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

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

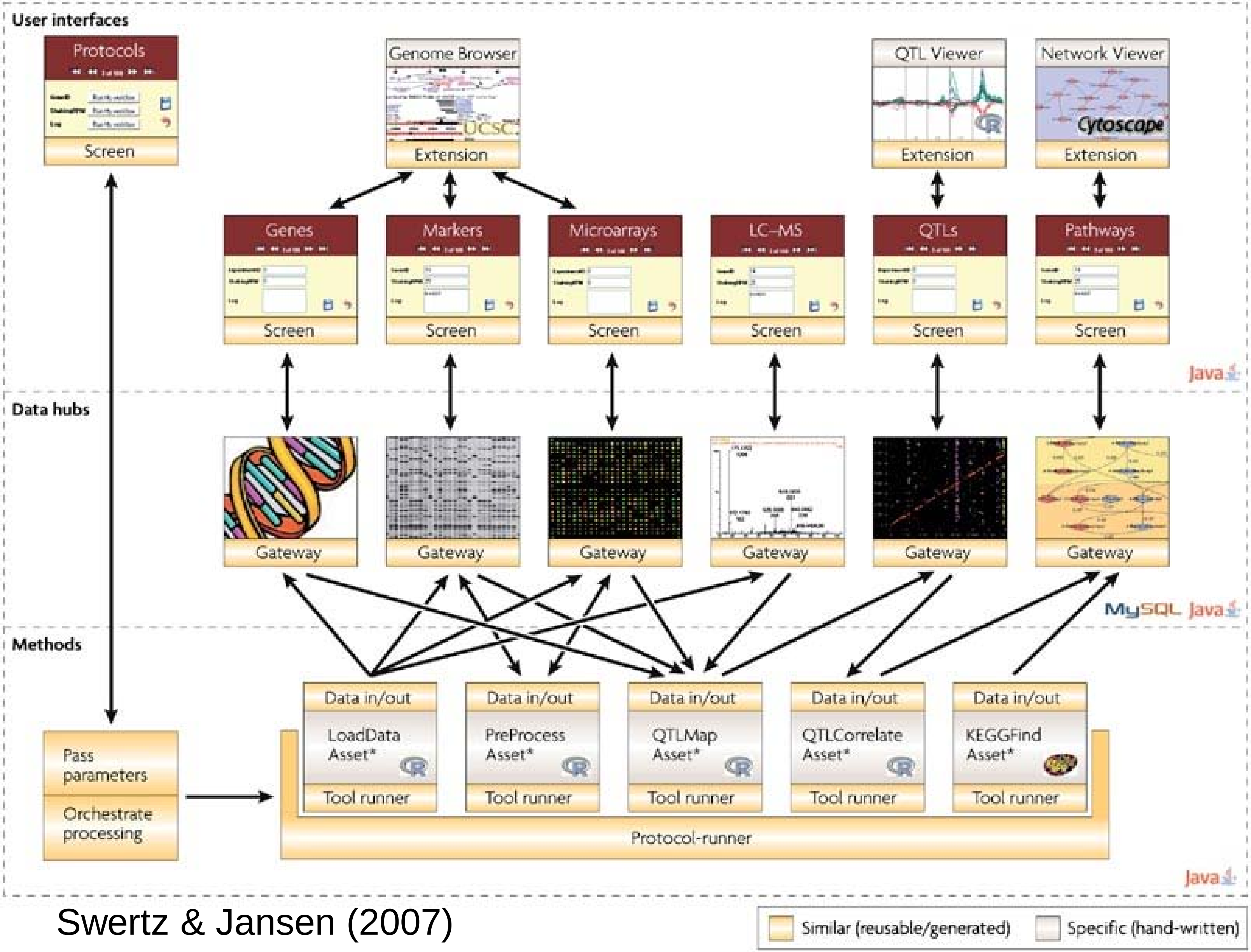
2- involve biologists early in development

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

Challenges Ahead:

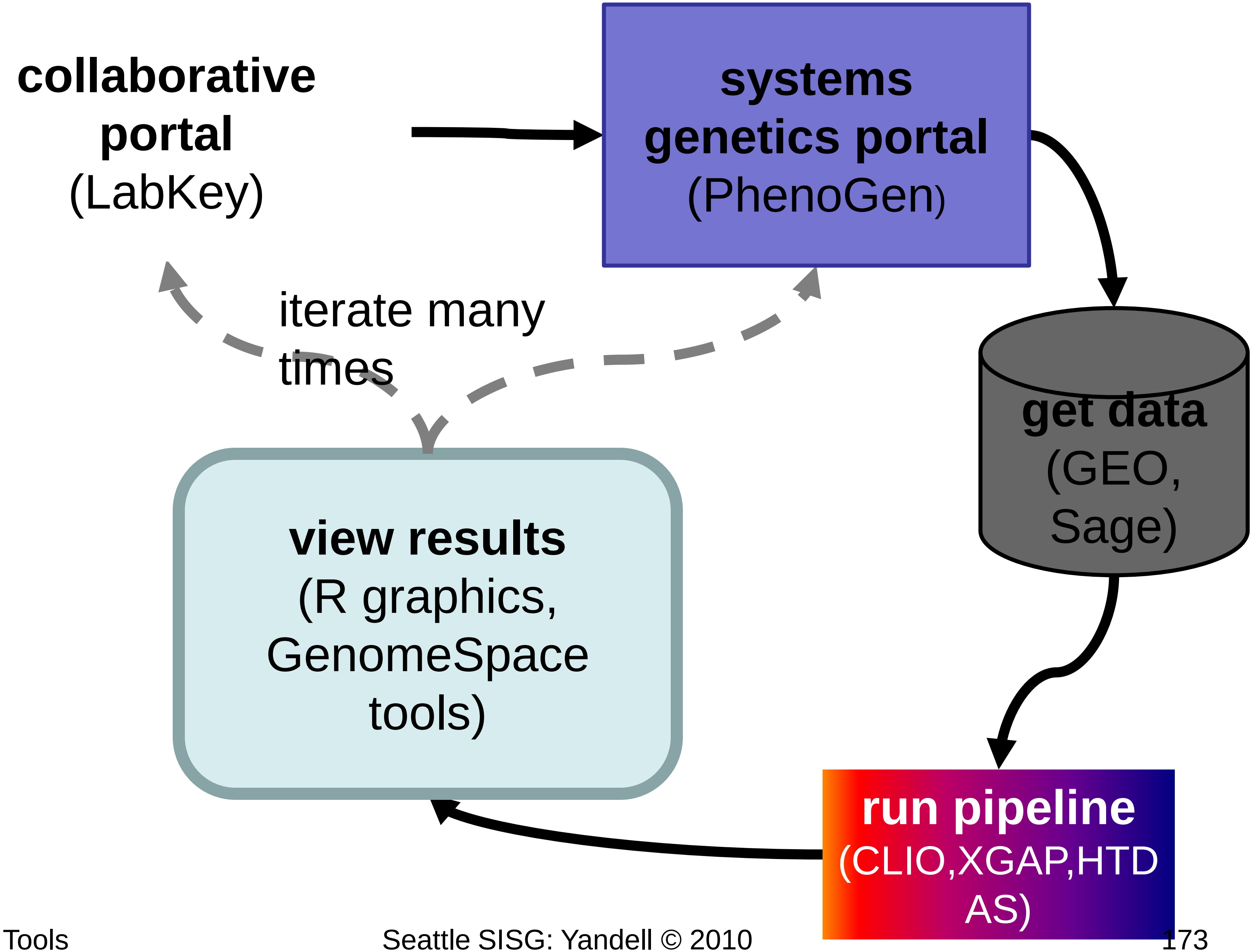
1- stitching together components as coherent system

2- ramping up to ever larger molecular datasets



Swertz & Jansen (2007)

Nature Reviews | Genetics



collaborative portal (LabKey)

systems genetics portal (PhenoGen)

