KOMP-312 LIMS and Data Presentation

Eric Engelhard, Ph.D.
Director of Informatics
UC Davis Mouse Biology Program
ekengelhard@ucdavis.edu

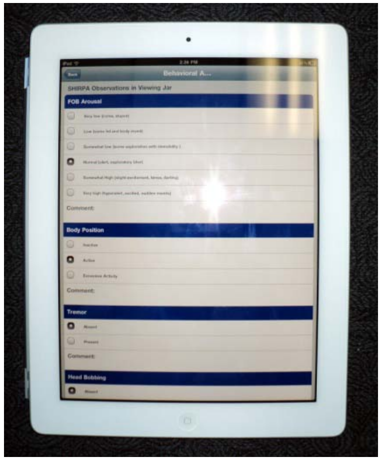
LIMS: Coordination

- Coordination between local and remote laboratory groups
  - Scheduling
  - Data integration and reporting

- Internal
  - MBAL - Mouse Behavioral Assessment Laboratory
  - MBML - Murine Biomethods Laboratory
  - MCRL - Murine Cryopreservation and Recovery Laboratory
  - MGAL - Murine Genetic Analysis Laboratory
  - MMCL - Murine Molecular Constructs Laboratory
  - MSCL - Murine Stem Cell Laboratory
  - MTGL - Murine Targeted Genomics Laboratory
- Remote
  - DTCC Consortium
  - Mouse Metabolic Phenotyping Center (MMPC)

LIMS: Data Acquisition

- Devices
  - Workstations
  - Hand-held devices
  - Laboratory automation
- Ergonomics
  - Custom views
  - Simplification
  - Feedback



