

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 $\sim 10^{40}$ 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

Welcome Guest, please [Log in](#)



[Search](#) [Explore](#) [Data](#) [Services](#) [About](#)

Phenotype Category
All Phenotypes

Gender
 Male / Female Male Female Hom / Het Het Hom

Exception Level: 90%

Legend
 No exceptions Less exceptions More exceptions

Additional Resources
View a [3D representation](#) of this data. [What is this?](#)

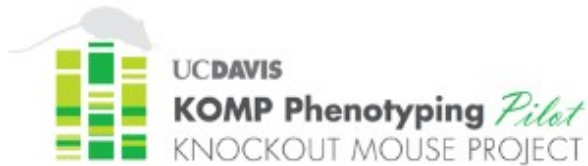
Currently displaying accumulative quantile exceptions at 90 percentile

0610907C21Rik	0610031J06Rik	1110012D08Rik	1110032E23Rik	1200015P23Rik	1500015O10Rik	1700040033Rik	2010209012Rik	2210415P13Rik	230002M23Rik	2310016C10Rik	2310044H10Rik	2310006D16Rik	2610109H07Rik	2610528A11Rik	2900046G09Rik	2900044A13Rik	3110079O15Rik	3830408D24Rik	443241G033Rik	4632428N05Rik
4930455C21Rik	4931408A02Rik	4932417I16Rik	5330417C23Rik	5930434B04Rik	6030498B19Rik	6030425E11Rik	6030607L17Rik	9130213B05Rik	9490015O10Rik	A430084P05Rik	A530080P10	Ace2	Acp1	Acp5	Acp2	Acsbg2	Acsr2	Acs5	Adem18	Adam30
Adam32	Adam33	Adams12	Adams11	Agpat4	Agpat6	Agr3	Ai402493	Ai747448	Alb	Alg2	Amica1	Angpt1	Angpt2	Angpt4	Angpt6	Angpt7	Aph1c	Aplin	Apoa5	
Ang	Aspn	Atp13a5	AU021092	BC004004	BC026682	BC037156	BC048599	BC059923	BC064033	Boc	Btn9	C1qtnf1	C1qtnf2	C1qtnf3	C1qtnf5	C1qtnf9	C330028O21Rik	Cadm1	Cadm3	
Calcao1	Car14	Car9	Cbln3	Cbln4	Ccbe1	Cdc126	Cd164f	Cd209r1	Cd300e	Cd300b	Cd309g	Cd320	Cdgn1	Cdcp1	Cdh19	Ceacam19	Chod1	Chfr	Chrt2	Chst14
Chyl1	Cica5	Cldn1	Cldn14	Cldn18	Cldn2	Cldn6	Cldn8	Clec12b	Clec14a	Clec1a	Clec1b	Clec3a	Clec4e	Clec4g	Clec7a	Crm2a	Cndp1	Cnm	Cnpy3	Cnpy4
Cog7	Colect11	Cpa4	Cpxm2	Crb3	Crelid1	Crelid2	Crm1	Crispid1	Crispid2	Crf3	Crtam	Csgalnct1	Csrp2b	Cst9	Cthrc1	Ctsl	Cuzd1	Cxcl14	Cxcl16	Cxcl17
Cyp4x1	Cyt11	D15Mg127	Dgat2	Dhrs3	Dhrs4	Dhrs7b	Dkk2	Dkk3	Dkk4	Dmsk	Dnajc10	Dok	Dpep2	Dpy19z	E430002G05Rik	Efemp2	EGG20393	EGG24219	Egfb	Egfb
Enp1	Enpp1	Enpp5	Enpp6	Enpp7	Enpp8	Epha6	Erpp1	Esam	Far1	Fcnsd1	Fcrl1	Fcrl5	Fcrs	Fgfr15	Fgfr21	Fgfr22	Fgf23	Fgfrr1	Fhcd1	Frt1
Fint2	Fint3	Fndc3b	Fat3	Fuca2	Fytd4	Gaint14	Gepd5	Gfral	Gkn2	Glib12	Glipr11	GRB1	GRB2	Gm128	Gm94	Gpa33	Gpc4	Gpc6	Gphbp1	Gpr113