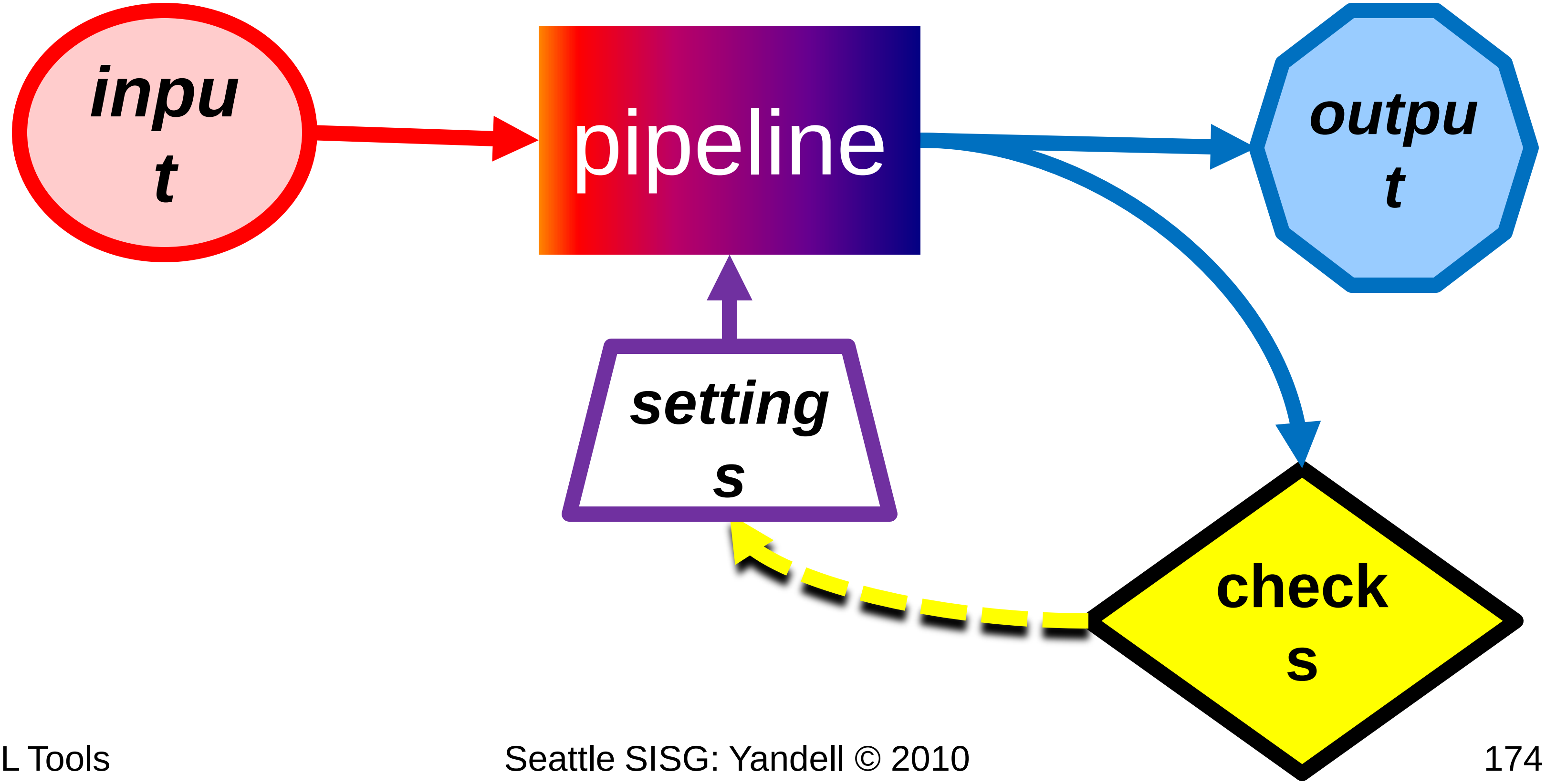
get data (GEO, Sage)

run pipeline (CLIO,XGAP,HTDAS)

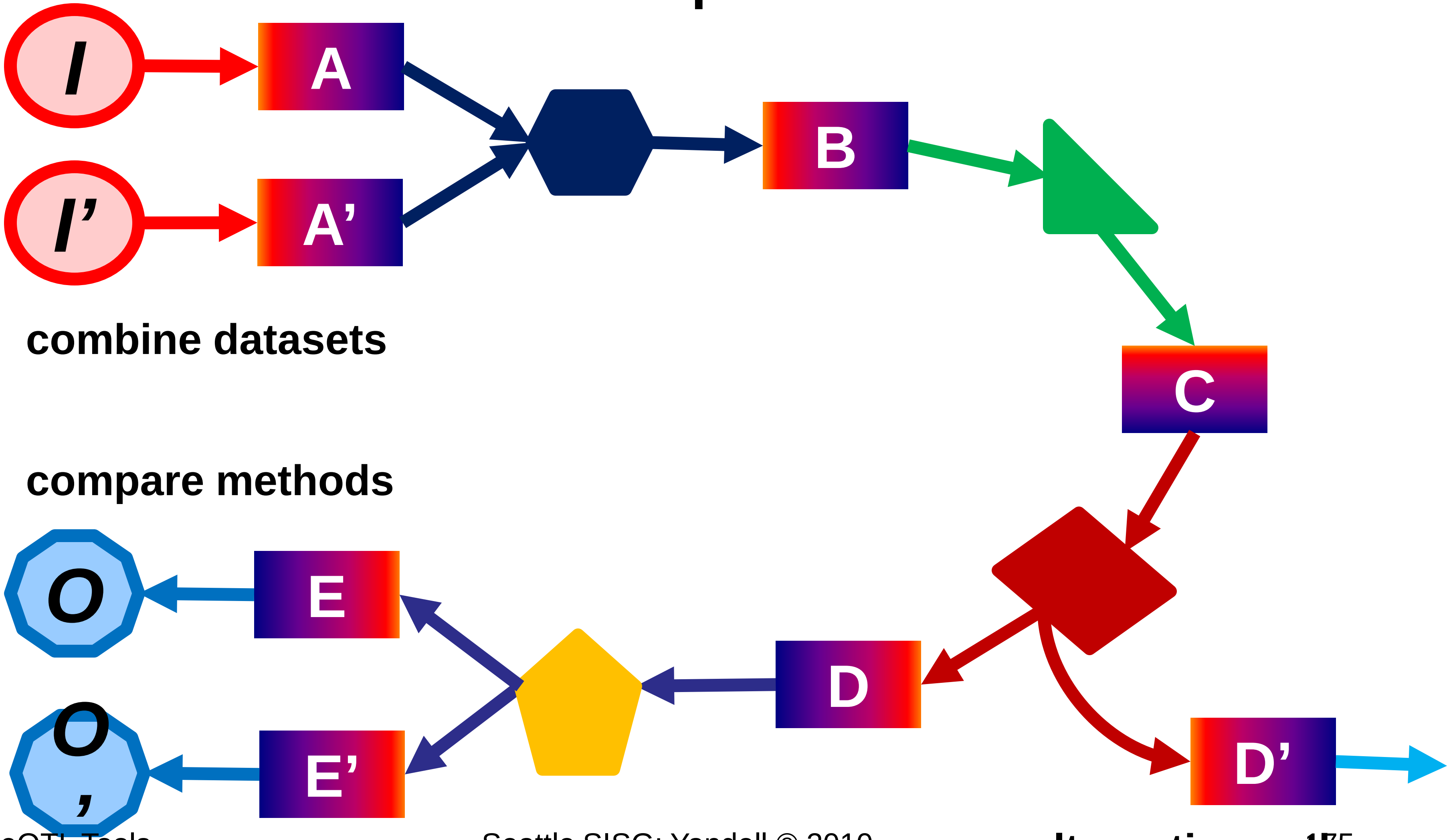
view results (R graphics, GenomeSpace tools)

iterate many times

analysis pipeline acts on objects (extends concept of GenePattern)



