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

A
Abnormal behavior

B
Psychoanalytic concepts

C
Formal diagnostic categories

D
Biological characterization of disorders

E
Refinement of diagnostic categories through biological investigation

Logan and Chesler, Int Rev Neurobiology, 2012
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
Midford PE. Bioinformatics. 2004
Many mouse genetic strategies associate genes to traits and phenotype terms

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

Biological characterization of disorders
Systems genetic analysis holistically connects traits to sets of genes and variants

Gene 1

Gene 2

Environment

Environment

Gene-Environment Interactions

Biological Processes

Gene-trait associations

Trait-trait Associations

Anxiety and Depression

Hyperactivity

Depression and Alcoholism

Chesler EJ and Langston, MA RECOMB Systems Biology and Regulatory Genomics 2005: 150-165
Systems Genetics and the ‘dark web’: gene-trait associations are available via web services


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
These data resources enable an alternative data-driven classification of traits and models based on underlying biology.
Cross-species and cross-population integration in GeneWeaver

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

Bubier et al, Mammalian Genome, 2015
Identifying promising new models by characterizing the ‘ignorome’

Bubier et al, Mammalian Genome, 2015
2900011008Rik mutation is available from KOMP repository – A model exists.
Aggregate analysis of many studies of alcohol preference

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

Data driven classification of psychiatric Disorders Using MeSH to Gene Annotations

Precision models

Refinement of diagnostic categories
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. Wiliam R. Lariviere
Dr. Bruce O’Hara
Dr. Gary Churchill
Dr. David Threadgill
Dr. Robert Williams

**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

**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

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

**Integrative Neuroscience Initiative On Alcoholism**
U24 AA13513
U01 AA13499
R01 AA18776

**International Mouse Phenotyping Consortium**
U54 HG006332

**Systems Genetics Group at ORNL**
Center for Genome Dynamics,
Nathan Shock Center on Aging
R01 DA 037927

**The Ellison Medical Foundation**
The Jackson Laboratory
P50 GM 76468
P30 AG38070

**Department of Energy**
The Jackson Laboratory
P50 GM 76468
P30 AG38070

**Office of Science**
The Ellison Medical Foundation

**Mouse Phenome Database**
Dr. Molly Bogue
Stephen Grubb
DA028420
The quest for consilience

“…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 $B$.