"/> <b>Tier III</b> <b>Human</b> <b>14 Genes</b> GS216653: Protein Biomarkers of Alcohol Abuse				0.044444
<input type="checkbox"/> <b>Tier V</b> <b>Mouse</b> <b>151 Genes</b> GS137124: 142 enriched CELF4				0.041353
<input type="checkbox"/> <b>Tier I</b> <b>Human</b> <b>GO</b> <b>4 Genes</b> GS199997: GO:0004024 alcohol dehydrogenase activity, zinc-dependent				0.031496
<input type="checkbox"/> <b>Rat</b> <b>277 Genes</b> GS216533: Learning related neural genes in dentate affected by age in Rat				0.030612
<input type="checkbox"/> <b>Tier III</b> <b>Mouse</b> <b>70 Genes</b> GS216898: INIA Ethanol				0.020725
<input type="checkbox"/> <b>Rat</b> <b>70 Genes</b> GS216438: Expression Dynamics in the Amygdala Central Nucleus During Alcohol Withdrawal				0.020725
<input type="checkbox"/> <b>Rat</b> <b>28 Genes</b> GS216537: Differential Expression in the Amygdala of NIH-HS "low-anxious" relative to NIH-HS "high-anxious" rat.				0.019737



# Identifying promising new models by characterizing the 'ignorome'

The screenshot shows the GeneWeaver.org interface. At the top left is the GeneWeaver.org logo with the tagline "A system for the integration of functional genomics experiments." and a navigation menu with "Home", "Search", "Manage GeneSets", "Analyze GeneSets", "About", and "Help". On the top right, it says "Welcome Guest! To enter to your data, please log in".

The main section is titled "Search for GeneSets". It has tabs for "General", "Tiers", "Species", and "Attributions". A search bar contains the ID "2900011O08Rik", which is circled in red. Below the search bar, there are checkboxes for "GeneSets", "Genes", "Abstracts", and "Ontologies", and a "Search" button. To the right of the search bar is an "Add Selected to Project" button.

Below the search bar, there is a section for "Global Filters" with options to include "provisional" (0) and "deprecated" (13) items, and a "Group" dropdown set to "Any (77)". A slider for "Geneset Size" is set between 23 and 29709. There is also a "Tiers" section with checkboxes for "No Tier (2)", "I: Resources (52)", "II: Pro-Curated (7)", "III: Curated (3)", "IV: Provisional (0)", and "V: Private (0)". A "Species" section is also visible.

The search results are displayed in a table with the following columns: "Select All Results", "Tier II", "Mouse", "Number of Genes", and "GeneSet Description". The results are sorted by "Relevance".

Select All Results	Tier II	Mouse	Number of Genes	GeneSet Description
<input type="checkbox"/>	Tier II	Mouse	149 Genes	GS33885: Whole Brain Gene expression correlates of Distance traveled (cm) during the first five minutes after ethanol in Males BXD
<input type="checkbox"/>	Tier II	Mouse	49 Genes	GS34289: Whole Brain Gene expression correlates of Activity in 30 second interval post 3rd tone shock pairing in Females BXD
<input type="checkbox"/>	Tier II	Mouse	23 Genes	GS35389: Whole Brain Gene expression correlates of Time below threshold in Females & Males BXD
<input type="checkbox"/>	Tier II	Mouse	101 Genes	GS36362: Whole Brain Gene expression correlates of Open Field - Total rears 0-5 minutes in Females & Males BXD
<input type="checkbox"/>	Tier II	Mouse	219 Genes	GS36477: Whole Brain Gene expression correlates of Morphine - Severity of ptosis in Males BXD
<input type="checkbox"/>	Tier II	Mouse	437 Genes	GS84292: nicotine sensitivity (Published QTL, Chr 16)
<input type="checkbox"/>	Tier II	Mouse	556 Genes	GS84293: METH responses for home cage activity (Published QTL, Chr 16)

Bubier et al, Mammalian Genome, 2015





# 2900011O08Rik mutation is available from KOMP repository – A model exists

## International Mouse Strain Resource (IMSR)



Search Repositories Participate Glossary Contact Us About Us Deposit Strains

### Summary

Search for:

2900011O08Rik

Search

Reset

Show Options

### You searched for:

Query: 2900011O08Rik

75 strains(s) match your unfiltered search.