pipeline is composed of many steps



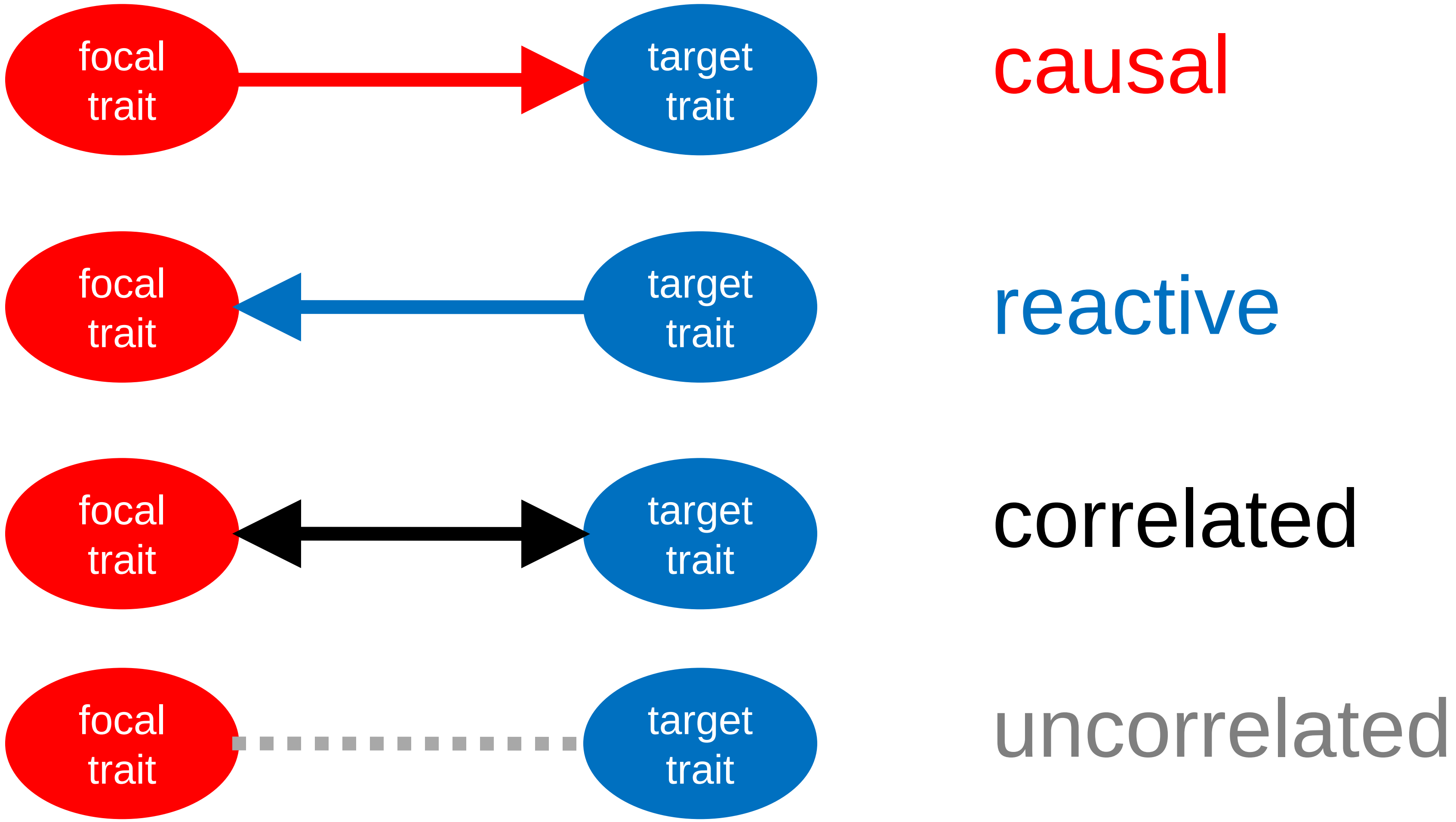
combine datasets

compare methods

alternative path

causal model selection choices

in context of larger, unknown network



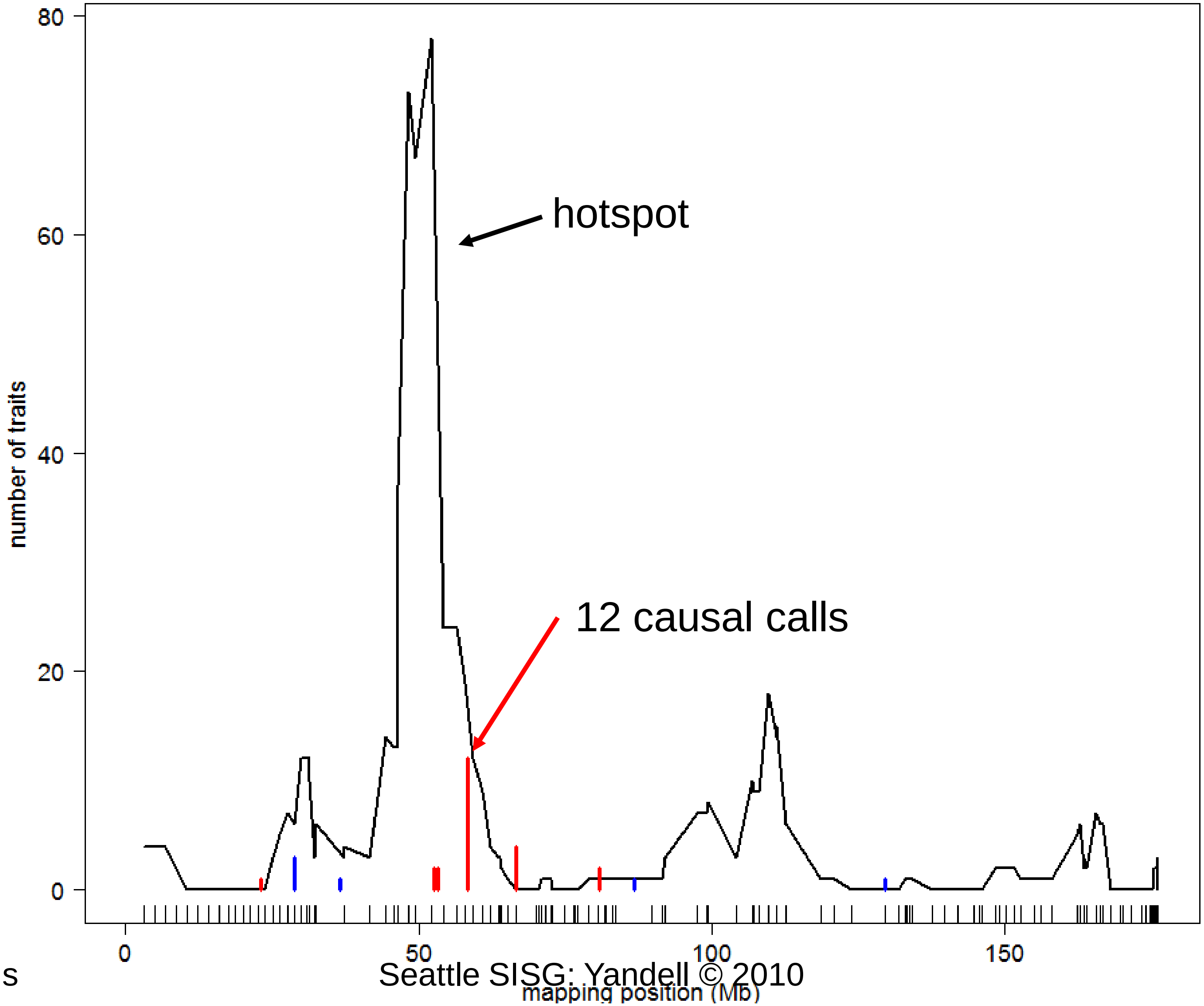
focal trait → target trait: causal

focal trait ← target trait: reactive

focal trait ↔ target trait: correlated

focal trait ···· target trait: uncorrelated

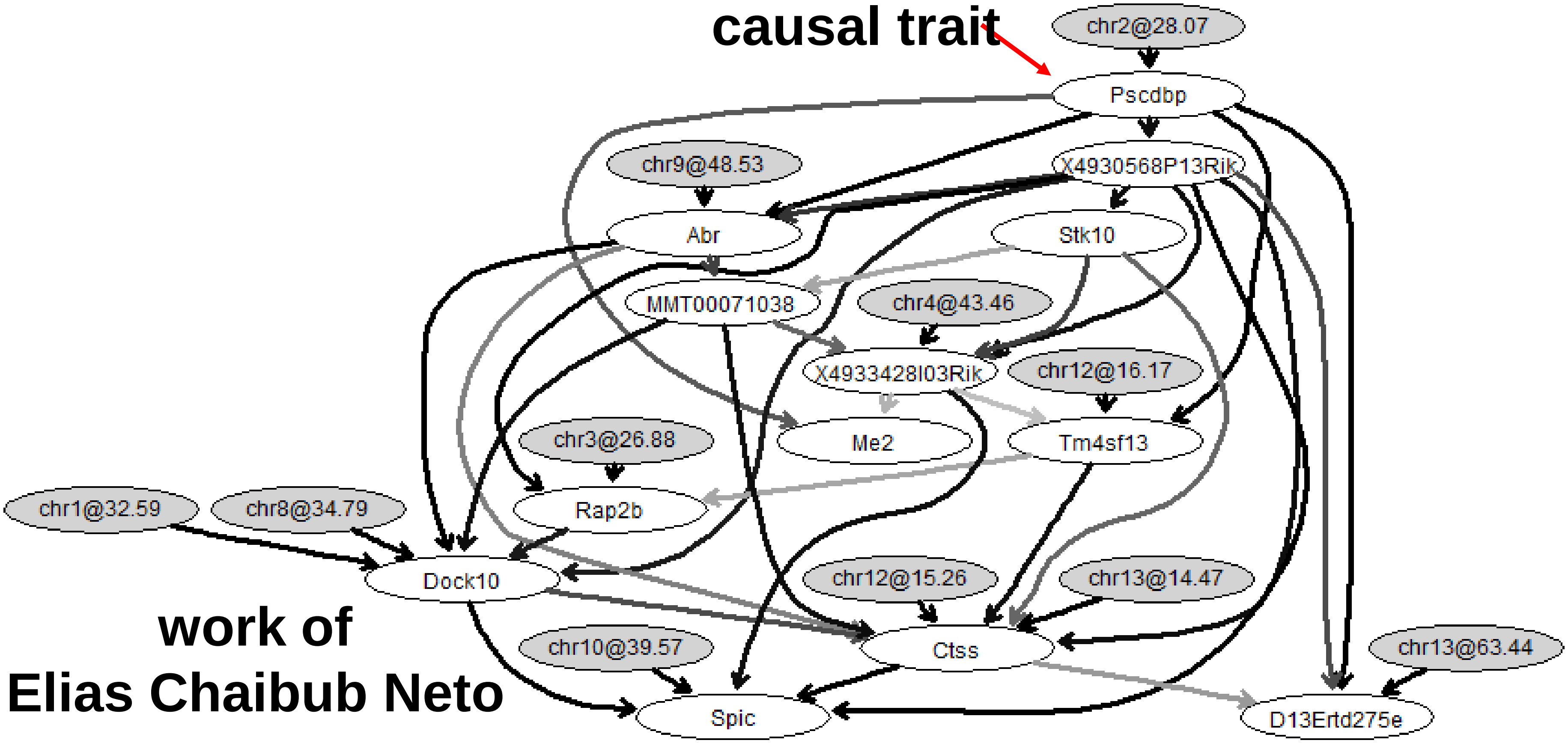
