Computational Infrastructure for Systems Genetics Analysis

Brian Yandell, UW-Madison

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

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

experimental context

- B6 x BTBR obese mouse cross
  - model for diabetes and obesity
  - 500+ mice from intercross (F2)
  - collaboration with Rosetta/Merck
- genotypes
  - 5K SNP Affymetrix mouse chip
  - care in curating genotypes! (map version, errors, …)
- phenotypes
  - clinical phenotypes (>100 / mouse)
  - gene expression traits (>40,000 / mouse / tissue)
  - other molecular phenotypes

how does one filter traits?

- want to reduce to “manageable” set
  - 10/100/1000: depends on needs/tools
  - How many can the biologist handle?
- how can we create such sets?
  - data-driven procedures
    - correlation-based modules
      - Zhang & Horvath 2005 *SAGMB*, Keller et al. 2008 *Genome Res*
      - Li et al. 2006 *Hum Mol Gen*
    - mapping-based focus on genome region
  - function-driven selection with database tools
    - GO, KEGG, etc
    - Incomplete knowledge leads to bias
  - random sample

why build Web eQTL tools?

- common storage/maintainence of data
  - one well-curated copy
  - central repository
  - reduce errors, ensure analysis on same data
- automate commonly used methods
  - biologist gets immediate feedback
  - statistician can focus on new methods
  - codify standard choices

how does one build tools?

- no one solution for all situations
- use existing tools wherever possible
  - new tools take time and care to build!
  - downloaded databases must be updated regularly
- human component is key
  - need informatics expertise
  - need continual dialog with biologists
- build bridges (interfaces) between tools
  - Web interface uses PHP
  - commands are created dynamically for R
- continually rethink & redesign organization

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

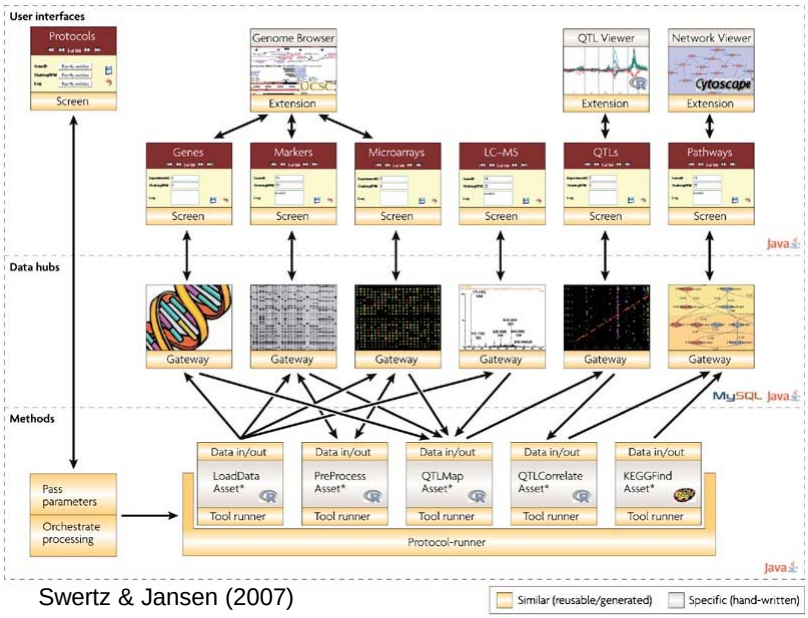
2- involve biologists early in development

3- repeated use of pipelines leads to
documented learning from experience
increased rigor in methods

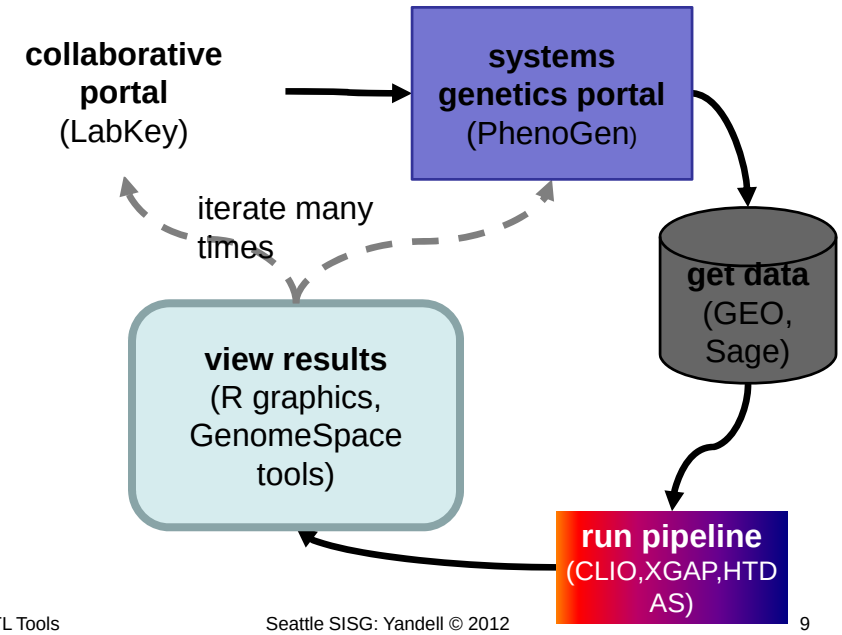
Challenges Ahead:

1- stitching together components as coherent system

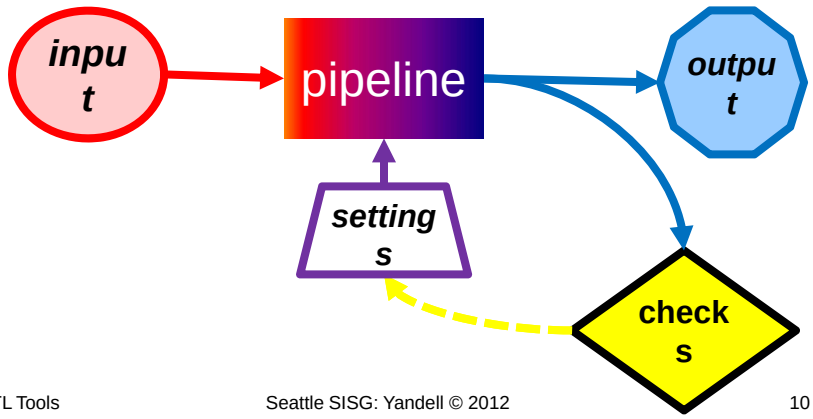
2- ramping up to ever larger molecular datasets



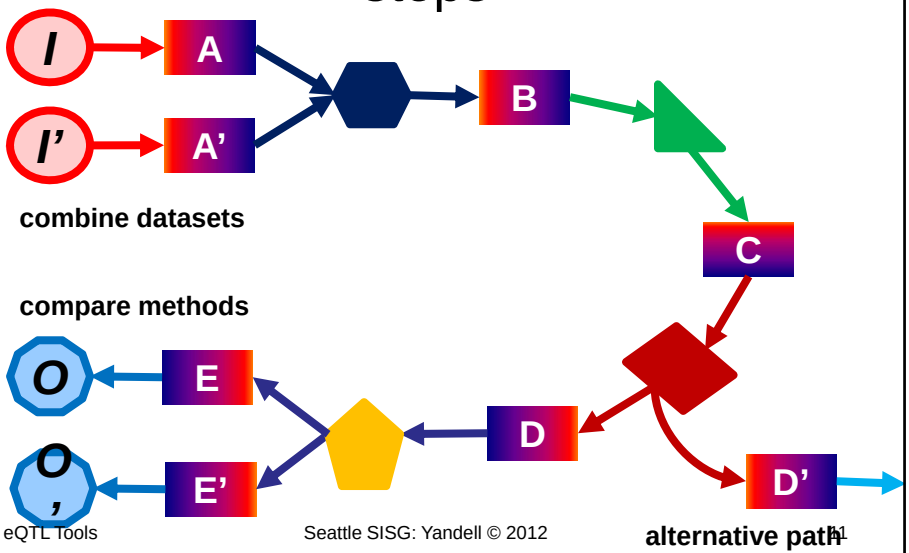
Swertz & Jansen (2007)



analysis pipeline acts on objects (extends concept of GenePattern)