Gpr114	Gpr125	Gpr183	Gpr56	Gpr89	Gpr7	Gremd1c	Hamp	Hapln3	Havcr1	Hdgfrp2	Hepacam3	Hgfac	Hhlp	Hhpg2	Hsd17b11	Hsd17b13	Hsd17b14	Ifnl1	Ifnl3	Igf13
Igf9	Ilf7b	Ilf7c	Irf4	Irf7	Ilf7b	Ilf7c	Ilf7d	Ilf7e	Ilf10	Ilf5	Ilf9	Ii20	Ii20b	Ii22	Ii22a1	Ii22a2	Ii3ra	Irfa1	Isir1	Irf5
Irfn1	Jam2	Jam3	Kazad1	Klf7	Kirrel2	Kirrel3	Kik1	Kik2	Klrl1	Layn	Lcn6	Lct1	Leprot1	Lgr6	Lhfp3	Lingo1	Lman1	Lnk1	Lrig3	Lrp10
Lrrc20	Lrrc33	Lrrc4	Lrrc4c	Lrrc8a	Lrrm1	Lrrm2	Lrrm3	Lrrm4c	Lrrtm1	Lrrtm3	Lrrtm4	Lycat	Lynx1	Lypd1	Lypd2	Lypd4	Lypd5	Lypd6	Lyp1a3	Lyz1
Magt1	Mai	Mai1	mammaglobin-1	Manc1	Megr9	Mendc2	Mett7a1	Mfsd2	Mfsd7	MST2446259	Mia3	Mmp27	Mxr1	Mxr2	Mxra3	Mxd	Npr105	Msa414	Msa4a6	Msa4c
Msa4a6	Msa4a6b	Msa4a6c	Msa4a6d	Muc15	Muc16	Muc20	Mupcdh	Mvra8	Myadm	Nbea212	Neto2	Nnat	novel MEN	novel secreted	Npat	Nrm1	Ntng1	Nxph4	Oit1	Oit3
Olfm3	Olfm4	Olfm5	Olfm78	ORF9	ORF9	Olfm1	Olfm1	Olfm1	Olfm1	Olfm1	Pcolce2	Pdgfc	Pdgfc	Per1d	Pgap1	Pglyrp2	Pil6	Pigt	Pkd11l1	Pivap
Pknox2	Pknox2	Pnk1	Podn	Pomnt1	Prap1	Prom2	Prss2	Prss3	Prss1	Rdn13	Reg3b	Reg4	Retn	Retnib	Retstat	Ric3	Robo4	Rspny1	Rtn4r	Sbsn
Sgpl3a1	Scn2b	Skep1	Scrp1	Sdcag10	Sell1	Sema4a	Sema4b	Sema6f	Serpine10	Sez6f	Sez6i	Sfrp2	Sh2a3c	Shisa2	Shisa4	Sid2	Sigrr	Siglec3	Siglec5	Siglec6
Slamf7	Slamf8	Slamf9	Sic25a23	Sic25a3	Sic2a3	Sic2a3	Sic25a3	Sic35e3	Sic39a14	Sic39a6	Sic39a8	Sic44a3	Sic44a4	Sic7a5	Silt3	Siltkr1	Sinox	Sost	Spaca4	Sox14
Spin6	Spin7	Spin1	Spin3	Stand3n	Steap2	Strc3	Strc6	Sulf1	Sulf2	Sulf2	Susd4	Tac2	Tb12	Tbsd1	Tmag1	Tir3	Tir7	TIR8	Tm4sf20	Tmc2
Tmco3	Tmco3	Tmem107	Tmem108	Tmem119	Tmem123	Tmem130	Tmem136	Tmem149	Tmem161b	Tmem179	Tmem204	Tmem207	Tmem21a	Tmem43	Tmem50a	Tmem50b	Tmem57	Tmem77	Tmem81	Tmem88
Tmprs2	Tmprs3	Tmprs4	Tmprs6	Tmub1	Tmub2	Tnfrsf19	Tnfrsf25	Tnmd	Tobg	Trem1	Trem2	Tnhd	Trpm2	Tsku	Tspan12	Tspan18	Txndc1	Txndc12	U40088	Unc5b
Uts2	Ux1	Vapb	Vasn	V5	Vtcn1	Vwc2	Wfdc12	Wfdc5	Wf1	Wnt9b	Zdhnc11	Zdhnc24	Zdhnc9							

The KOMP Phenotyping Project is funded by an ARRA grant to UC Davis and CHORI
Questions? Comments? Please contact us: 1-888-KOMP-MICE or service@komp.org



KOMPphenotype.org



- [Search](#)
- [Explore](#)
- [Data](#)
- [Services](#)
- [About](#)

Gene Information

Fgf23

Name: fibroblast growth factor 23

Resources: [MGI:1891427](#) [NCBI:64654](#) [OMIM](#), [PubMed](#), [Gene Cloud](#), [Additional resources at KOMP](#)