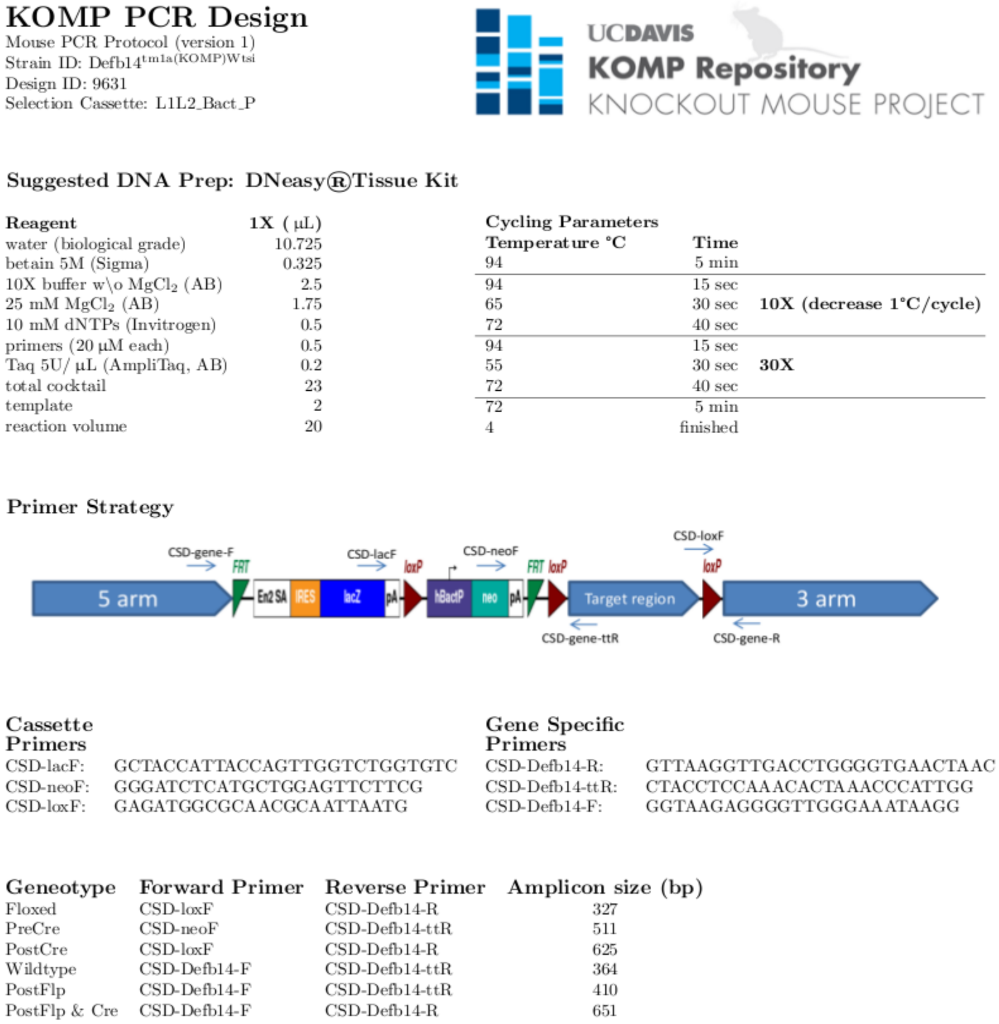
Infrastructure for Service Availability

- LIMS
- Public Services
  - KOMP.org
  - KOMPPhenotype.org
  - MMRRC.org
  - gentrap.org
  - mousebiology.org

- LAMP stack
- Virtual Machines
- Redundancies
  - SAN storage, distributed server hardware, multiple networks, redundant firewalls
- Disaster recovery
  - Full data and VM backups
  - Remote storage and hosting partners

Automation: Primer Design

- Consolidation of reference data
  - IKMC
  - Ensembl
- Primer3 wrapper
  - Universal PCR conditions
  - All associated mutation types
- LaTeX generator for both web and print output



KOMP PCR Design

Mouse PCR Protocol (version 1)
Strain ID: Defb14$^{\mathrm{tm1a(KOMP)Wtsi}}$
Design ID: 9631
Selection Cassette: L1L2_Bact_P

Suggested DNA Prep: DNeasy®Tissue Kit

Reagent	1X (µL)
water (biological grade)	10.725
betain 5M (Sigma)	0.325
10X buffer w\o $MgCl_2$ (AB)	2.5
25 mM $MgCl_2$ (AB)	1.75
10 mM dNTPs (Invitrogen)	0.5
primers (20 µM each)	0.5
Taq 5U/ µL (AmpliTaq, AB)	0.2
total cocktail	23
template	2
reaction volume	20

Cycling Parameters

Temperature °C	Time	
94	5 min	
94	15 sec	
65	30 sec	**10X (decrease 1°C/cycle)**
72	40 sec	
94	15 sec	
55	30 sec	**30X**
72	40 sec	
72	5 min	
4	finished	

Primer Strategy

Cassette Primers	
CSD-lacF:	GCTACCATTACCAGTTGGTCTGGTGTC
CSD-neoF:	GGGATCTCATGCTGGAGTTCTTCG
CSD-loxF:	GAGATGGCGCAACGCAATTAATG

Gene Specific Primers	
CSD-Defb14-R:	GTTAAGGTTGACCTGGGGTGAACTAAC
CSD-Defb14-ttR:	CTACCTCCAAACACTAAACCCATTGG
CSD-Defb14-F:	GGTAAGAGGGGTTGGGAAATAAGG

Geneotype	Forward Primer	Reverse Primer	Amplicon size (bp)
Floxed	CSD-loxF	CSD-Defb14-R	327
PreCre	CSD-neoF	CSD-Defb14-ttR	511
PostCre	CSD-loxF	CSD-Defb14-R	625
Wildtype	CSD-Defb14-F	CSD-Defb14-ttR	364
PostFlp	CSD-Defb14-F	CSD-Defb14-ttR	410
PostFlp & Cre	CSD-Defb14-F	CSD-Defb14-R	651

