Computational Infrastructure for Systems Genetics Analysis
Brian Yandell, UW-Madison

high-throughput analysis of systems data enable biologists & analysts to share tools

UW-Madison: Yandell, Attie, Broman, Kendziorski
Jackson Labs: Churchill
U Groningen: Jansen, Swertz
UC-Denver: Tabakoff
LabKey: Igra
A collaborative portal (LabKey) is connected to a systems genetics portal (PhenoGen) by an iterative process. The process involves:

1. Iterating many times
2. Getting data (GEO, Sage)
3. View results (R graphics, GenomeSpace tools)
4. Running a pipeline (CLIO, XGAP, HTDAS)
analysis pipeline acts on objects (extends concept of GenePattern)
pipeline is composed of many steps

I → A → B → C

I' → A' → B → C

combine datasets

O → E → D → D'

O' → E' → D → D'

compare methods

alternative path
causal model selection choices in context of larger, unknown network

- **focal trait** → **target trait** causal
- **focal trait** ← **target trait** reactive
- **focal trait** ↔ **target trait** correlated
- **focal trait** ↔ **target trait** uncorrelated
BxH ApoE-/- chr 2: causal architecture
BxH ApoE-/- causal network for transcription factor Pscdbp

causal trait

work of Elias Chaibub Neto
collaborative portal (LabKey)

iterate many times

systems genetics portal (PhenoGen)

view results (R graphics, GenomeSpace tools)

get data (GEO, Sage)

run pipeline (CLIO, XGAP, HTDAS)

update periodically

develop analysis methods & algorithms

byandell@wisc.edu
Pipeline diagram:

- **Input**
- **Pipeline**
- **Output**

- **Settings**
- **Checks**
- **Preserve History**
- **Package**
- **Raw Code**
- **R&D**
Model/View/Controller (MVC) software architecture

- isolate domain logic from input and presentation
- permit independent development, testing, maintenance
perspectives for building a community where disease data and models are shared

Benefits of wider access to datasets and models:
1- catalyze new insights on disease & methods
2- enable deeper comparison of methods & results

Lessons Learned:
1- need quick feedback between biologists & analysts
2- involve biologists early in development
3- repeated use of pipelines leads to documented learning from experience increased rigor in methods

Challenges Ahead:
1- stitching together components as coherent system
2- ramping up to ever larger molecular datasets
www.stat.wisc.edu/~yandell/statgen
byandell@wisc.edu

• UW-Madison
  – Alan Attie
  – Christina Kendziorski
  – Karl Broman
  – Mark Keller
  – Andrew Broman
  – Aimee Broman
  – YounJeong Choi
  – Elias Chaibub Neto
  – Jee Young Moon
  – John Dawson
  – Ping Wang
  – NIH Grants DK58037, DK66369, GM74244, GM69430, EY18869

• Jackson Labs (HTDAS)
  – Gary Churchill
  – Ricardo Verdugo
  – Keith Sheppard

• UC-Denver (PhenoGen)
  – Boris Tabakoff
  – Cheryl Hornbaker
  – Laura Saba
  – Paula Hoffman

• Labkey Software
  – Mark Igra

• U Groningen (XGA)
  – Ritsert Jansen
  – Morris Swertz
  – Pjotr Pins
  – Danny Arends

• Broad Institute
  – Jill Mesirov
  – Michael Reich
Systems Genetics Analysis Platform
Brian Yandell, UW-Madison

high-throughput analysis of systems data enable biologists & analysts to share tools

UW-Madison: Attie, Broman, Kendziorski
Jackson Labs: Churchill
U Groningen: Jansen, Swertz
UC-Denver: Tabakoff
LabKey: Igra