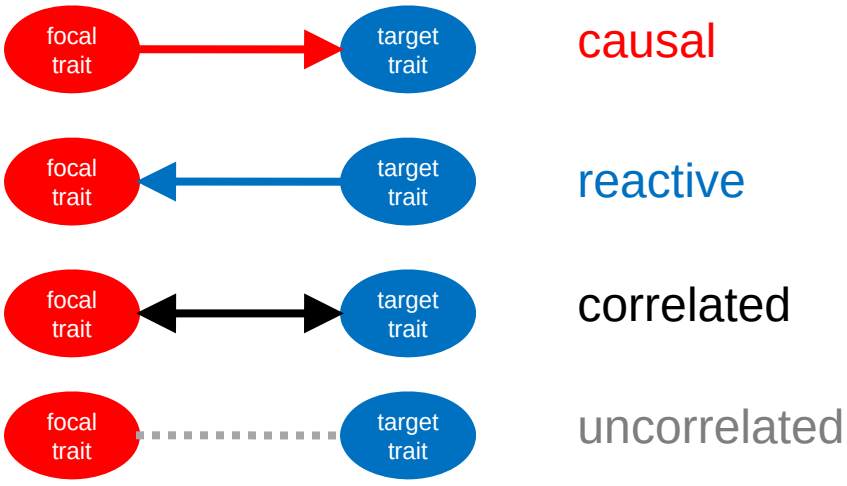


pipeline is composed of many steps

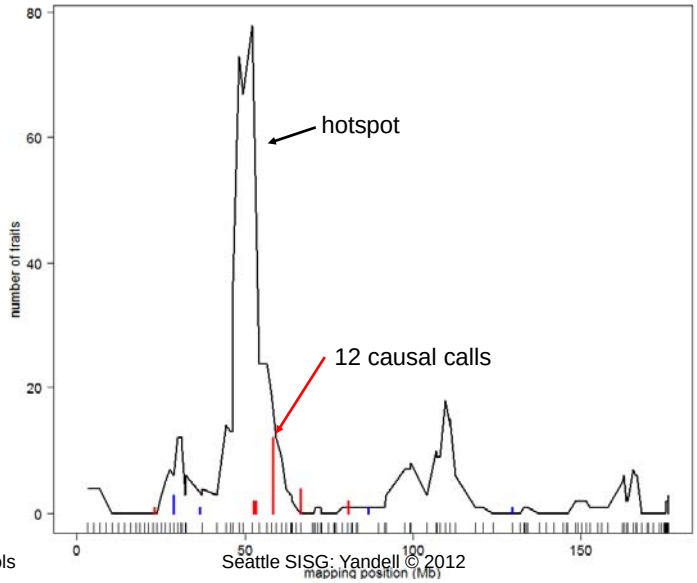


causal model selection choices

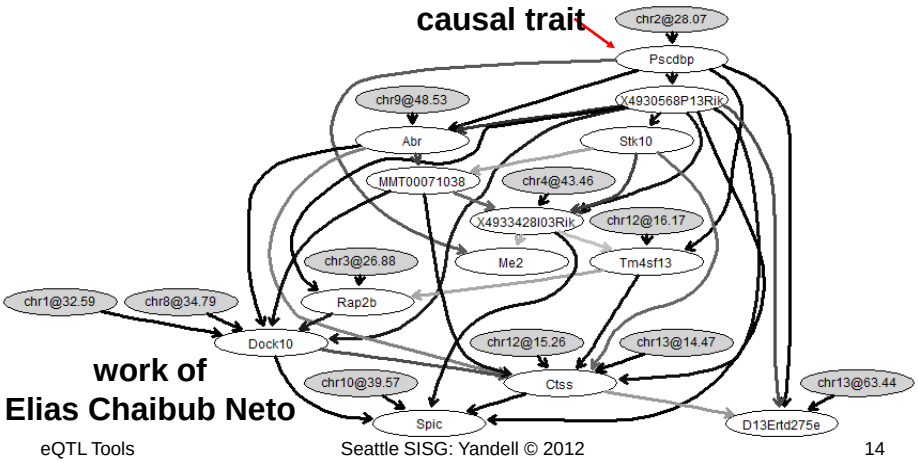
in context of larger, unknown network



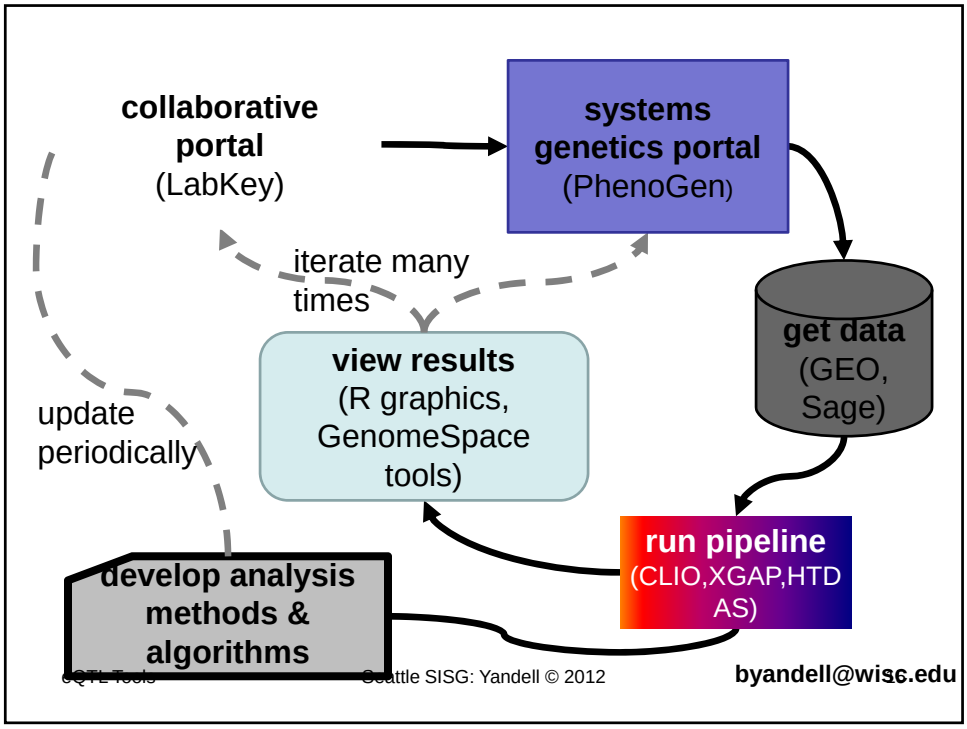
BxH ApoE-/- chr 2: causal architecture

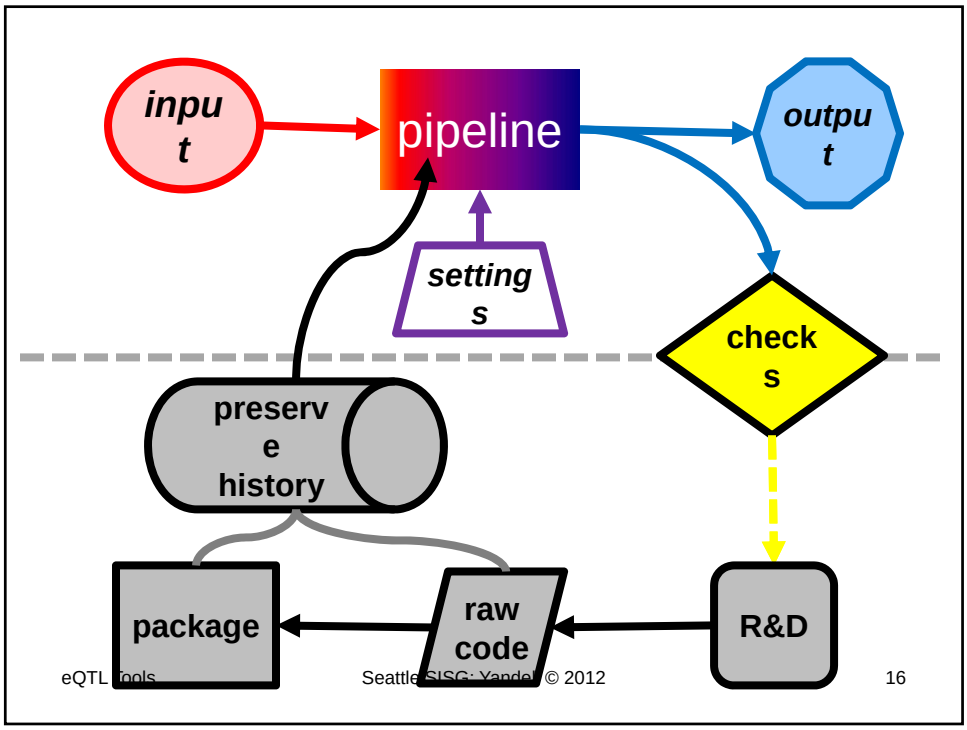