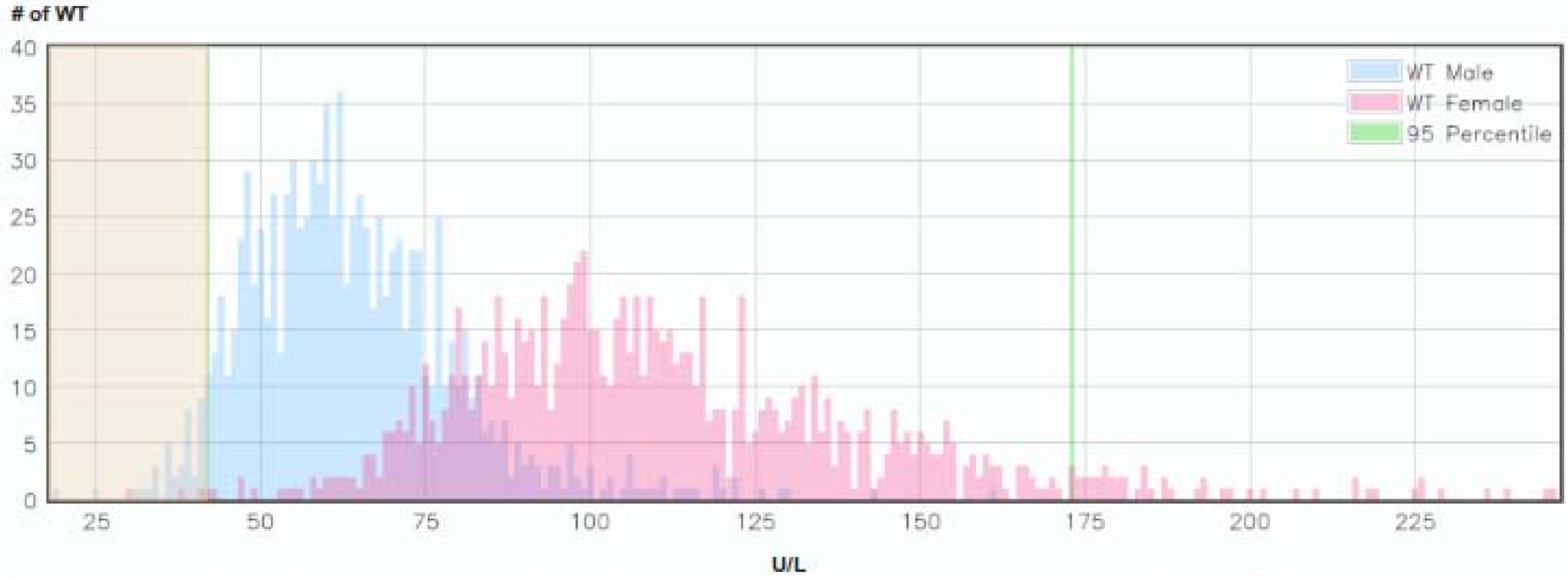
Data Presentation

- Gene-centric annotations becoming deeper and more numerous
- Unified pipelines create new opportunities for data-driven result presentation
  - User defined search images
  - Pattern recognition agents