Export: Filter by: State Type Provider Mutation

N	Strain Name	Synonyms	States	Repository
?	C57BL/6N-2900011O08Rik-GH(IST12125G1)Tigm		ES Cell	TIGM
?	C57BL/6N-2900011O08Rik-GH(IST11876F7)Tigm		ES Cell	TIGM



SEARCH

ABOUT IMPC

NEWS & EVENTS

CONTACT

MY IMPC

Login Register

Home » Search

Filter your search

- Genes 1
- IMPC Phenotyping Status
  - Approved 0
  - Started 0
  - Attempt Registered 0
  - Legacy 0
- IMPC Mouse Production Status
- IMPC Mouse Production Center
- IMPC Mouse Phenotype Center
- Subtype

"2900011O08Rik"

[View example search](#)

Found 1 gene

Show 10 entries

Download

Gene

Production Status Phenotype Status

**2900011O08Rik**

name: RIKEN cDNA 2900011O08 gene  
human ortholog: C16orf45  
synonym: MINP

Mice

Interest

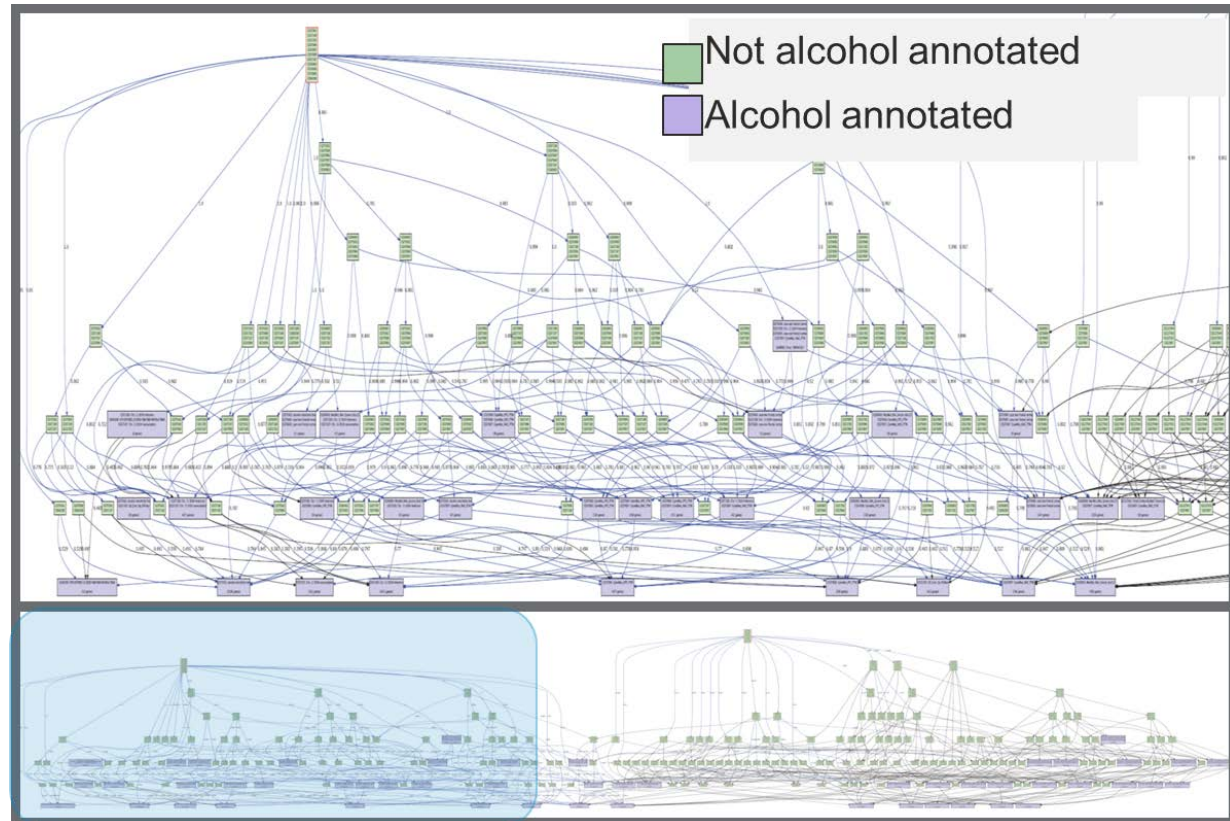
Showing 1 to 1 of 1 entries

← First 1 Last →



# Aggregate analysis of many studies of alcohol preference

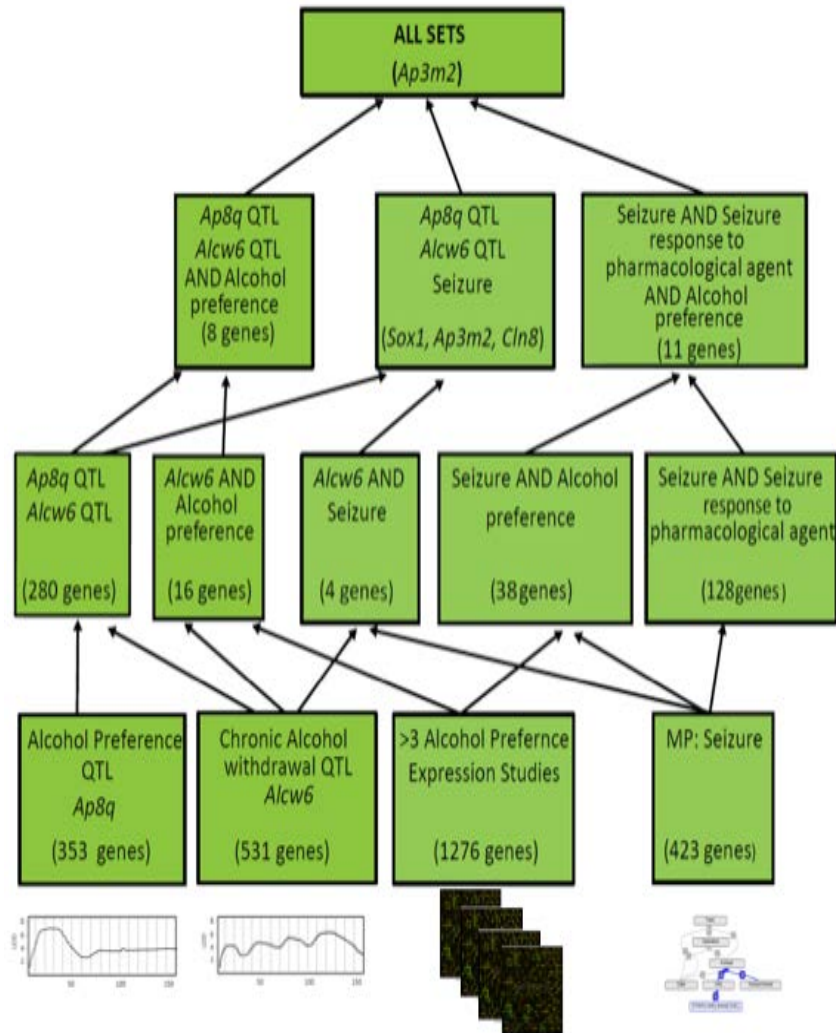
↑  
connectivity



The most frequently represented genomic results in alcohol preference studies are not currently associated with alcoholism.