BxH ApoE-/- causal network for transcription factor Pscdbp



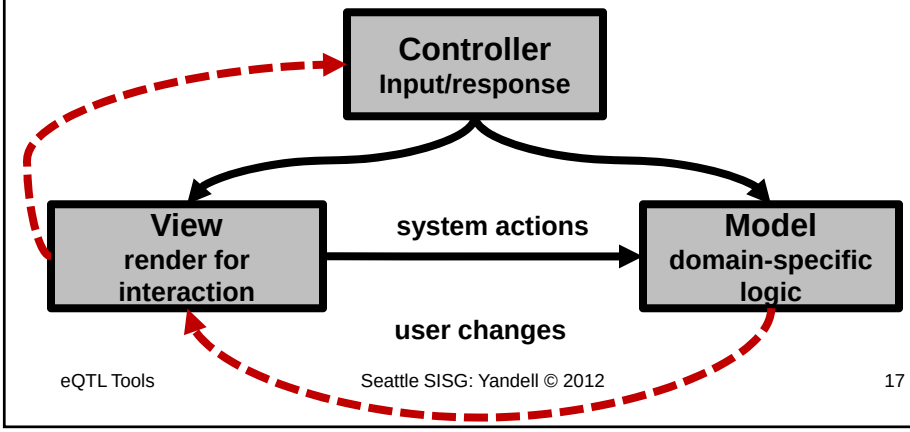
work of Elias Chaibub Neto

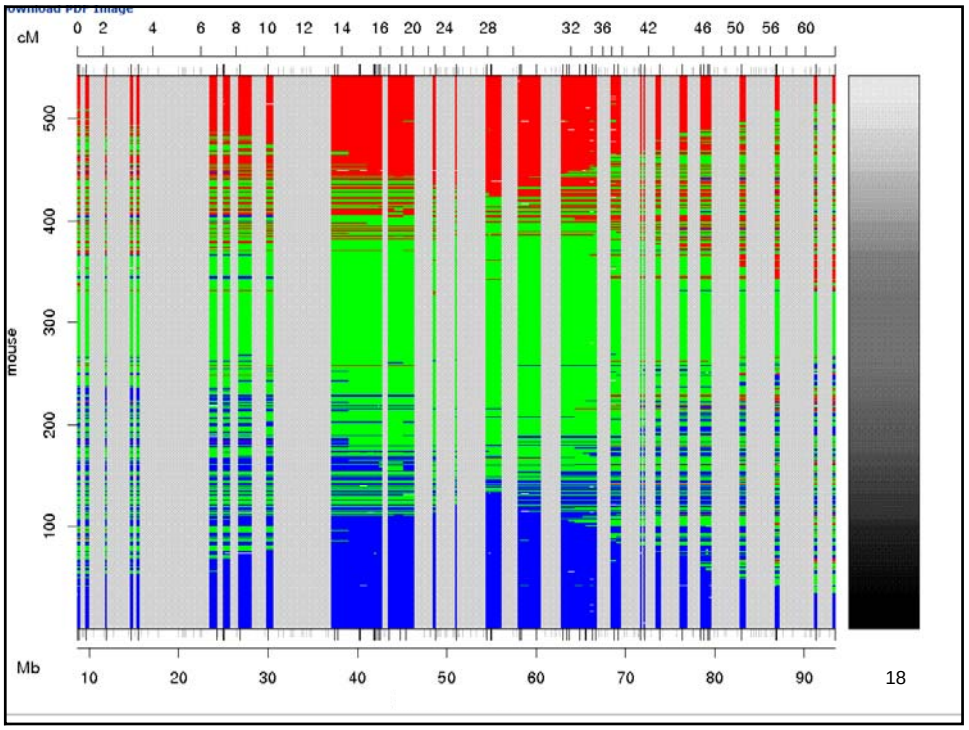

collaborative portal (LabKey)
systems genetics portal (PhenoGen)
iterate many times
get data (GEO, Sage)
view results (R graphics, GenomeSpace tools)
update periodically
run pipeline (CLIO,XGAP,HTDAS)