BxH ApoE-/- chr 2: causal architecture



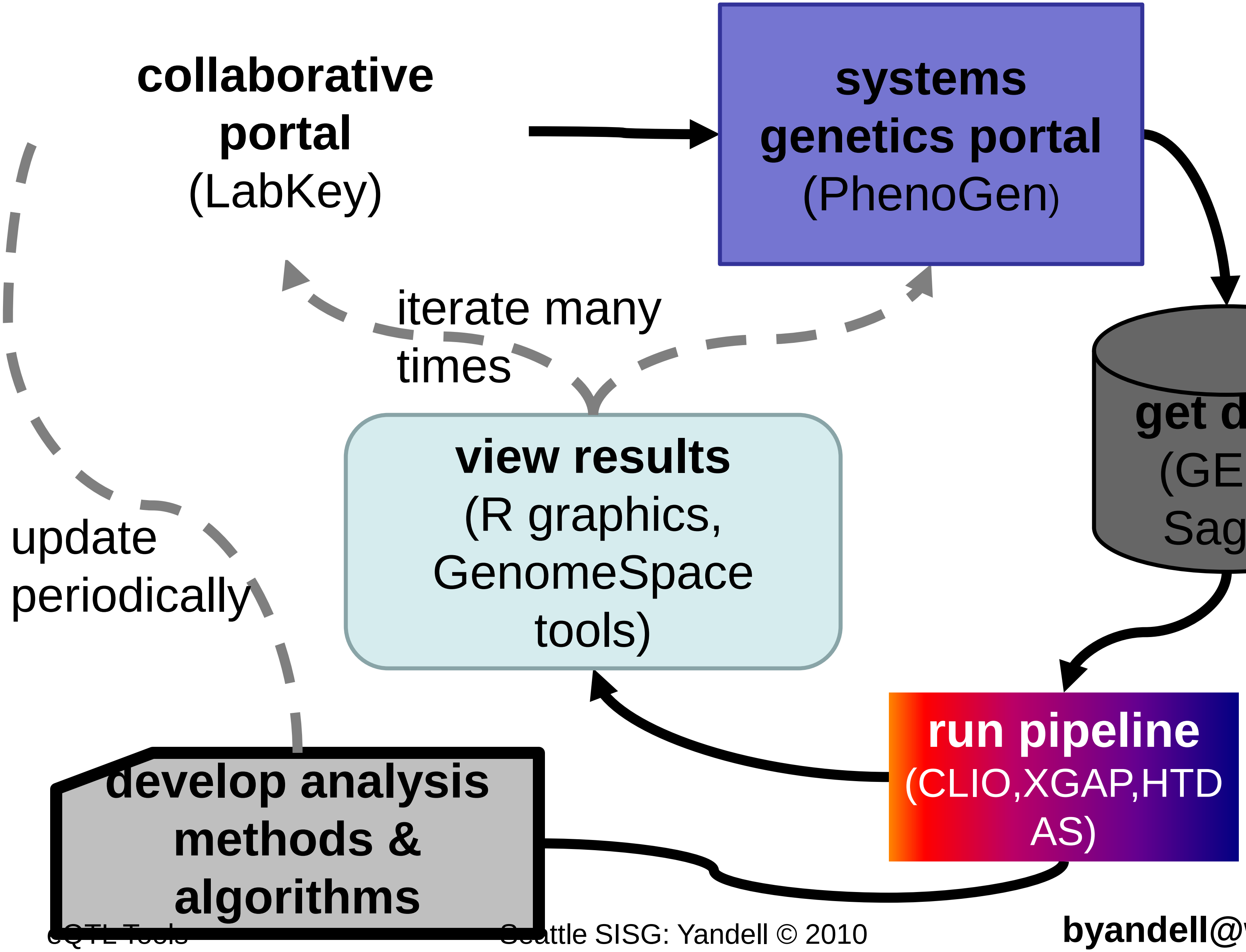
hotspot

12 causal calls

BxH ApoE-/- causal network for transcription factor Pscdbp



work of Elias Chaibub Neto



collaborative portal (LabKey)

systems genetics portal (PhenoGen)

get data (GEO, Sage)

run pipeline (CLIO,XGAP,HTDAS)

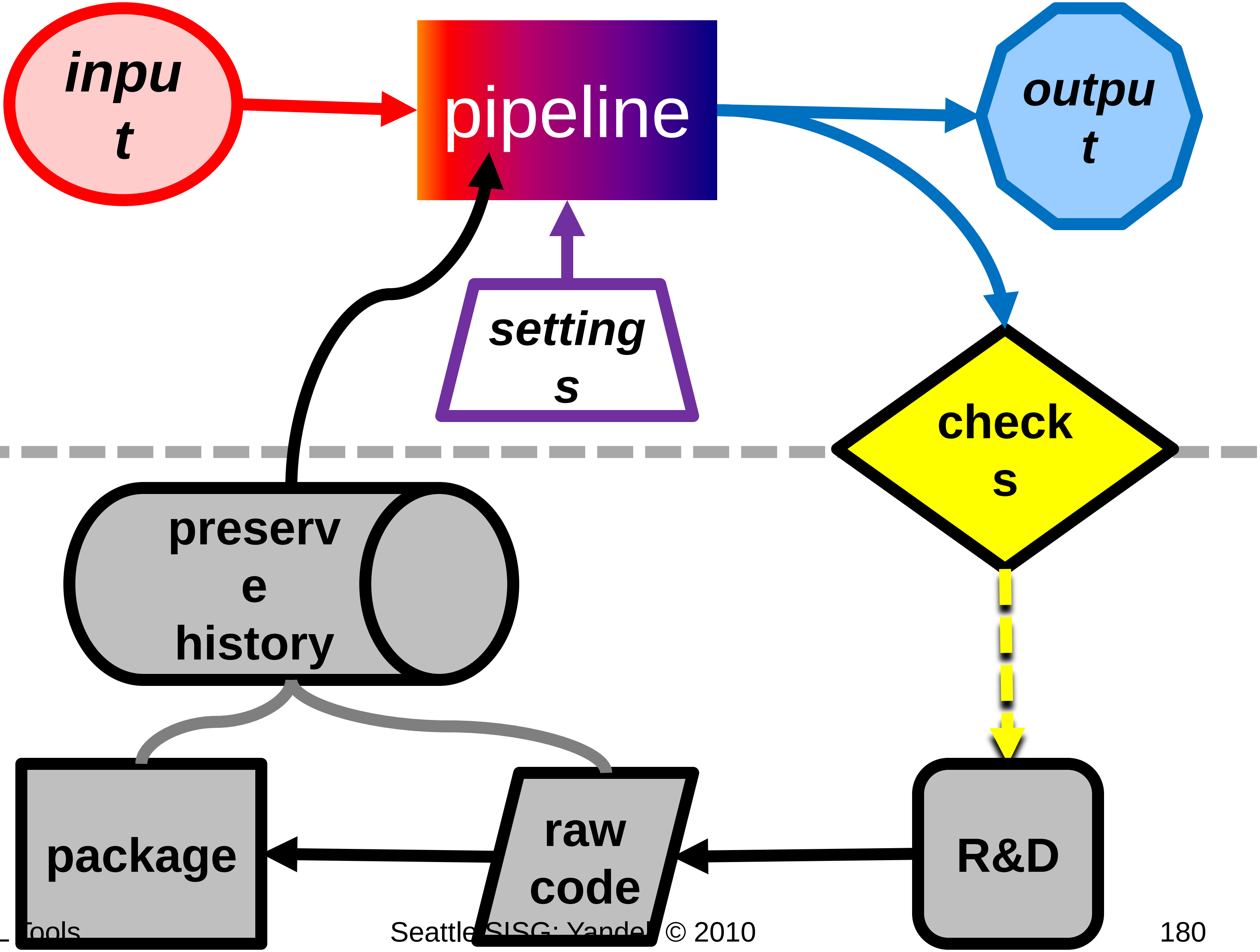
view results (R graphics, GenomeSpace tools)