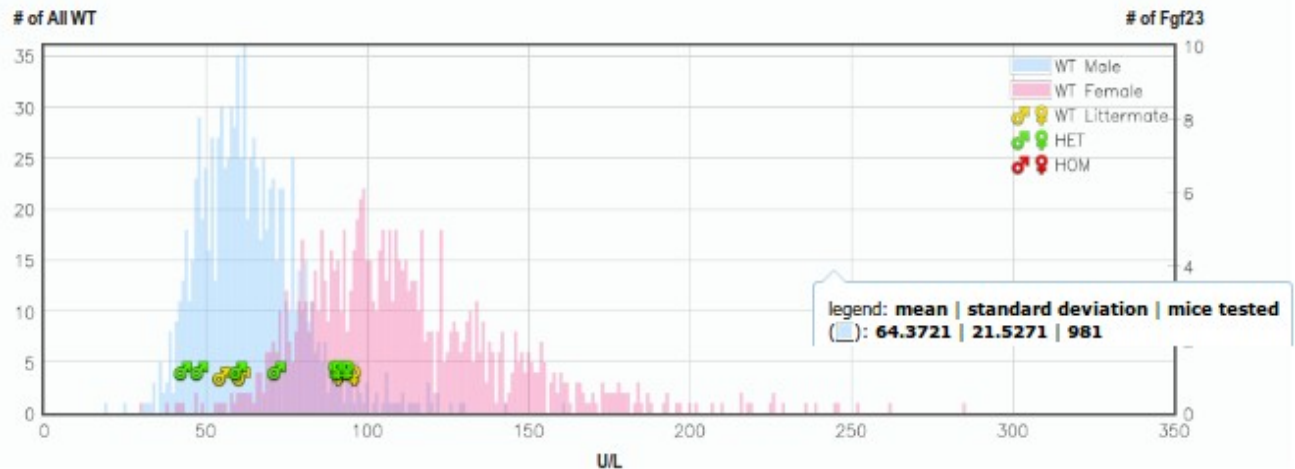
Products: [Cryo-archive](#)

- [Summary](#)
- [LacZ Expression](#)
- [Blood Chem](#)
- [Cardiology](#)
- [Growth/Gross/Histology](#)
- [Immunology](#)
- [Metabolism](#)
- [Neurology & Behavior](#)

Jump to:

[View Heatmap](#) [Expand All](#) [Collapse All](#)

- [Blood Chemistry - Alanine Aminotransferase](#)
- [Blood Chemistry - Albumin](#)
- [Blood Chemistry - Alkaline Phosphatase](#)