develop analysis methods & algorithms
eQTL Tools
Seattle SISG: Yandell © 2012
byandell@wisc.edu

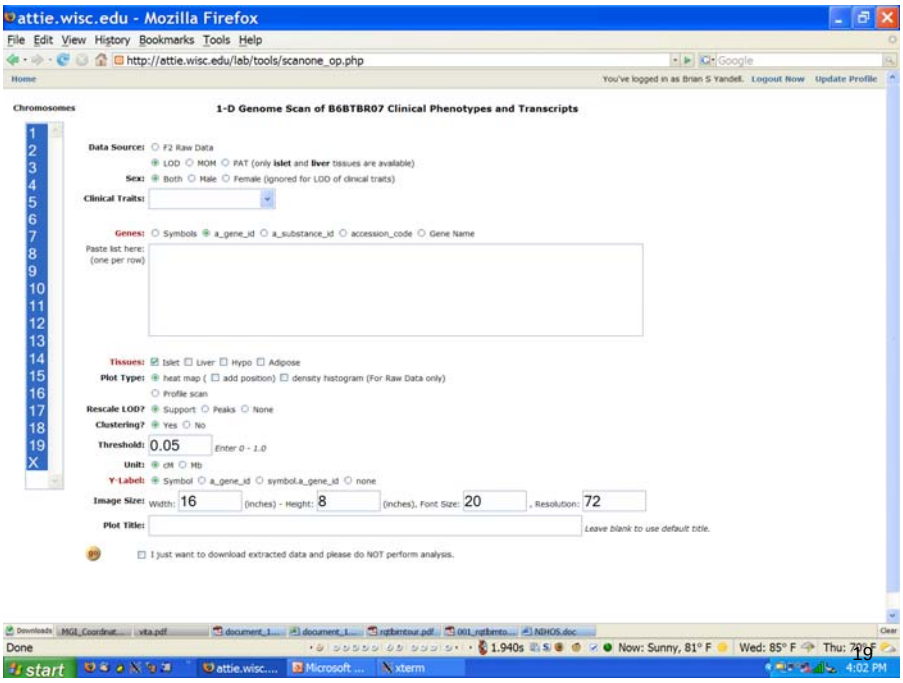

input
pipeline
output
setting s
check s
preserve history
package
raw code
R&D
eQTL Tools
Seattle SISG: Yandell © 2012
16

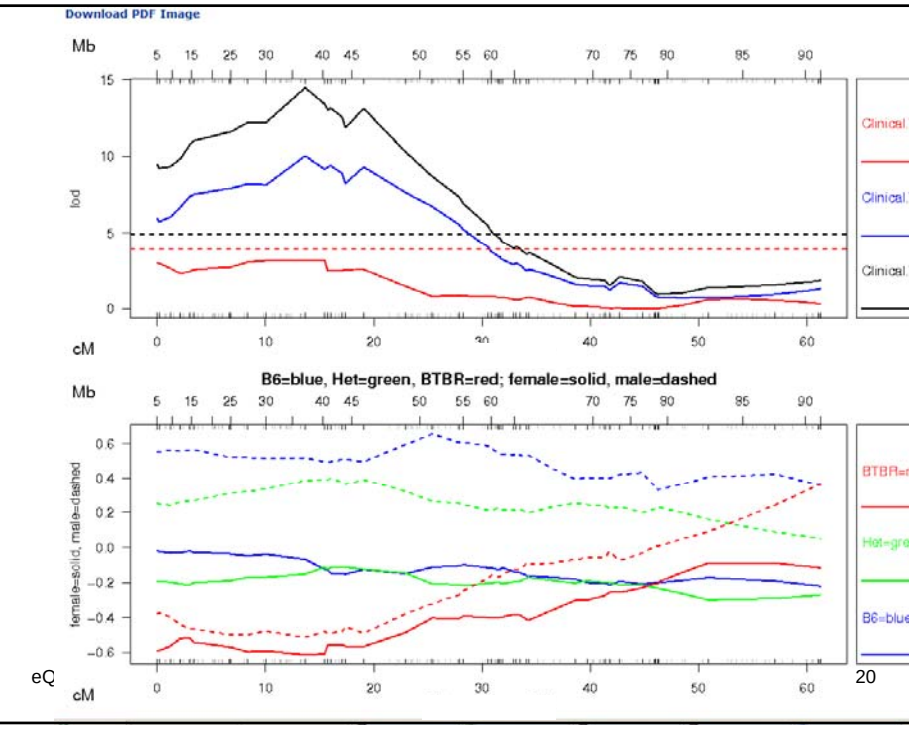
Model/View/Controller (MVC)
software architecture