develop analysis methods & algorithms

iterate many times

update periodically

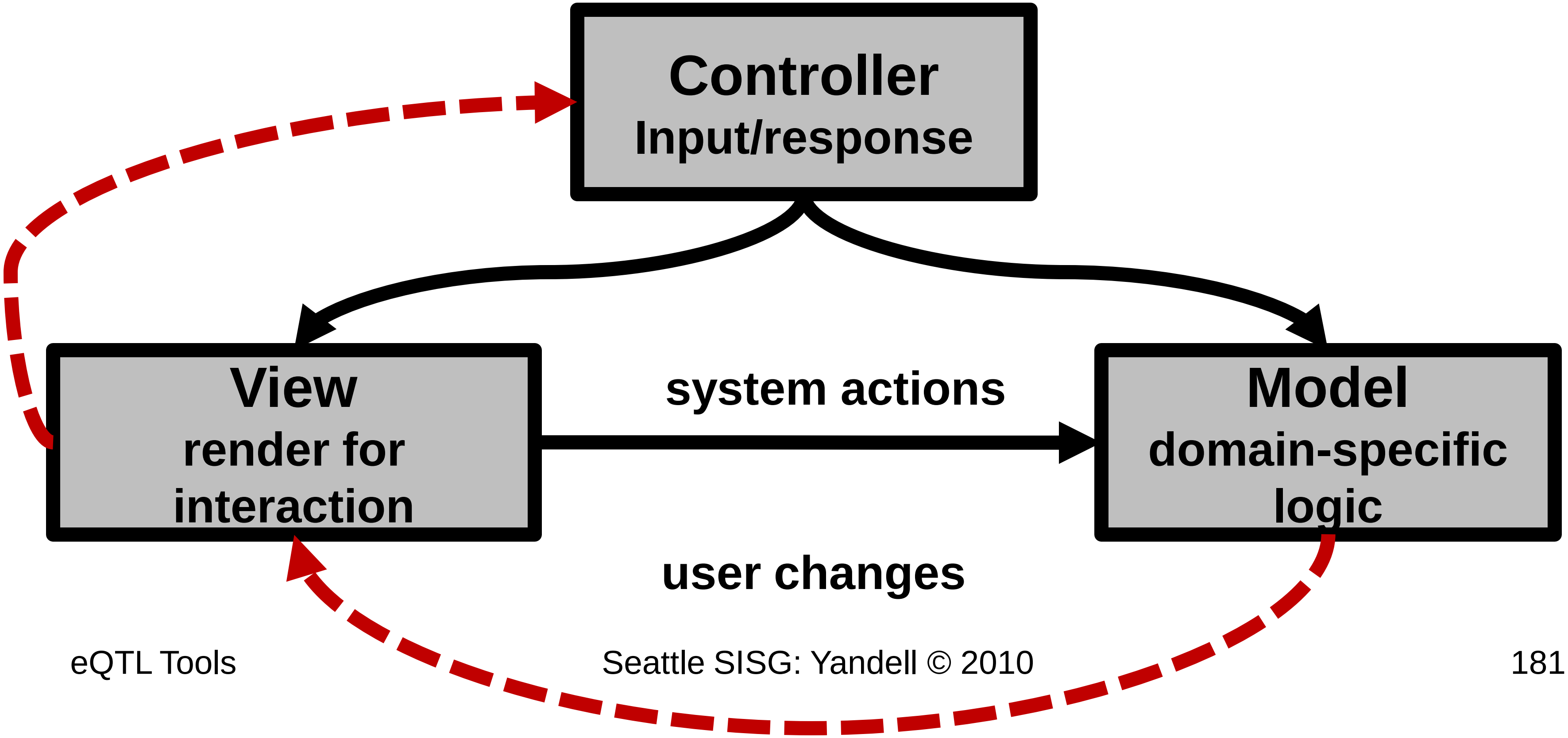
byandell@wisc.edu


inpu
t
pipeline
outpu
t
setting
s
check
s
preserv
e
history
package
raw
code
R&D

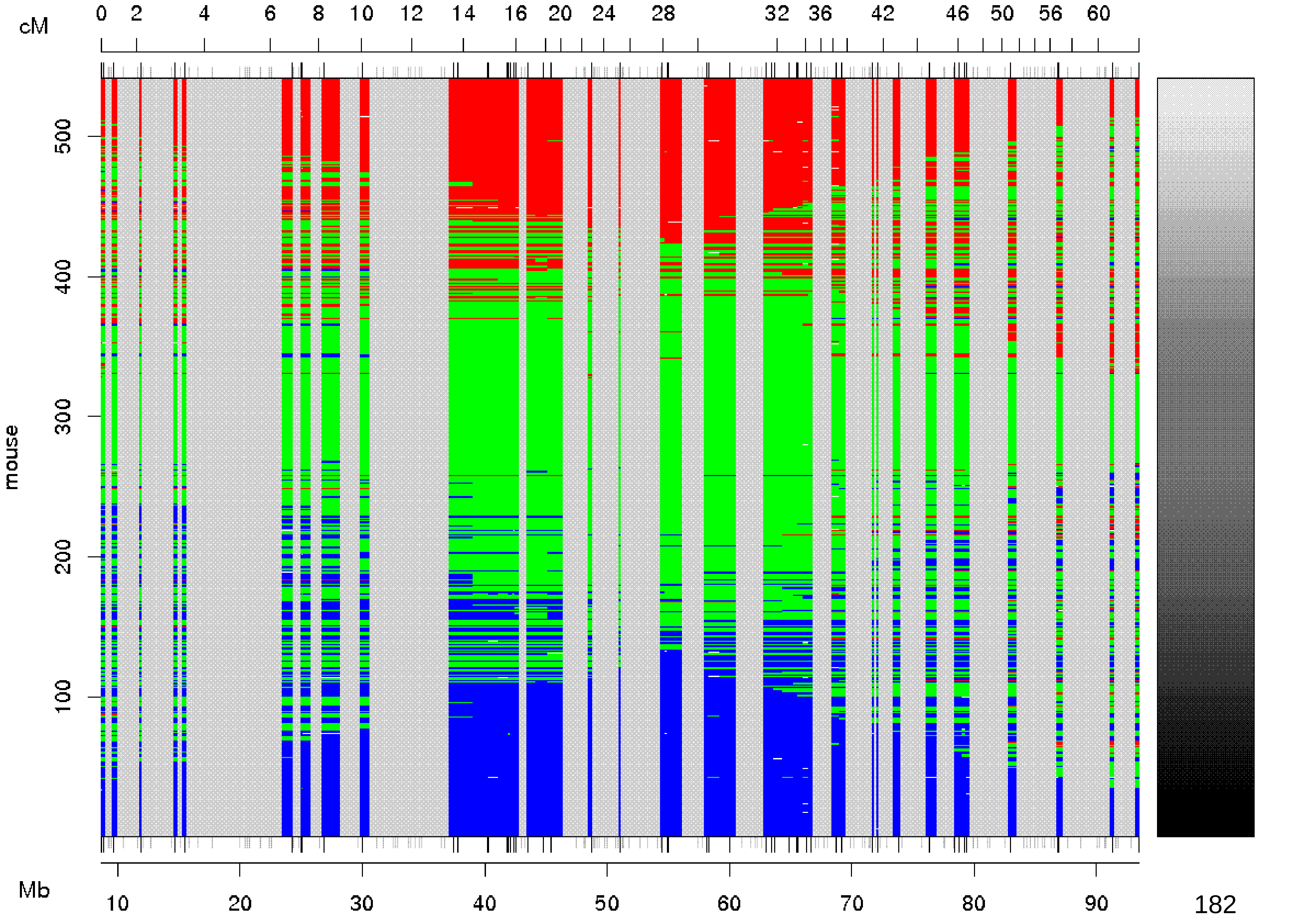
Model/View/Controller (MVC)