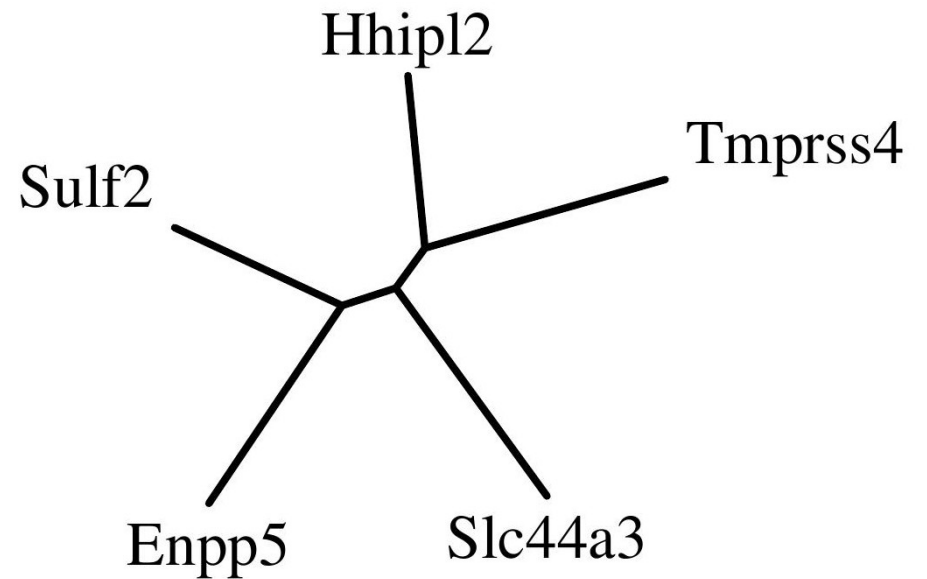
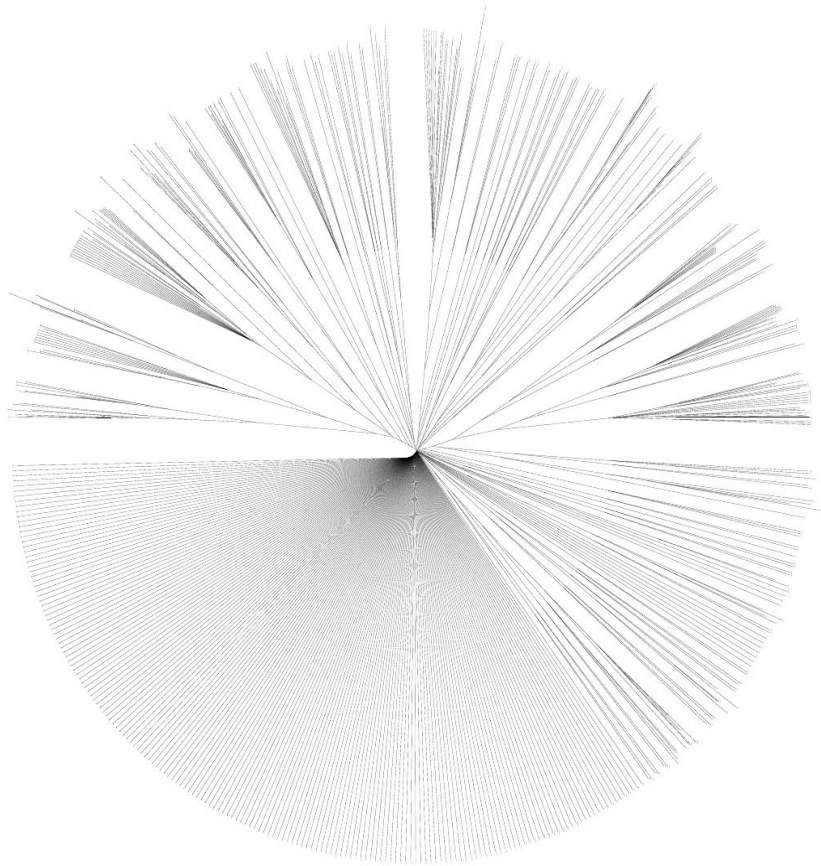


[Download Data](#) | [Heatmap](#) | [Top](#)

