KOMPCluster: A Pattern Recognition and 3D Visualization System for Phenotyping Projects

Eric Engelhard, Ph.D.
Director of Informatics
Mouse Biology Program
University of California, Davis

Overview

- Large, complex data sets are driving the need for new data presentations and query methods
- Unsupervised pattern recognition provides methods for quickly discovering and presenting relationships within the data
- A case study with mouse phenotype data

Querying the Phenotype Databases

- Gene centric
  - What are the phenotypes associated with this knockout?
  - Gene centric annotations are getting deeper and more numerous
- Phenotype centric
  - What are the knockouts associated with this phenotype?
  - Smaller sets of gene centric annotations

Complex Queries

- Simple queries with few genes or phenotypes can be manually combined using current tools
  - What are the knockouts associated with this phenotype exception AND this other phenotype exception?
- More complex queries requires interfaces supporting Boolean logic
  - What are the knockouts displaying X, Y, and Z phenotype exceptions in males, but not the K phenotype exception?

Access to Raw Data Increases Complexity

- Ontologies
  - Data compatibility and simplification
- Serving raw data allows far greater query flexibility
  - User defined trigger points
  - Which knockouts have blood sodium levels higher than X?
  - ...which can be included in complex queries
  - Which male knockouts have sodium levels between X and Y and the following behavioral abnormalities...?
- All of this can be handled readily standard SQL queries, but flexible, ergonomic web interfaces are a challenge to design and implement.

Complex Queries have a Large Solution Space

- Large number of observations
- Combinatorial
  - No repetition, order doesn't matter
  - Sum of n!/r!(n-r)! for each r chosen observations from n observations
- 133 observations results in ~10^{40} possible combinations

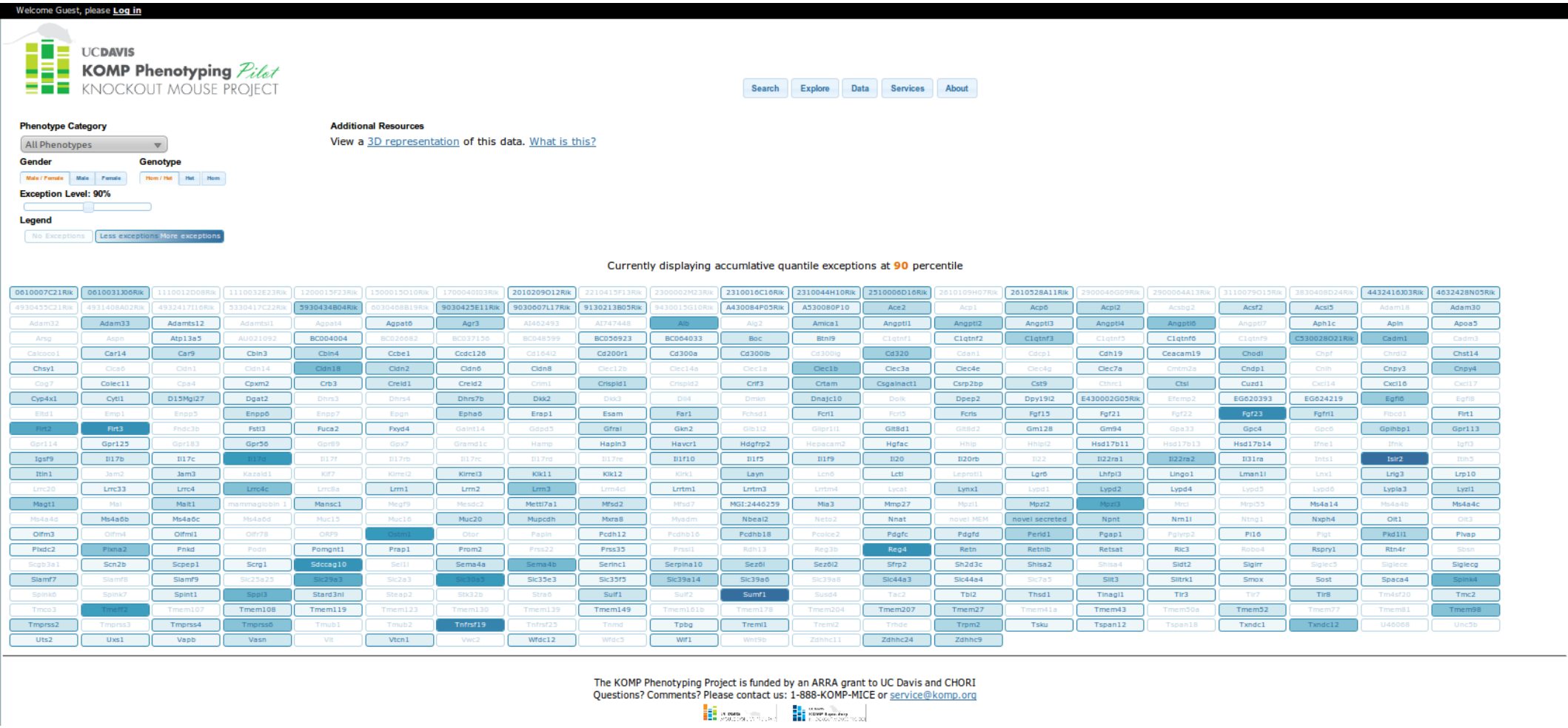
Why Use Unsupervised Pattern Recognition?