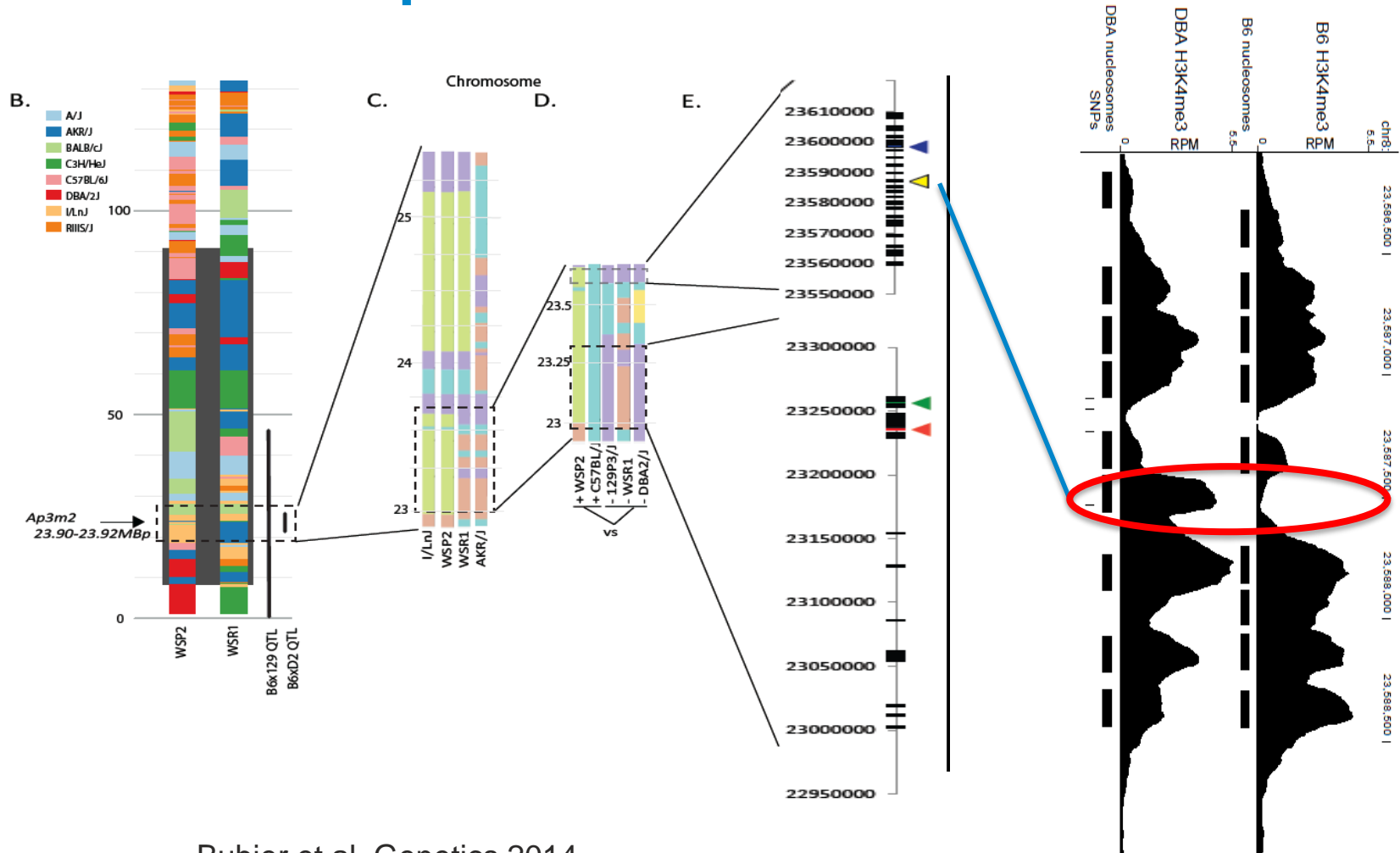
# Finding models for related facets of alcohol use disorder.