- isolate domain logic from input and presentation
- permit independent development, testing, maintenance





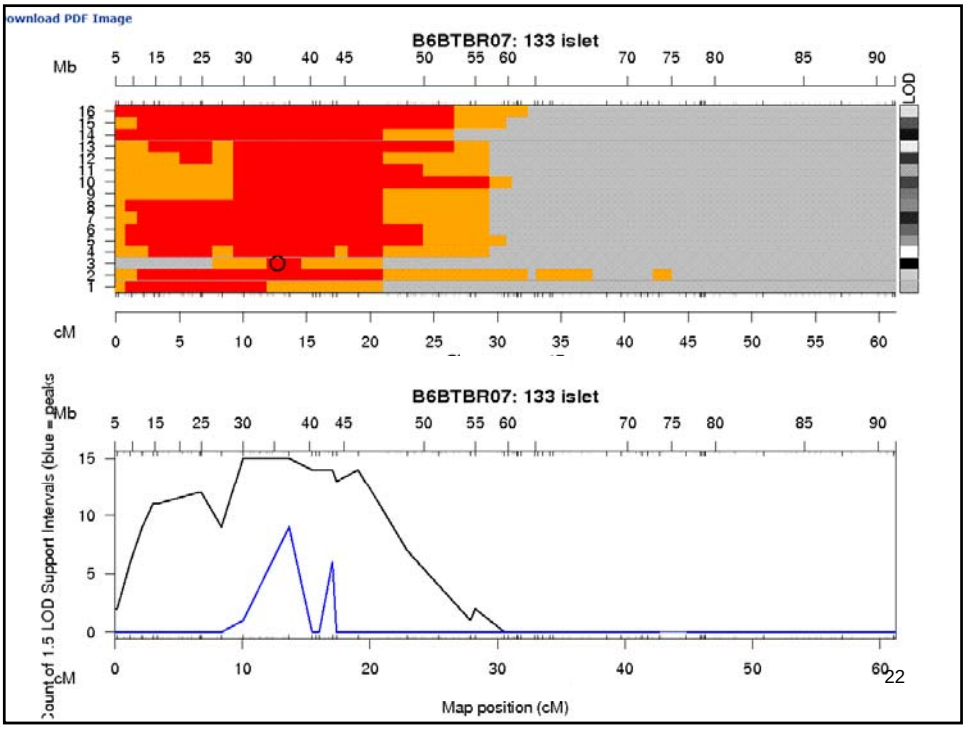

attie.wisc.edu - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://attie.wisc.edu/lab/tools/scanone_op.php
Home
You've logged in as Brian S Yandell. Logout Now Update Profile
Chromosomes
1-D Genome Scan of B6BTBR07 Clinical Phenotypes and Transcripts
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Data Source: F2 Raw Data
LOD MOM PAT (only islet and liver tissues are available)
Sex: Both Male Female (ignored for LOD of clinical traits)
Clinical Traits:
Genes: Symbols a_gene_id a_substance_id accession_code Gene Name
Paste list here: (one per row)
Tissues: Islet Liver Hypo Adipose
Plot Type: heat map (add position) density histogram (For Raw Data only)
Profile scan
Rescale LOD? Support Peaks None
Clustering? Yes No
Threshold: 0.05 Enter 0 - 1.0
Unit: cM Mb
Y Label: Symbol a_gene_id symbol.a_gene_id none
Image Size: Width: 16 (inches) - Height: 8 (inches), Font Size: 20 , Resolution: 72
Plot Title: Leave blank to use default title.
I just want to download extracted data and please do NOT perform analysis.
Done
1.940s
Now: Sunny, 81° F
Wed: 85° F
Thu:
start
attie.wisc....
Microsoft ...
xterm
4:02 PM