Customizable Search Images

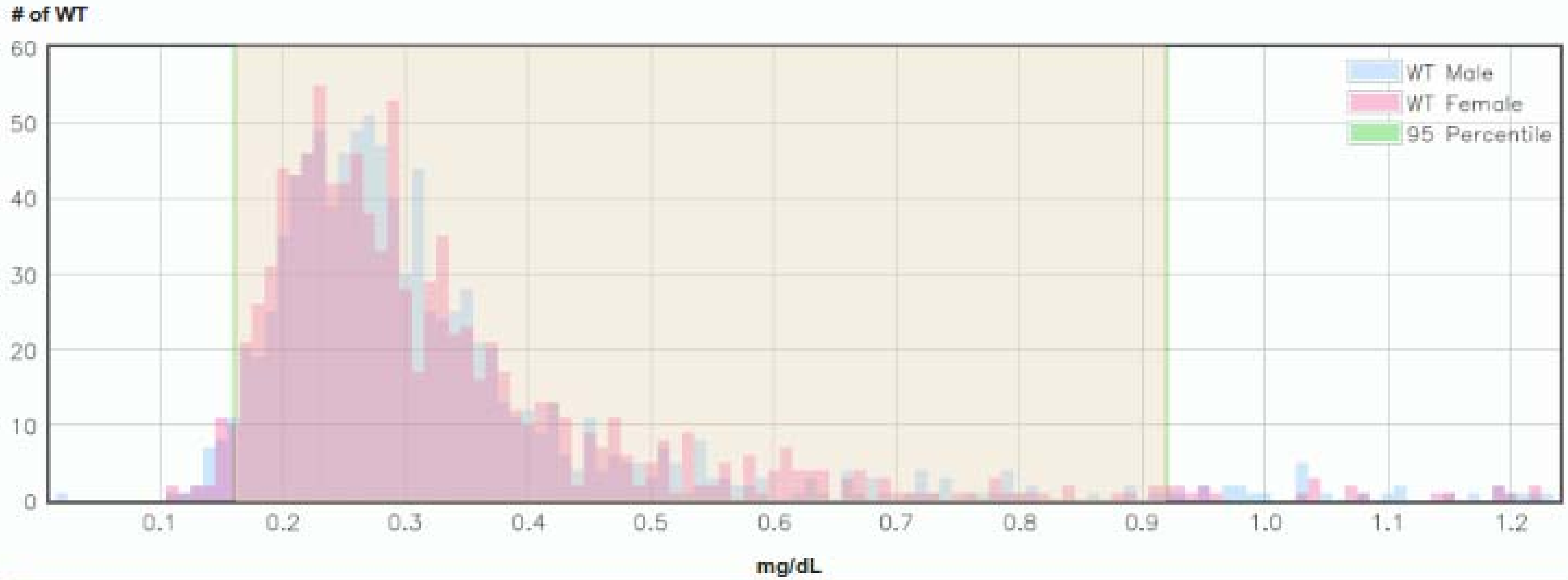
- Visual Query Builder
  - Effective communication of data structure
  - Easy range selection
  - Boolean combinations
    - AND, OR, NOT
    - grouping
  - Feedback for both individual constraints and combined filters