- Complex Boolean queries place a heavier knowledge burden on the user
  - Which phenotypes should be entered into a query?
  - Which COMBINATION of phenotypes should be entered into a query?
- Undefined queries allow the user to OBSERVE the data structure and answer questions like:
  - Are there any knockouts that share any four or more phenotype exceptions?
  - What are the phenotypes that group these knockouts?
  - Are there subgroups of knockouts within this group which share more phenotype exceptions?
  - Are there groups of knockouts that share some, but not all of these exceptions?
  - If there is more than one external group, which one is more closely related to the primary group?

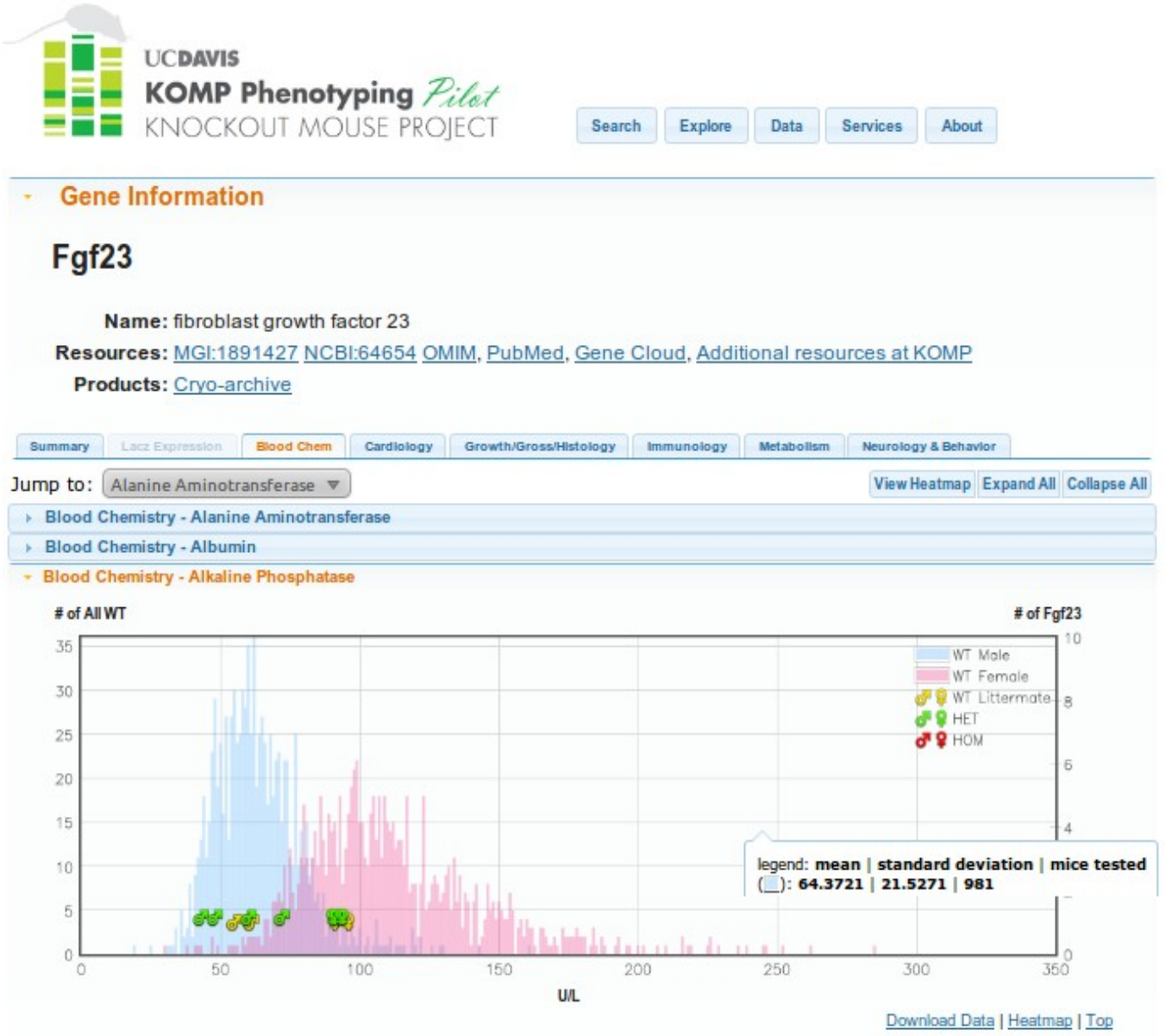
A Clustering Case Study

- Establish the relationship a group of ~500 knockout mouse lines based on all phenotype effects