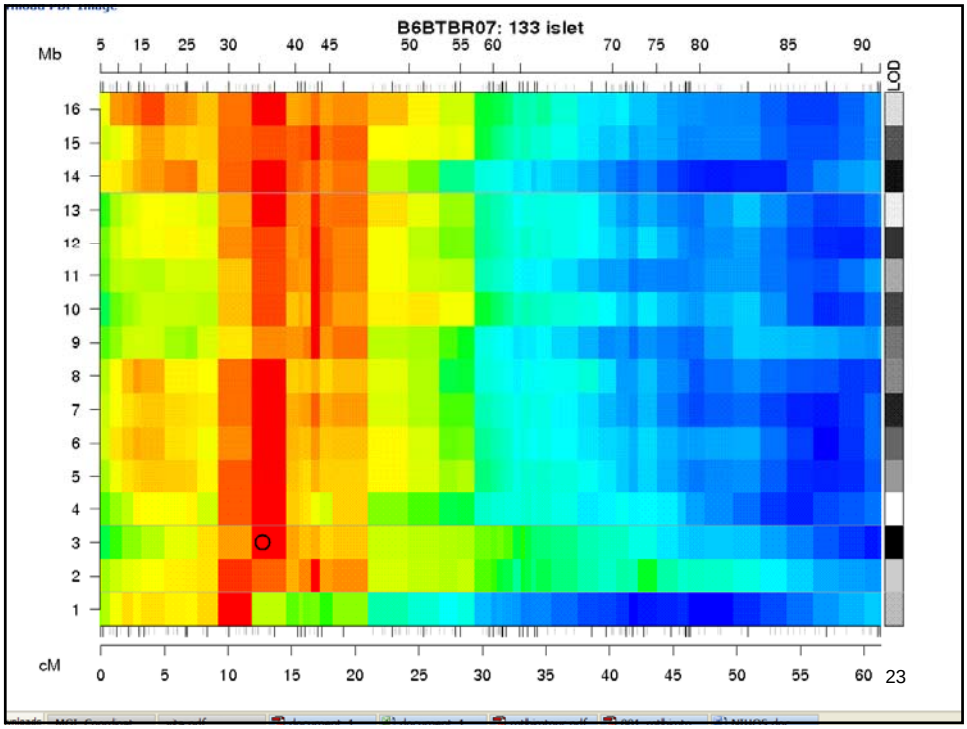

Download PDF Image
Mb
lod
cM
Clinical.Tot
Clinical.Tot
Clinical.Tot
B6=blue, Het=green, BTBR=red; female=solid, male=dashed
Mb
female=solid, male=dashed
cM
BTBR=red
Het=green
B6=blue

eQ

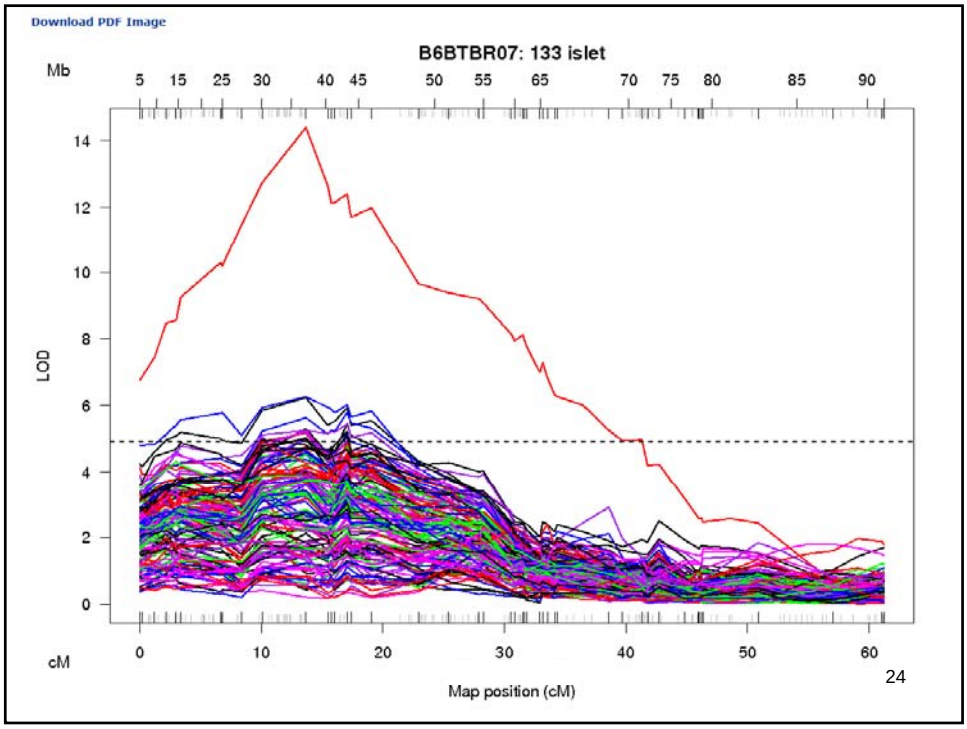
automated R script

```
library('B6BTBR07')

out <- multtrait(cross.name='B6BTBR07',
  filename = 'scanone_1214952578.csv',
  category = 'islet', chr = c(17),
  threshold.level = 0.05, sex = 'both',)

sink('scanone_1214952578.txt')
print(summary(out))
sink()

bitmap('scanone_1214952578%03d.bmp',
  height = 12, width = 16, res = 72, pointsize = 20)
plot(out, use.cM = TRUE)
dev.off()
```