Alkaline Phosphatase

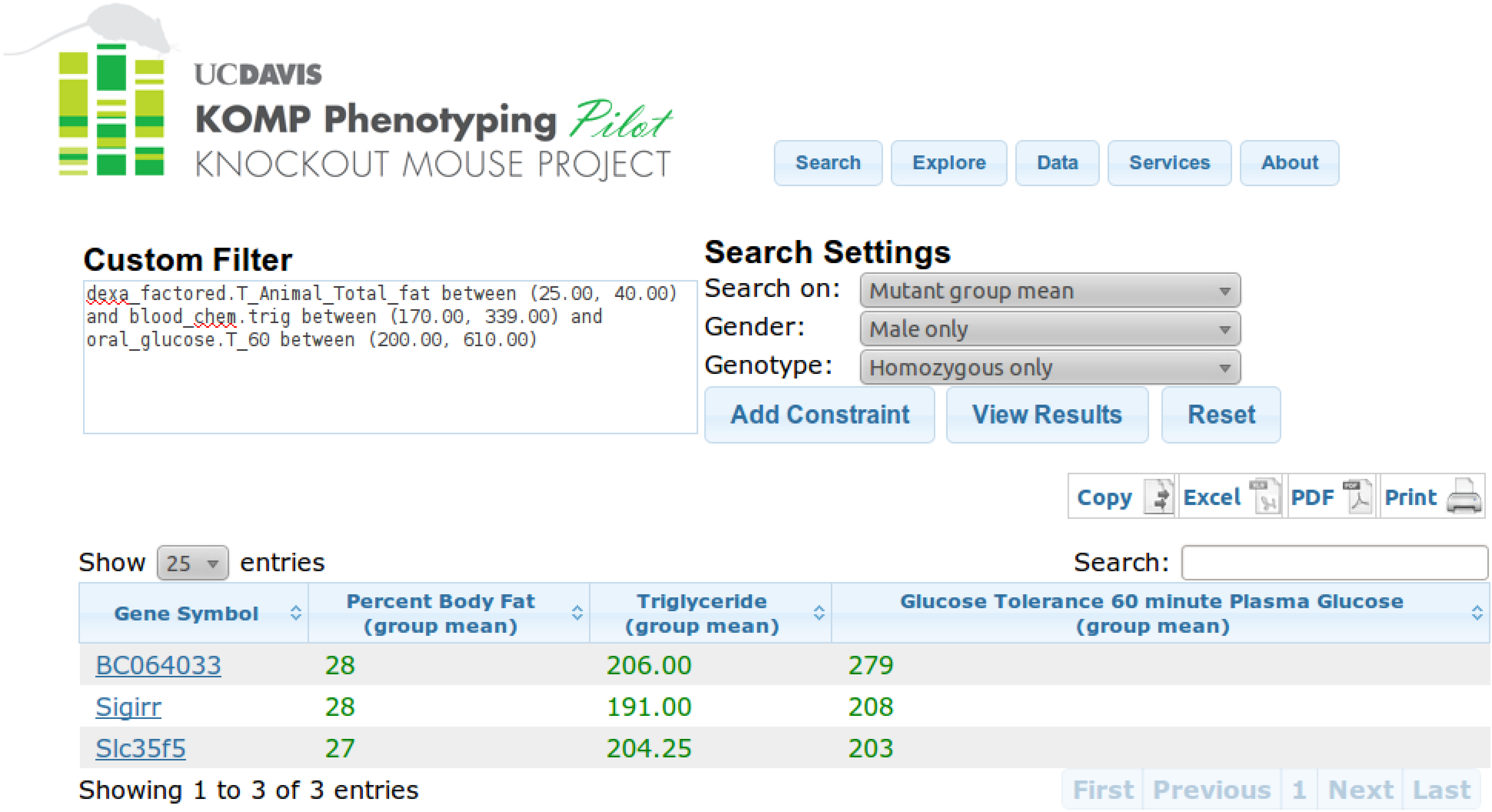


123 genes have mutants testing within the range (18.00, 41.90). 0 genes intersect with existing filter. Add to filter
487 genes have mutants testing outside of the range (18.00, 41.90).0 genes intersect with existing filter. Add to filter

Bilirubin Total



487 genes have mutants testing within the range (0.16, 0.92). 0 genes intersect with existing filter. Add to filter
232 genes have mutants testing outside of the range (0.16, 0.92).0 genes intersect with existing filter. Add to filter



Custom Filter

```
dexa_factored.T_Animal_Total_fat between (25.00, 40.00)
and blood_chem.trig between (170.00, 339.00) and
oral_glucose.T_60 between (200.00, 610.00)
```

Search Settings

Search on: Mutant group mean
Gender: Male only
Genotype: Homozygous only

Show 25 entries

Search:

Gene Symbol	Percent Body Fat (group mean)	Triglyceride (group mean)	Glucose Tolerance 60 minute Plasma Glucose (group mean)
BC064033	28	206.00	279
Sigirr	28	191.00	208
Slc35f5	27	204.25	203