KOMPphenotype.org



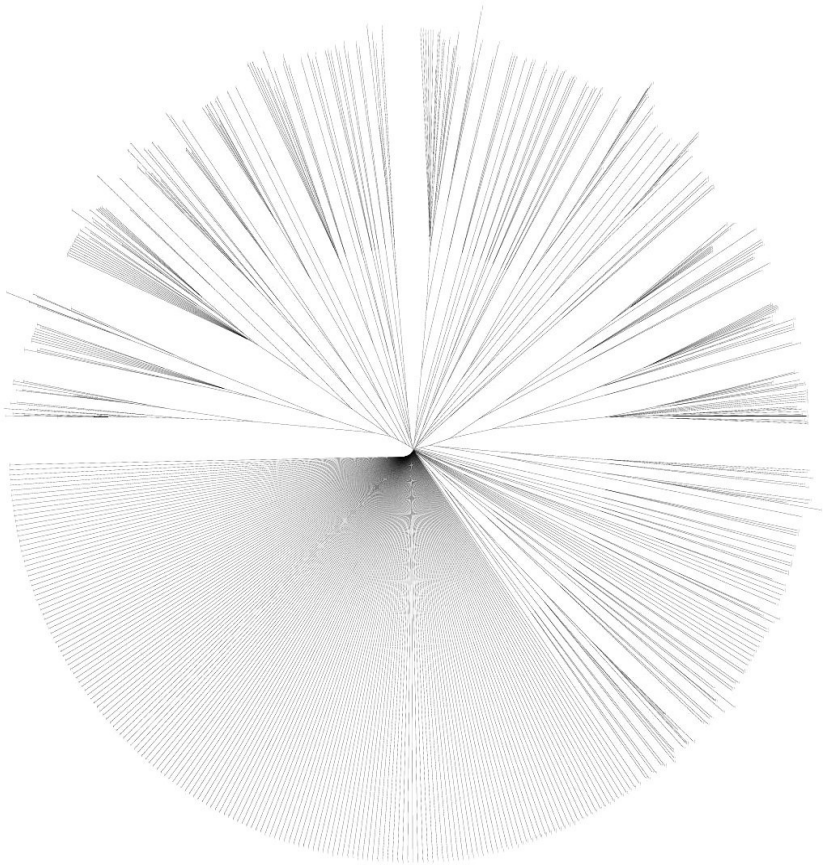
KOMPphenotype.org

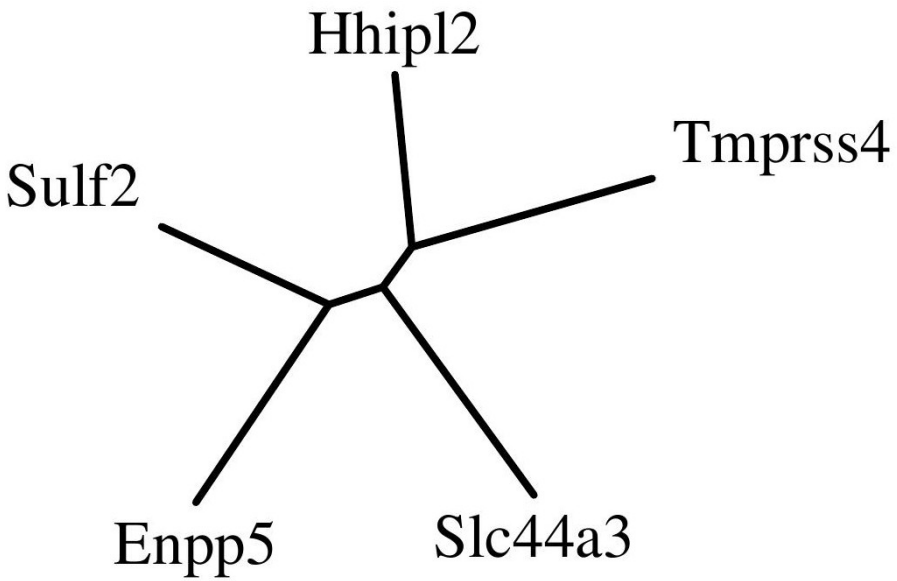


Clustering

- Set Theory
  - For a given gender, genotype, genetic background each KO line has a set of flagged phenotypes
  - The distance between any two KO lines is inversely proportional to the intersection of their sets
- Distance Matrix
  - First order relationships between all KO lines
- Neighbor Joining
  - Very fast bottom up algorithm for phenogram construction