Bubier et al, Genetics 2014



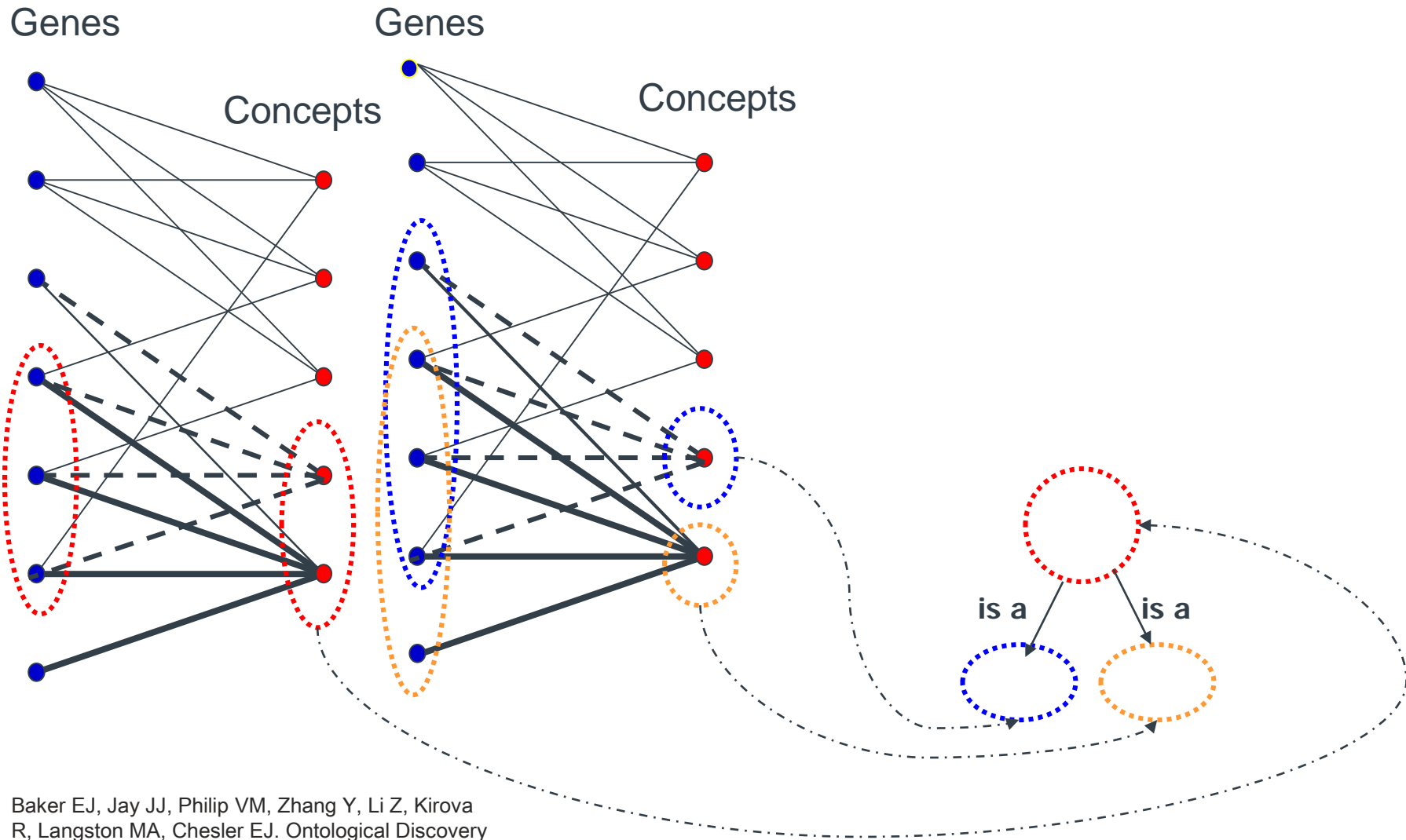
# Convergent evidence across populations and traits enables causal SNP identification and design of precision mouse models