Showing 1 to 3 of 3 entries

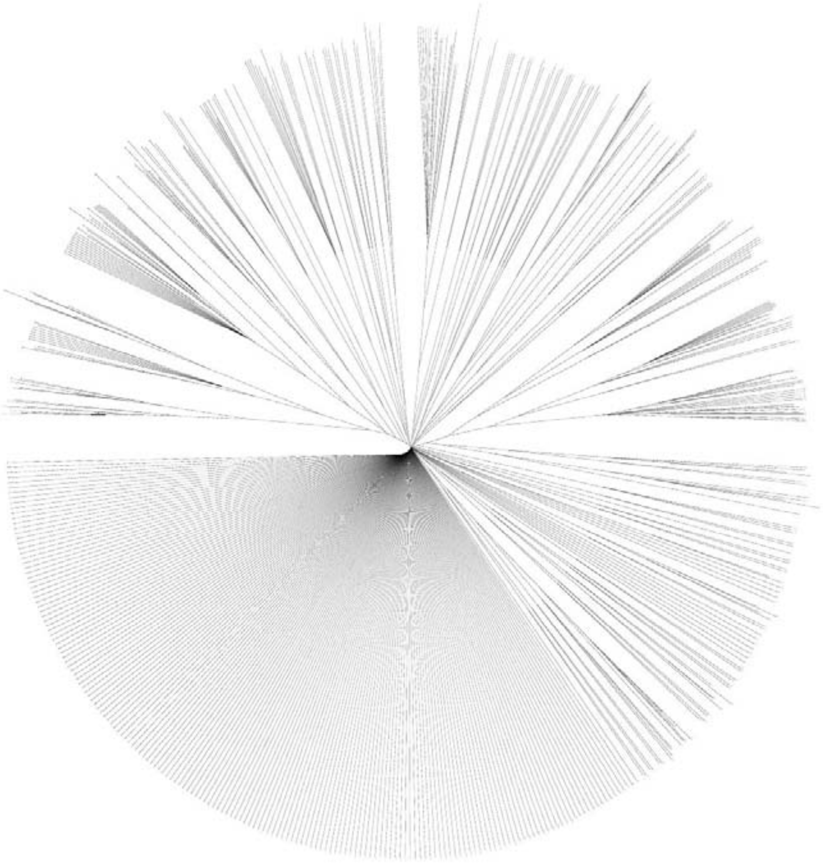
Unsupervised Pattern Recognition

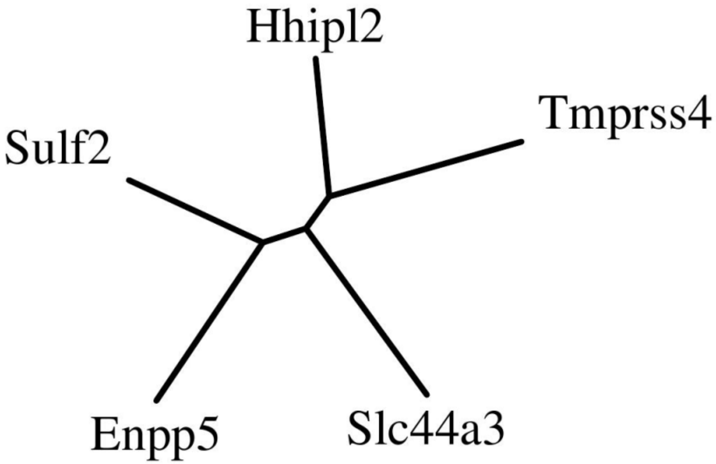
- Undefined queries allow the user to <u>observe</u> the data structure and answer questions like:
    - Are there any knockouts that share any three or more phenotype similarities (with my candidate knockout)?
    - What are the phenotypes that group these knockouts?
    - Are there subgroups of knockouts 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?

KOMPCluster

- Set Theory
  - For a given gender, genotype, genetic background each KO line is decorated with set of associated phenotype characteristics
  - Distance inversely proportional to size of intersections
- Distance Matrix
  - First order relationships between each KO line
- Neighbor Joining
  - Fast, bottom-up algorithm for establishing higher order relationships

Phenograms



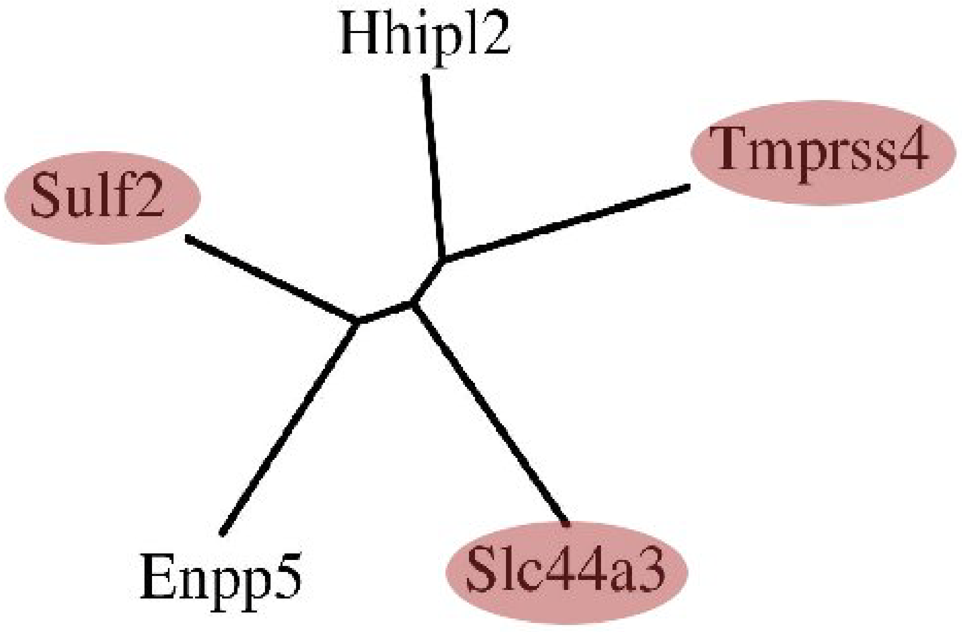


- Branch lengths indicate degrees of relatedness
- Clusters defined by specifying branch lengths