Phenograms: Trees and Clusters



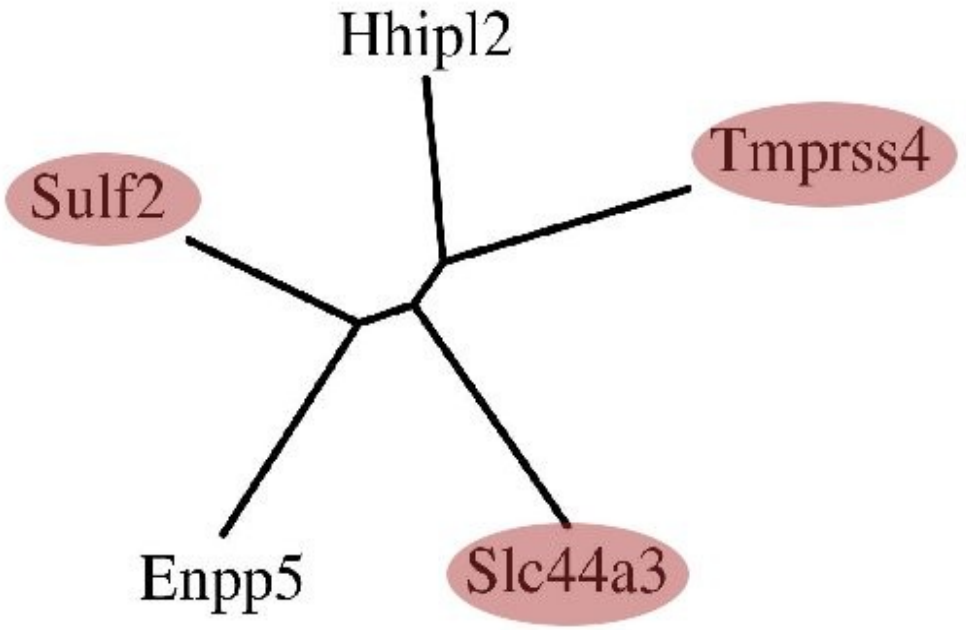


Clusters defined by specifying branch distances

Tumor Suppressors

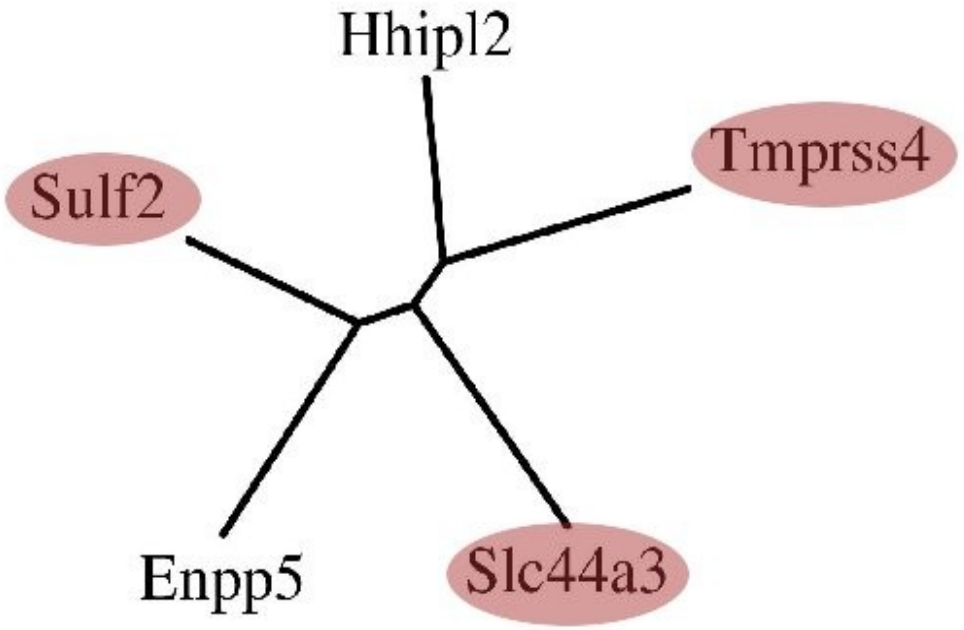
- Data from the Retrovirus Tagged Cancer Gene Database (RTCGD) mapped against the current mouse genome assembly and annotations
- Conservative algorithm flagged multiple, independent exon/intron insertions (truncated gene products)
- 849 candidates of which 16 were represented within the data

Co-Clustering



- Sulf2, Tmprss4, and Slc44a3 in same cluster
- Enpp5 may play a role in neuronal cell communication (SwissProt)
- Hhipl2 is a homolog to HHIPL2, a gene deregulated in gastric carcinomas
- p approx 5×10^{-7}