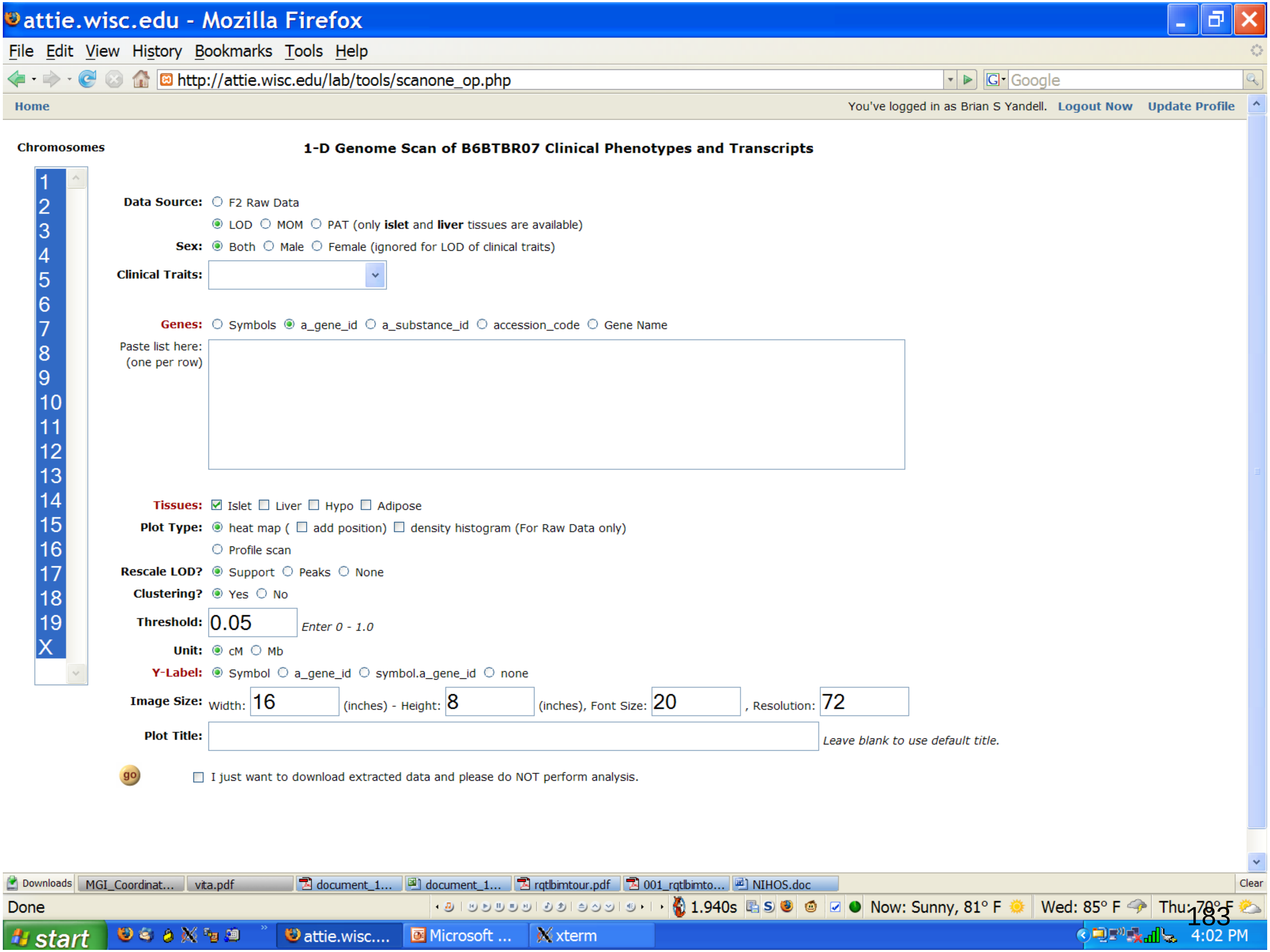
software architecture

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



Download PDF Image





1-D Genome Scan of B6BTBR07 Clinical Phenotypes and Transcripts

Chromosomes

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

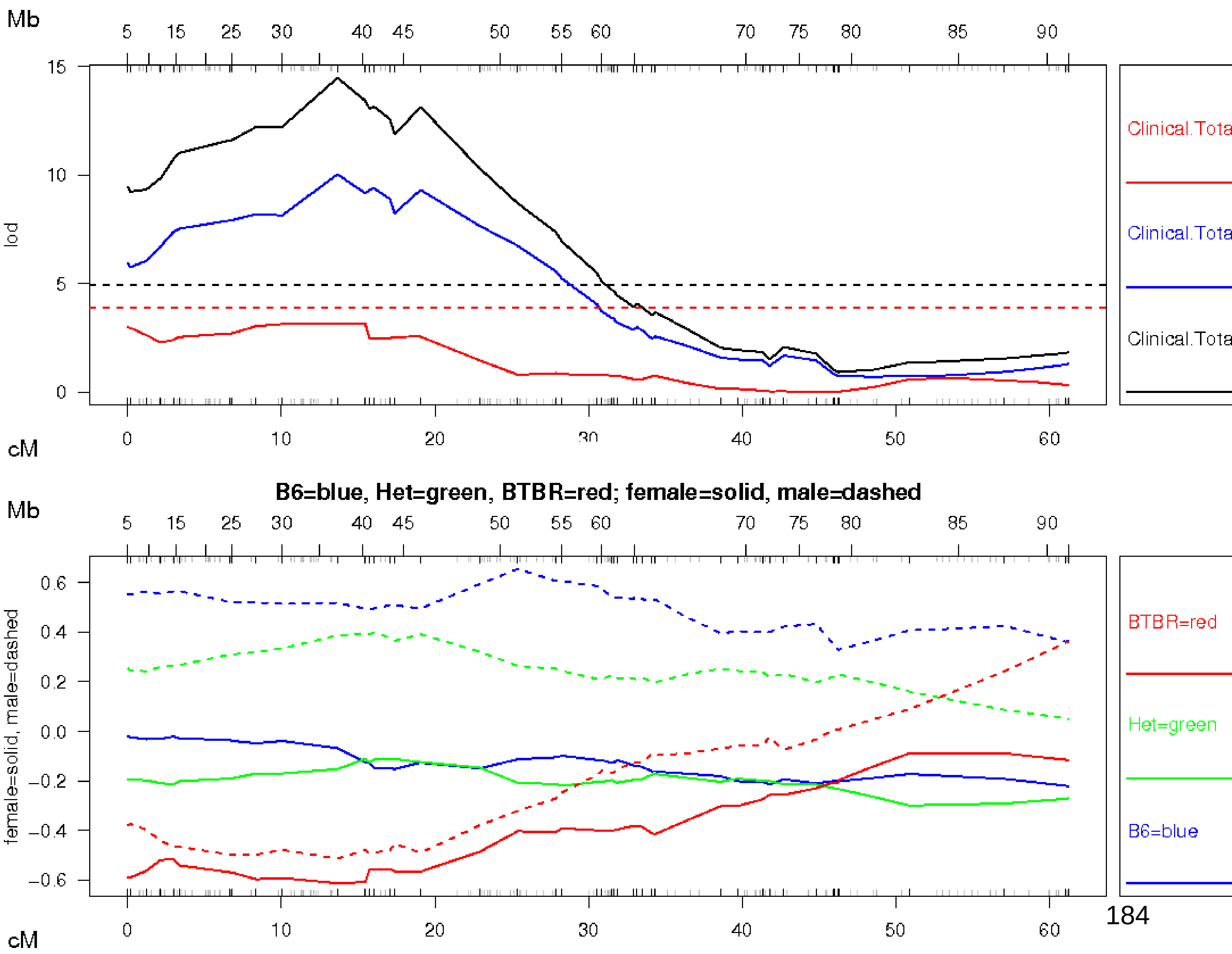
Y-Label: ● Symbol ○ a_gene_id ○ symbol.a_gene_id ○ none

Image Size: Width: 16 (inches) - Height: 8 (inches), Font Size: 20 , Resolution: 72

Plot Title: *Leave blank to use default title.*

go □ I just want to download extracted data and please do NOT perform analysis.