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

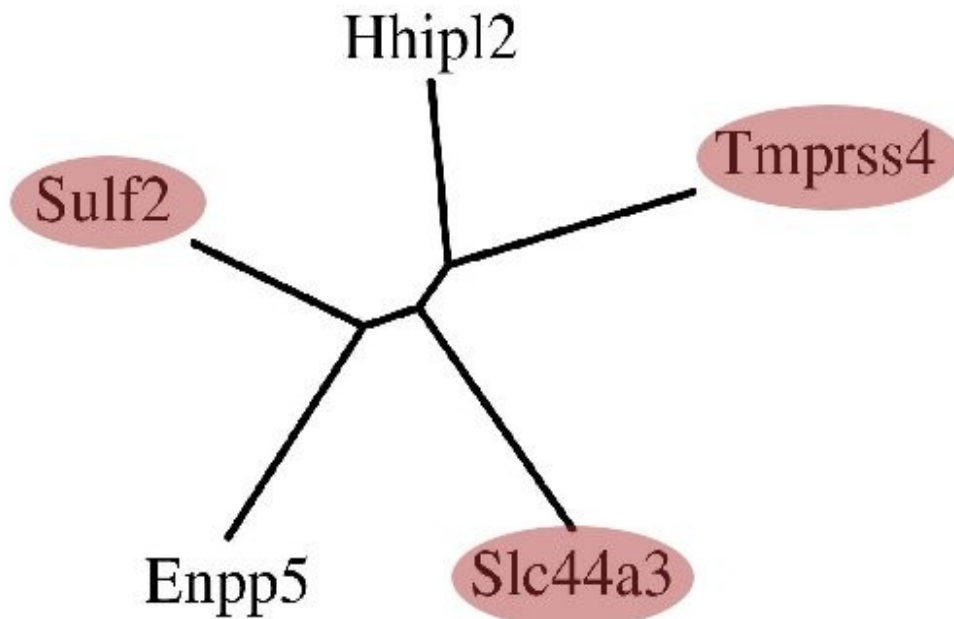


Clusters defined by specifying branch distances

Tumor Suppressors

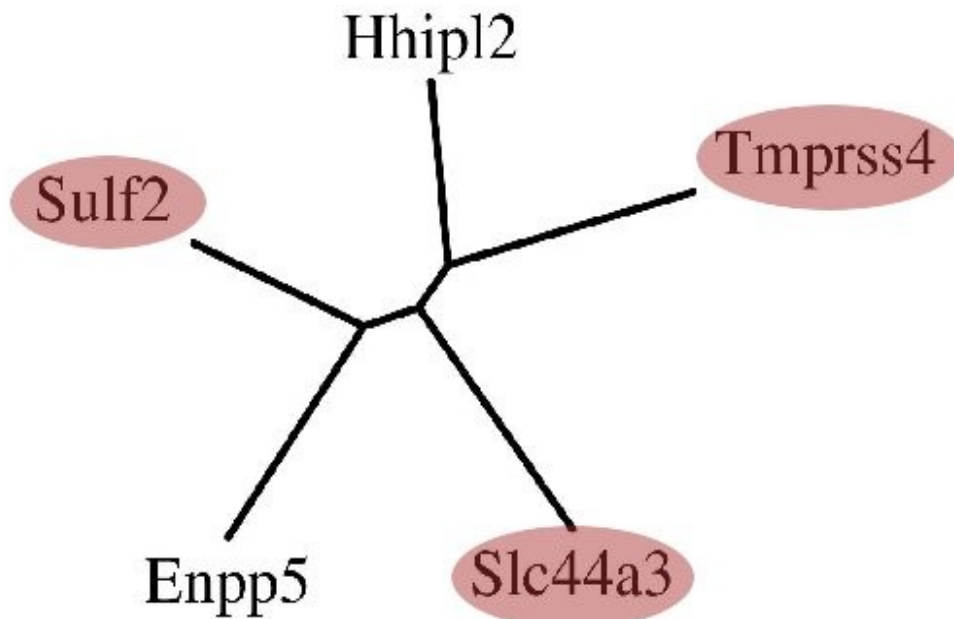
- Data from the Retrovirus Tagged Cancer Gene Database (RTC GD) mapped against the current mouse genome assembly and annotations
- Conservative algorithm flagged multiple, independent exon/intron insertions (truncated gene products)
- 849 candidates of which 16 were represented within the data

Co-Clustering



- Sulf2, Tmprss4, and Slc44a3 in same cluster
- Enpp5 may play a role in neuronal cell communication (SwissProt)
- Hhip12 is a homolog to HHIP12, a gene deregulated in gastric carcinomas
- p approx 5×10^{-7}