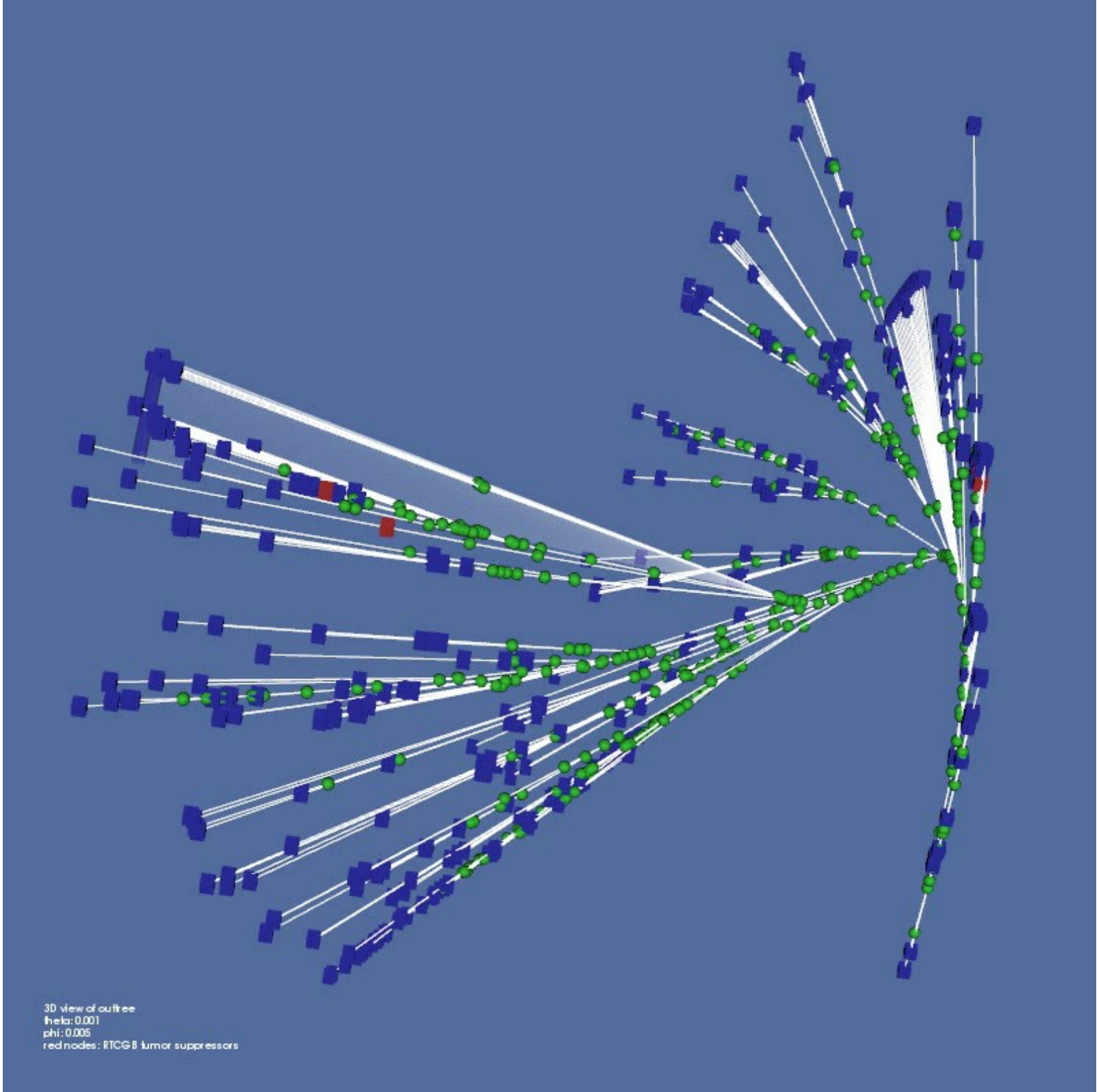
Clustered by Behavioral Phenotypes



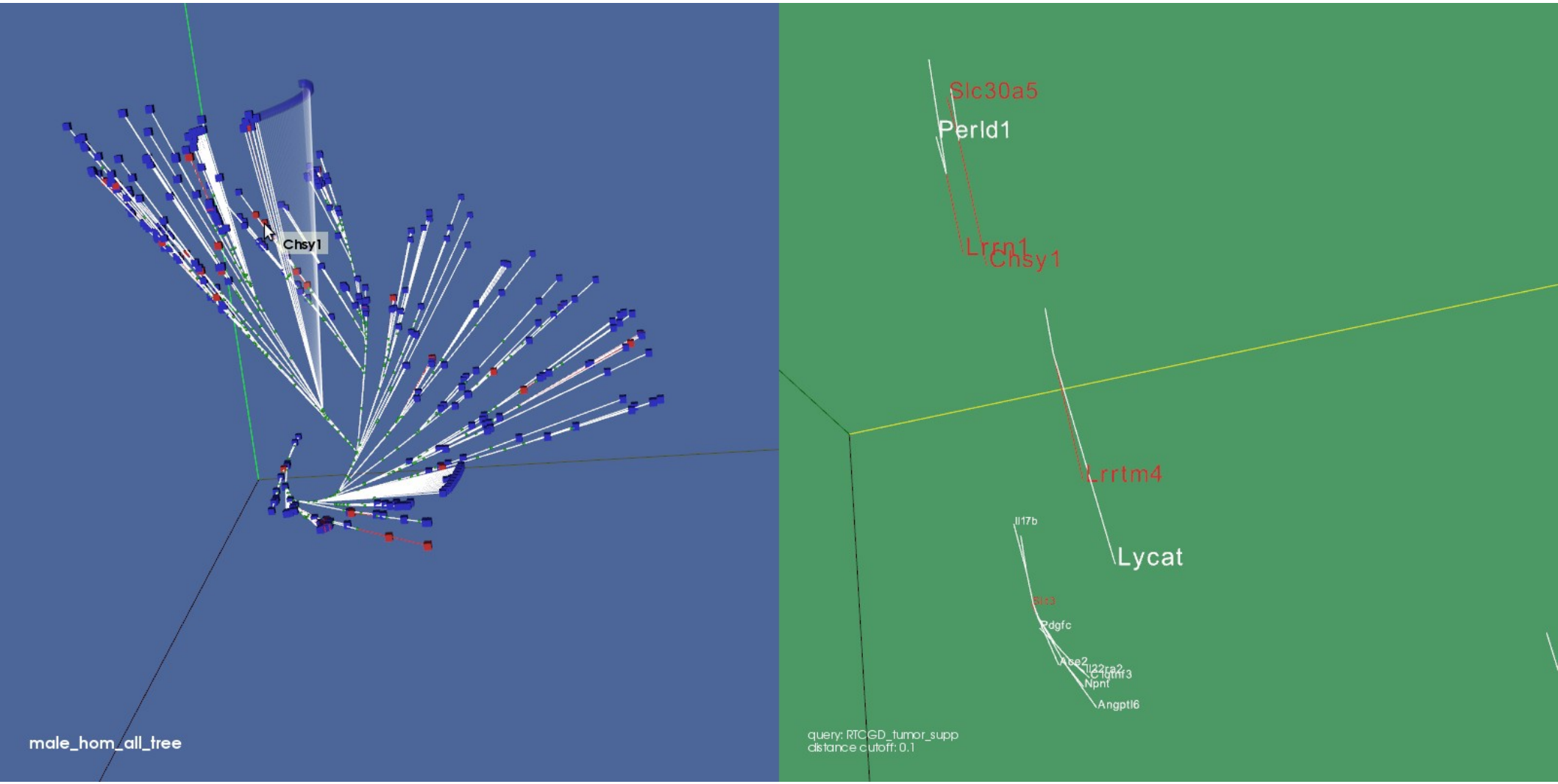
- Phenotypes
  - Piloerection
  - Exophthalmus
  - Freezing
  - Rearing
  - Abnormal Gait
  - Whiskers
- Many oncogenes and tumor suppressors are involved in neural development

Interactive 3D Visualization

- Flexible cluster building
- Output will overwhelm 2D phenogram space
- Lists of genes is a cumbersome option
- 3D system
  - Visualization Toolkit
  - Originally developed at GE
  - Open Source
  - C++ library, very OOP
  - Python, Java, and Tcl hooks


3D view of outtree
theta: 0.001
phi: 0.005
red nodes: RTCGB tumor suppressors

Visual Probing for Subtrees



KOMPCluster as a Web Service

- 3D display on the web remains difficult
- X3D group is working on HTML standard inclusion, but not there yet
  - Stand alone application
  - Flash
- 2D projections
- Take advantage of federated databases

UC Davis Mouse Biology Program

- Kent Lloyd
- David West

- MouseBiology.org
- KOMP.org
- KOMPPhenotype.org
- GeneTrap.org
- GeneCloud.org

- Informatics Group
  - Eric Engelhard
  - Jared Rapp
  - Bowen Li
  - Patrick Fish
  - Dave Clary