Bubier et al, Genetics 2014



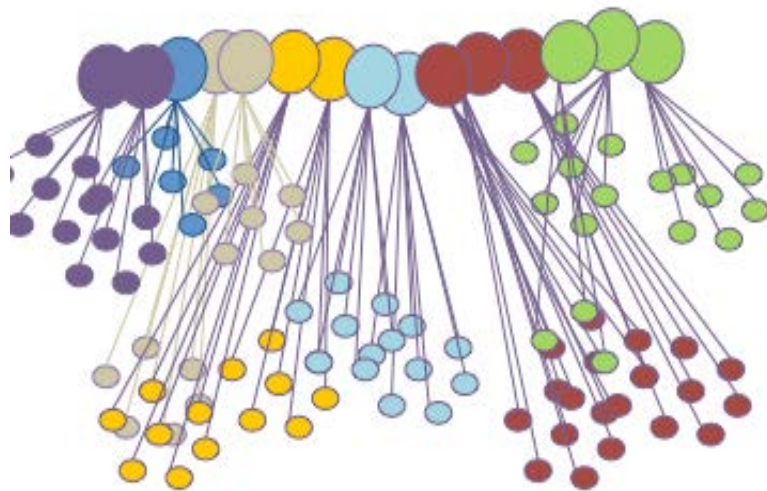
# Construction of a latent ontology from empirical genomic evidence



Baker EJ, Jay JJ, Philip VM, Zhang Y, Li Z, Kirova R, Langston MA, Chesler EJ. Ontological Discovery Environment: a system for integrating gene-phenotype associations. Genomics. 2009

# Data driven classification of psychiatric Disorders Using MeSH to Gene Annotations

Precision models

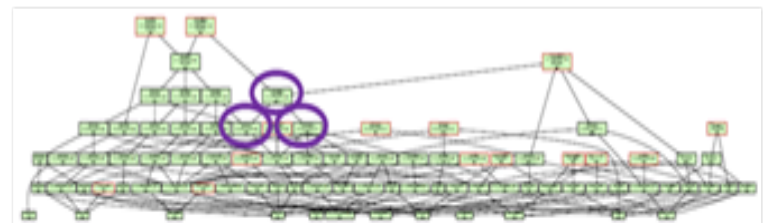


Refinement of diagnostic categories

GS128772: Autistic disorder  
 GS128327: Depressive disorder/alcohol use comorbidity  
 GS128721: Mood disorders  
 GS128731: Anxiety disorders  
 GS128747: Schizophrenia and disorders with psychotic features  
**(GRIN2A, HTR1B)**

GS128772: Autistic disorder  
 GS128721: Mood disorders  
**GS128731: Anxiety disorders**  
 GS128747: Schizophrenia and disorders with psychotic features  
**(20 Genes)**

GS128772: Autistic disorder  
**GS128327: Depressive disorder/alcohol use comorbidity**  
 GS128721: Mood disorders  
 GS128747: Schizophrenia and disorders with psychotic features  
**(ABCB1, GRIN2A, HTR1B, OXTR)**



# Summary

- Linking animal models to human disease through phenotypes often exploits face validity.
- ‘Construct validity’ is the desired characteristic.
- Underlying construct similarity can be obtained through genome wide comparison of assays and models.
- A wealth of data sources from mouse and other organisms exist.
- Cross-species integrative functional genomics enables global comparison of animal models, assays and diseases based on underlying biology.



# Acknowledgements

## Collaborative Cross

Dr. Vivek Philip  
Jason S. Spence  
Melissa M. Beckman  
Dr. Roumyana Kirovva  
Dr. Cymbeline Culiat  
Darla Miller  
Dr. Brynn H. Voy  
Dr. William R. Lariviere  
Dr. Bruce O'Hara  
Dr. Gary Churchill  
Dr. David Threadgill  
Dr. Robert Williams

Systems Genetics Group at  
ORNL

The Ellison Medical  
Foundation

Department of Energy  
Office of Science

## Diversity Outbred

Dr. Ryan W. Logan  
Dr. Ray F. Robledo  
Dr. Jill Recla  
Dr. Dan Gatti  
Dr. Narayanan Raghupathy  
Dr. Matthew A. Hibbs  
Dr. Carol Bult  
Dr. Andrew Holmes  
Dr. Gary A. Churchill  
Kathryn Mc Naughton  
Dr. Price Dickson  
Andrew Gallup