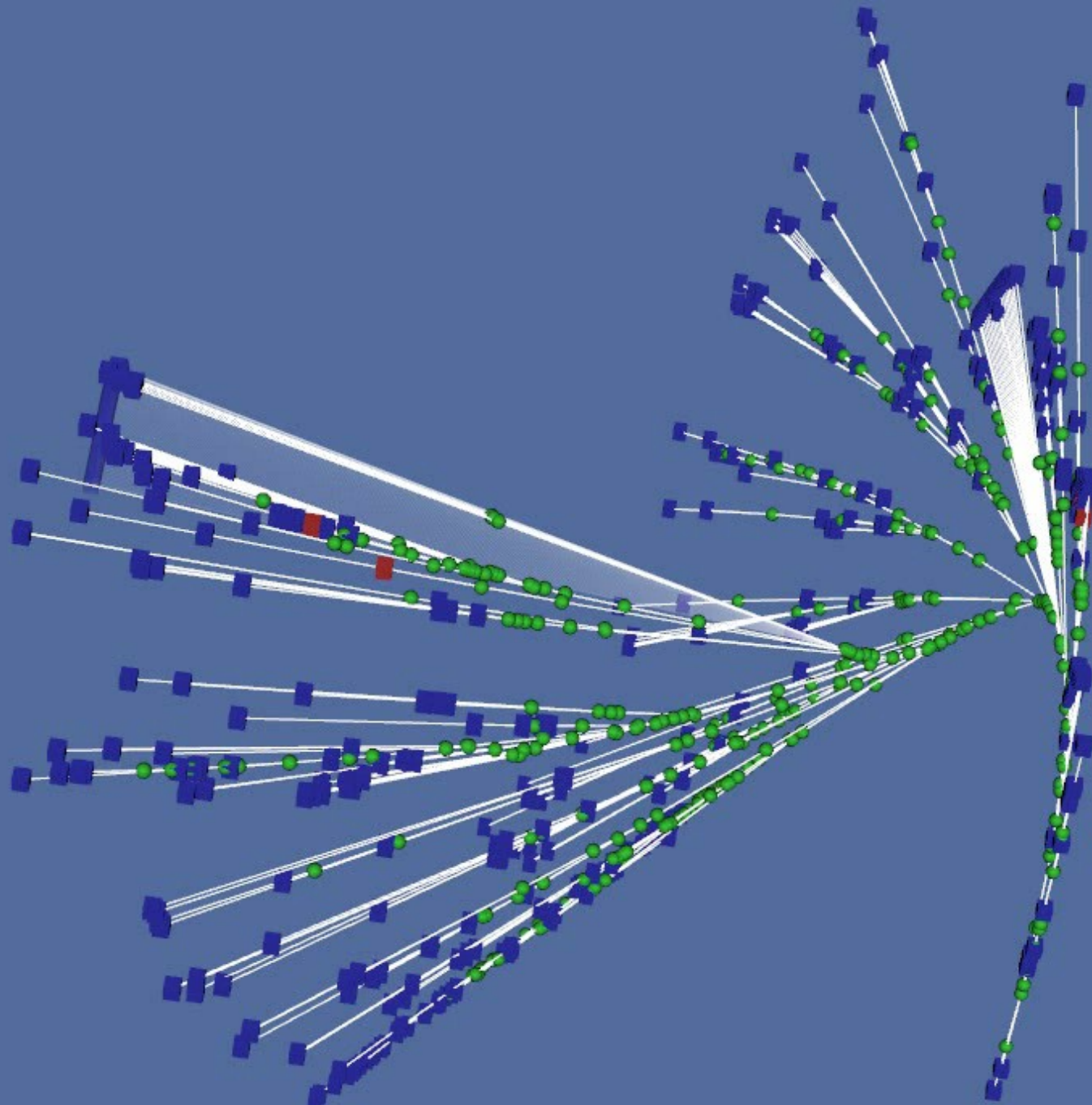
Clustered by Behavioral Phenotypes



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

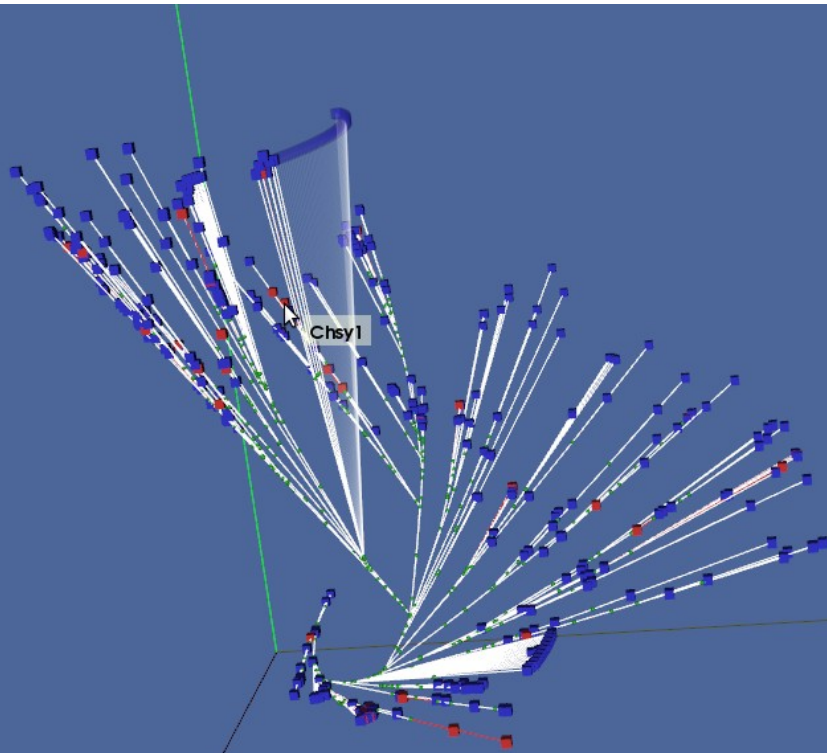
Interactive 3D Visualization

- 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: R1CG8 tumor suppressors

Visual Probing for Subtrees



male_hom_all_tree

Slc30a5

Per1d1

Lrrn1
Chsy1

Lrrtm4

Lycat

Il17b

Slc3

Pdgfc

Ase2

Il2ra2

Il11r3

Npnt

Angptl6

query: RTCGD_tumor_supp
distance cutoff: 0.1

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