Download PDF Image



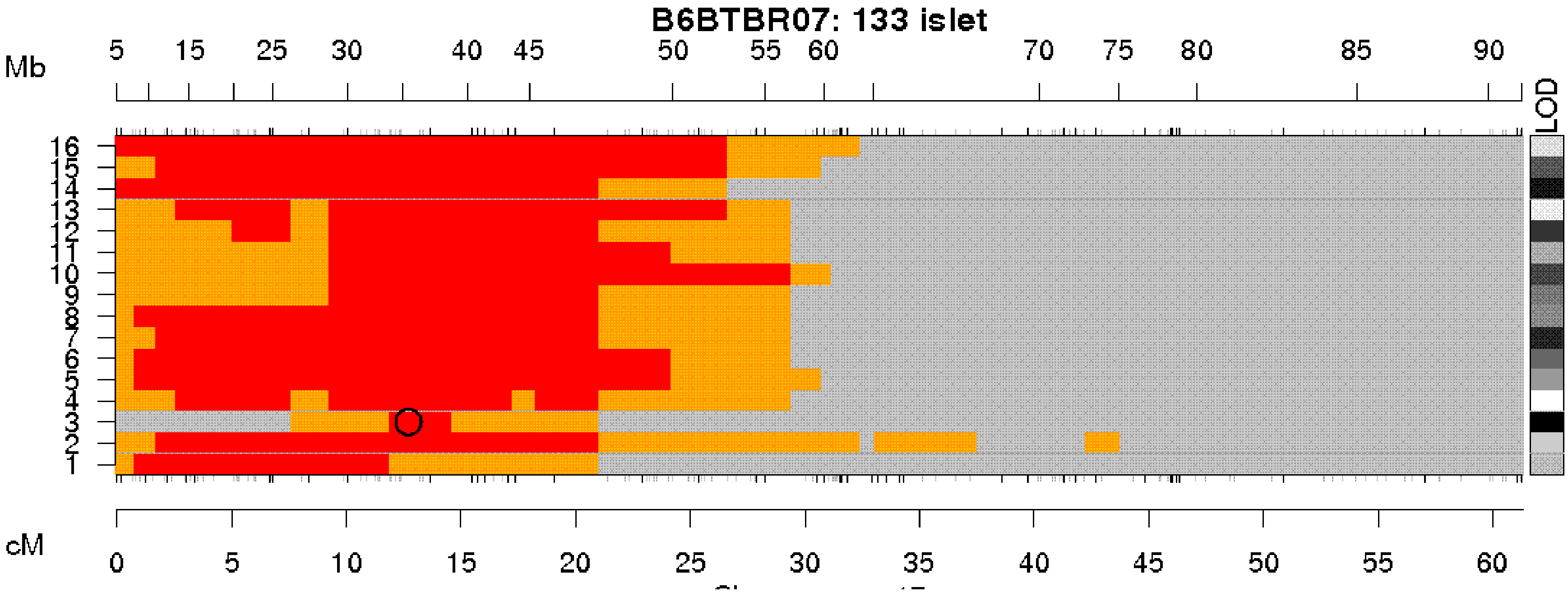
eQ

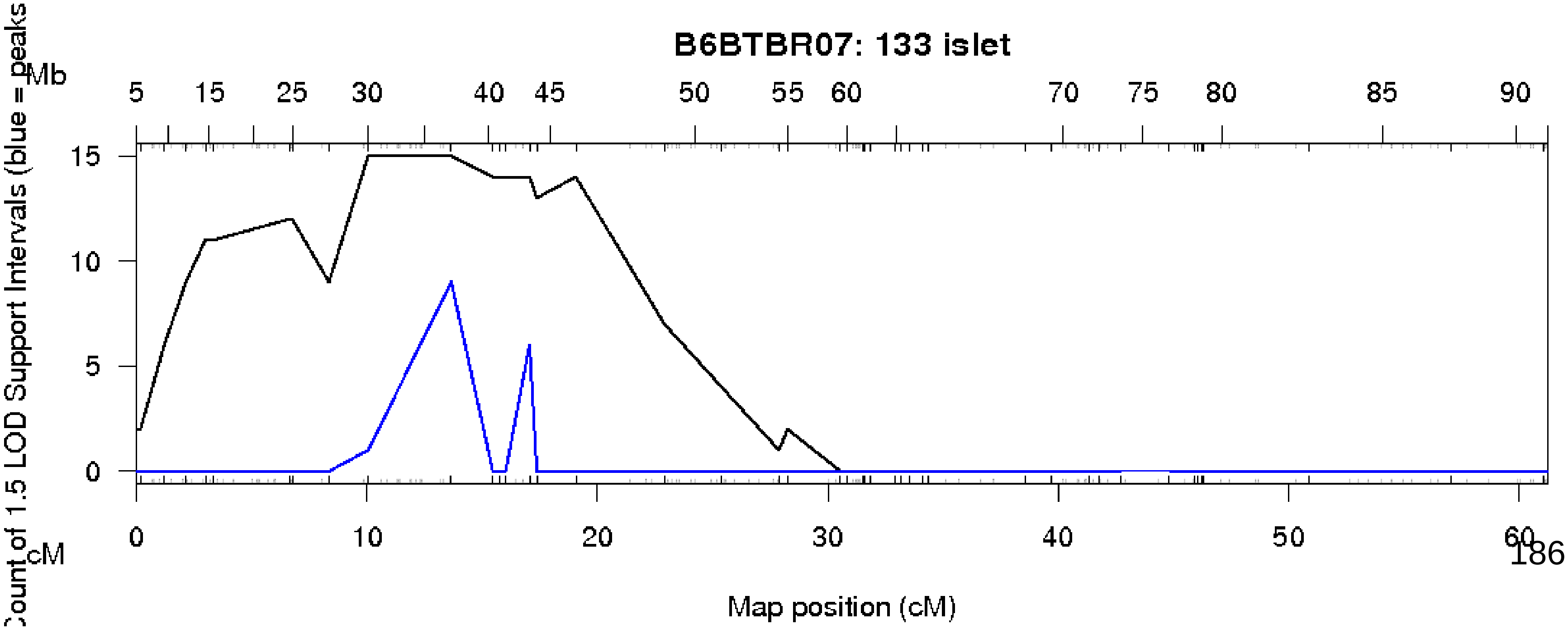
184

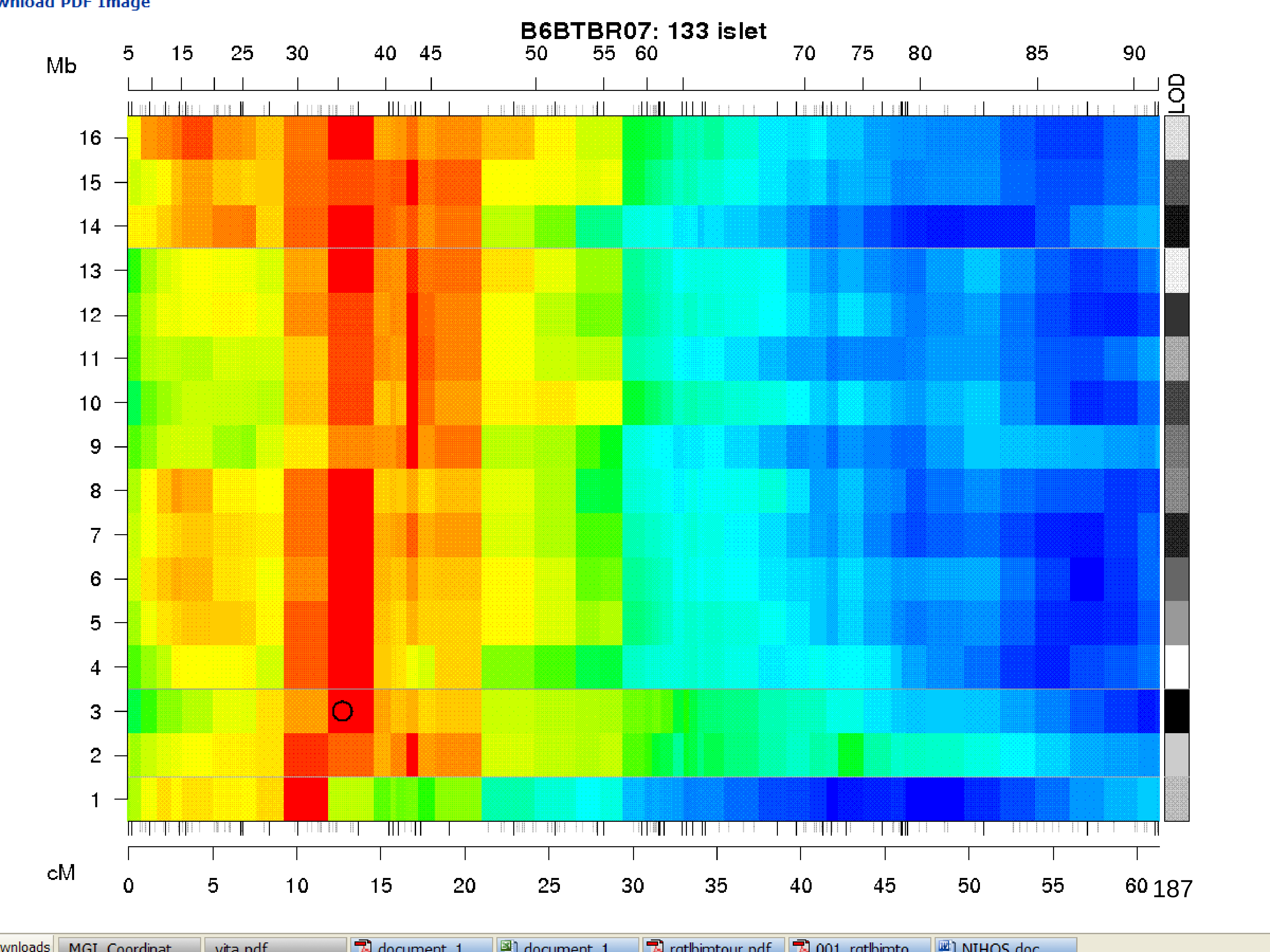
automated R script

```
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()
```