Center for Genome Dynamics,  
Nathan Shock Center on Aging

**R01 DA 037927**

## The Jackson Laboratory

**P50 GM 76468**  
P30 AG38070

## GeneWeaver

Dr. Jeremy Jay  
Dr. Jason A. Bubier  
Dr. Erich Baker  
Dr. Michael A. Langston  
Dr. Judith Blake  
Dr. Yun Zhang  
Dr. John C. Crabbe  
Pamela Metten

Integrative Neuroscience  
Initiative On Alcoholism

U24 AA13513  
U01 AA13499  
**R01 AA18776**

## Mouse Phenome Database

Dr. Molly Bogue  
Stephen Grubb  
**DA028420**

## KOMP

Troy Wilcox  
James Clark  
Laura Anderson  
James DeNegre  
Dr. Patsy Nishina  
Chuck Donnelly  
Dr. Karen Svenson  
Dr. Bob Braun  
Dr. Vivek Kumar

International Mouse  
Phenotyping Consortium

**U54 HG006332**





# The quest for consilience



*affectionibus quibus  
Whewell*

“...the evidence in favour of our induction is of a much higher and more forcible character when it enables us to explain and determine cases of a kind different from those which were contemplated in the formation of our hypothesis...

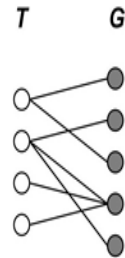
***No accident could give rise to such extraordinary coincidence.”***

-W. Whewell, 1847

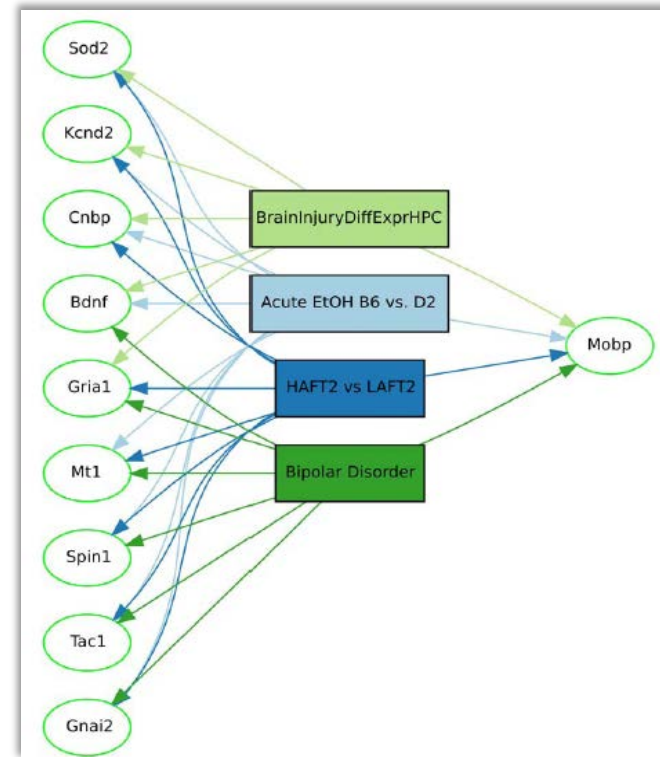


# Representing the data for integration

- Relationships are discrete. The corresponding adjacency matrix,  $M$ , may be weighted or unweighted.
- Gene lists are represented as a *bi-partite graph*,  $B = \langle T, G, E \rangle$ ,



- Genes (list members) connected to phenotypes (set names) by *edges*. A set of genes is defined by a term, and denotes those genes adjacent to the term. That is, for  $t$  in  $T$ ,  $S_t$  denotes the set of  $t$ 's neighbors in  $G$ .
- A set of sets of genes is denoted by  $S'$ . A set of sets of sets of genes is defined similarly, and denoted as  $S''$ .
- Most existing GeneWeaver functions operate on  $B$ .



Baker EJ, Jay JJ, Philip VM, Zhang Y, Li Z, Kirova R, Langston MA, Chesler EJ. Ontological Discovery Environment: a system for integrating gene-phenotype associations. *Genomics*. 2009 Dec;94(6):377-87.