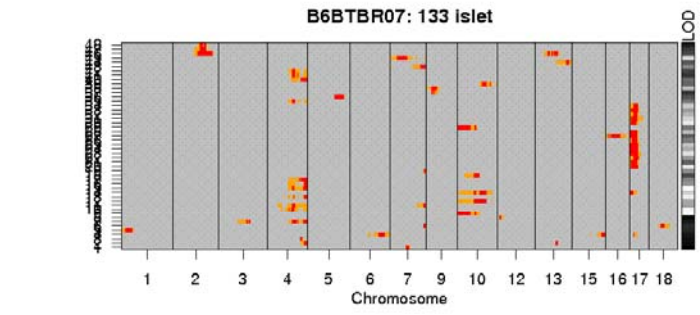



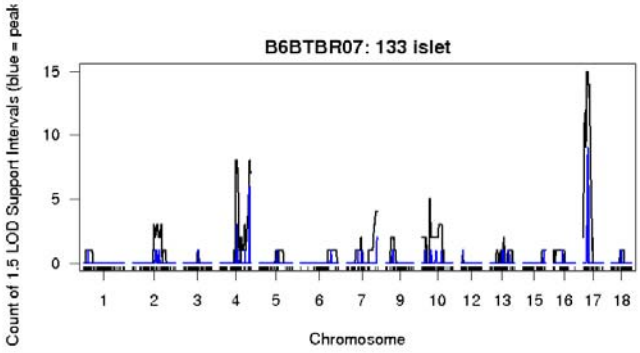

B6BTBR07: 133 islet
Mb
5 15 25 30 40 45 50 55 60 70 75 80 85 90
LOD
16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
cM
0 5 10 15 20 25 30 35 40 45 50 55 60

23

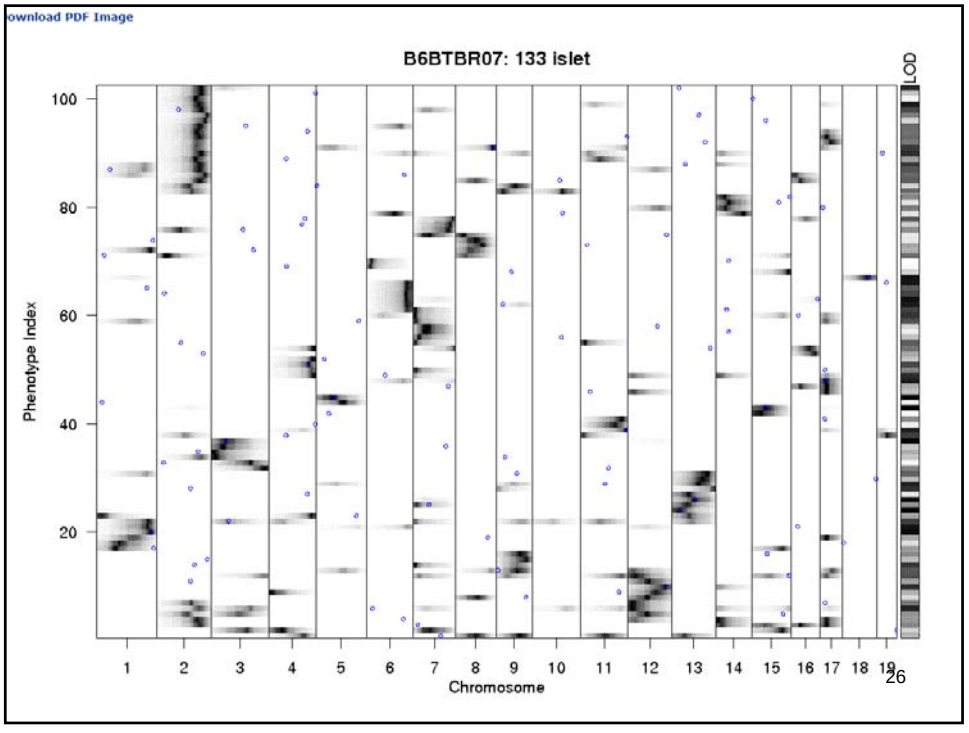

Download PDF Image
B6BTBR07: 133 islet
Mb
5 15 25 30 40 45 50 55 65 70 75 80 85 90
LOD
14 12 10 8 6 4 2 0
cM
0 10 20 30 40 50 60
Map position (cM)

24


B6BTBR07: 133 islet
LOD
1 2 3 4 5 6 7 9 10 12 13 15 16 17 18
Chromosome


B6BTBR07: 133 islet
Count of 1.5 LOD Support Intervals (blue = peak
15
10
5
0
1 2 3 4 5 6 7 9 10 12 13 15 16 17 18
Chromosome

25


ownload PDF Image
B6BTBR07: 133 islet
LOD
Phenotype Index
100
80
60
40
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Chromosome