Functional Inference Example

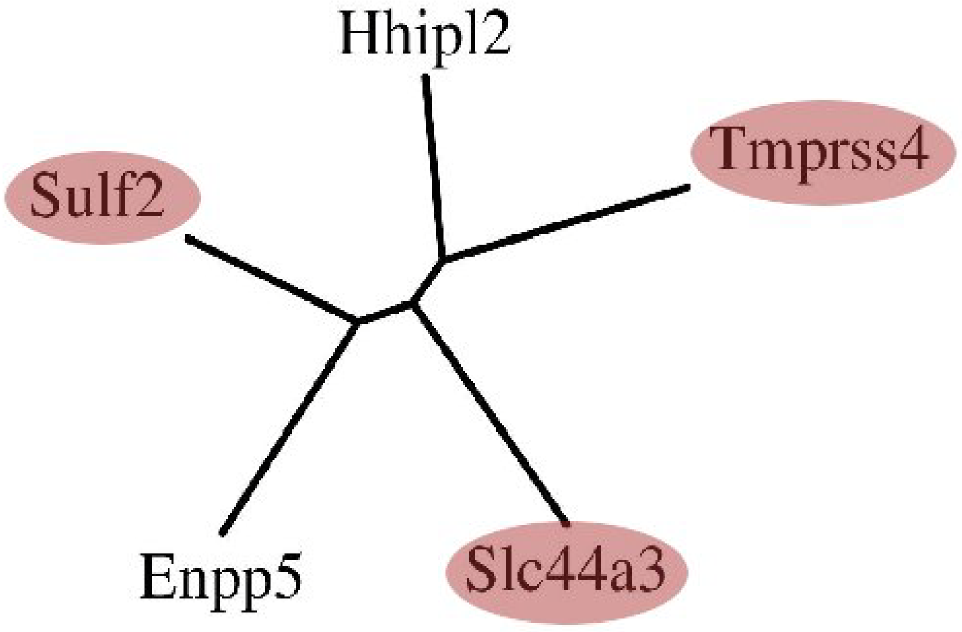
- Genentech Lexicon knockout mouse library
  - ~500 lines and associated phenotypes
  - KOMPCluster: male, hom, all phenotype exceptions
- Tumor Suppressors from the Retrovirus Tagged Cancer Gene Database (RTCGD) used to probe Lexicon Genentech
  - 849 candidates of which 16 were represented within the clustered data

Co-Clustering



- Putative tumor suppressor genes Sulf2, Tmprss4, and Slc44a3 in local cluster
- Enpp5 may play a role in neuronal cell communication (SwissProt)
- Hhipl2 is a homolog to HHIPL2, a human gene deregulated in gastric carcinomas
- $p = 5 \times 10^{-7}$

Identification of Clustering Factors



- Phenotypes
    - Piloerection
    - Exophthalamus
    - Freezing
    - Rearing
    - Abnormal Gait
    - Whiskers
- Refinement of search images

Advancing Data Discovery

- Structure determination
- Reinforcement learning
- Data Visualization and Interaction

Summary

- KOMP-312 project facilitated the development of LIMS and automation
- The KOMP2 DTCC will be building on these capabilities for both Production and Phenotyping efforts
- Large data sets can benefit from more advanced search interfaces and data-driven approaches

Acknowledgements

- KOMP Phenotyping Pilot
  - Kent Lloyd
  - David West
- UC Davis MBP Informatics Group
  - Eric Engelhard
  - Jared Rapp
  - Bowen Li
  - Patrick Fish
  - David Clary
  - Aric Horstman