Download PDF Image



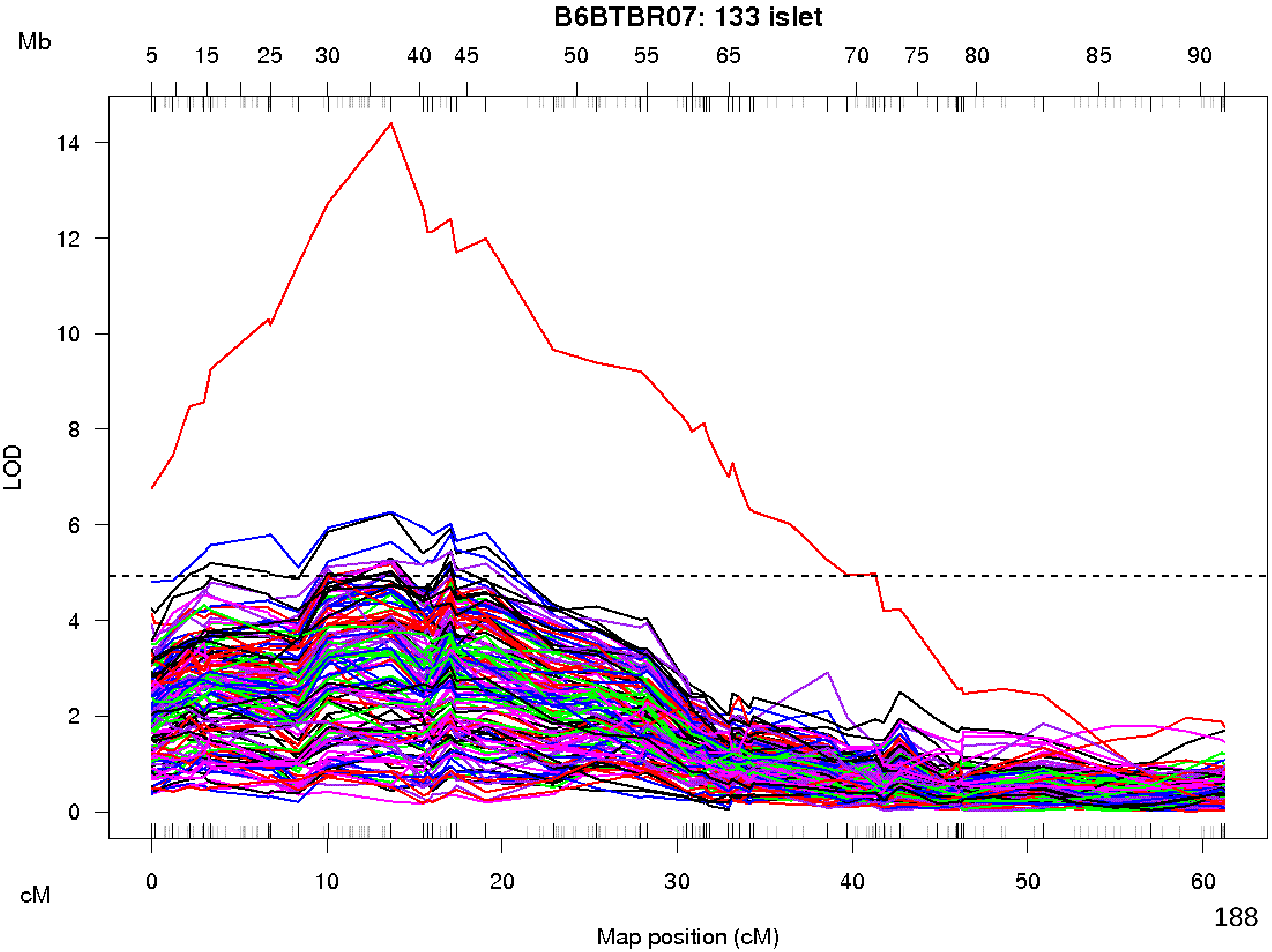




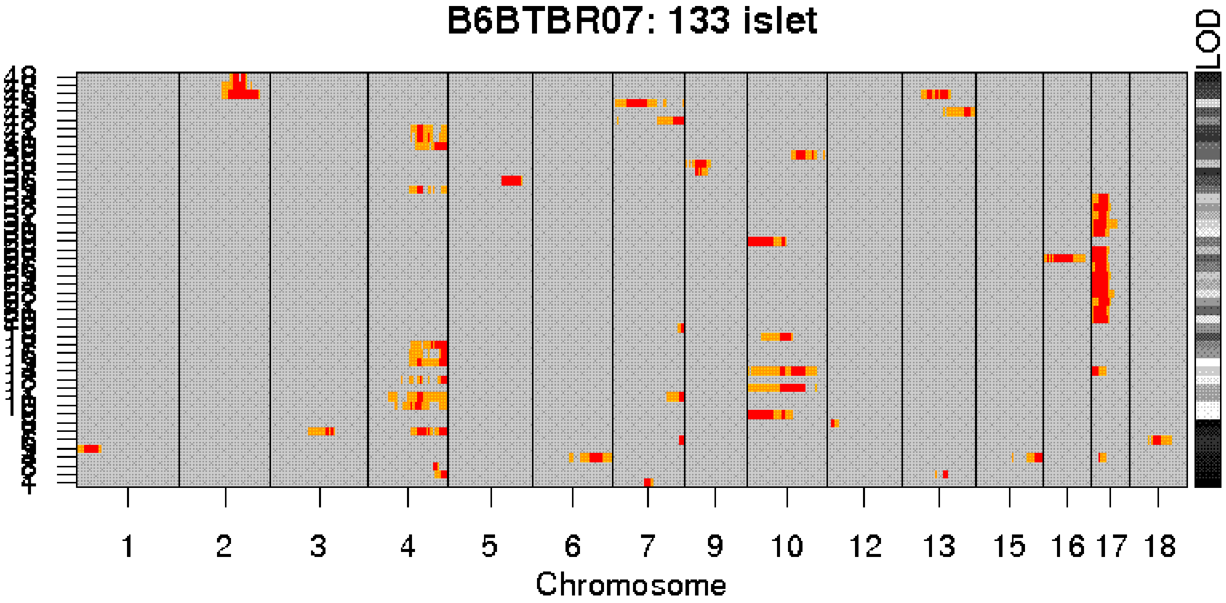
wnload PDF Image

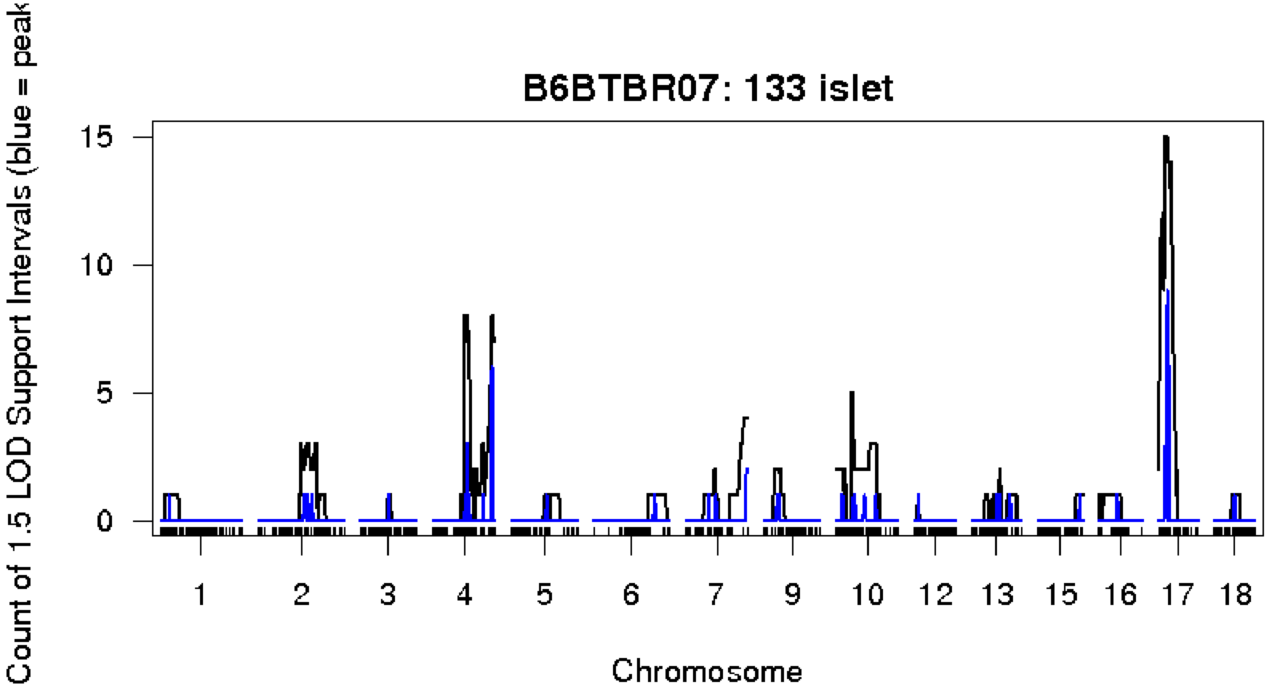
B6BTBR07: 133 islet

Download PDF Image

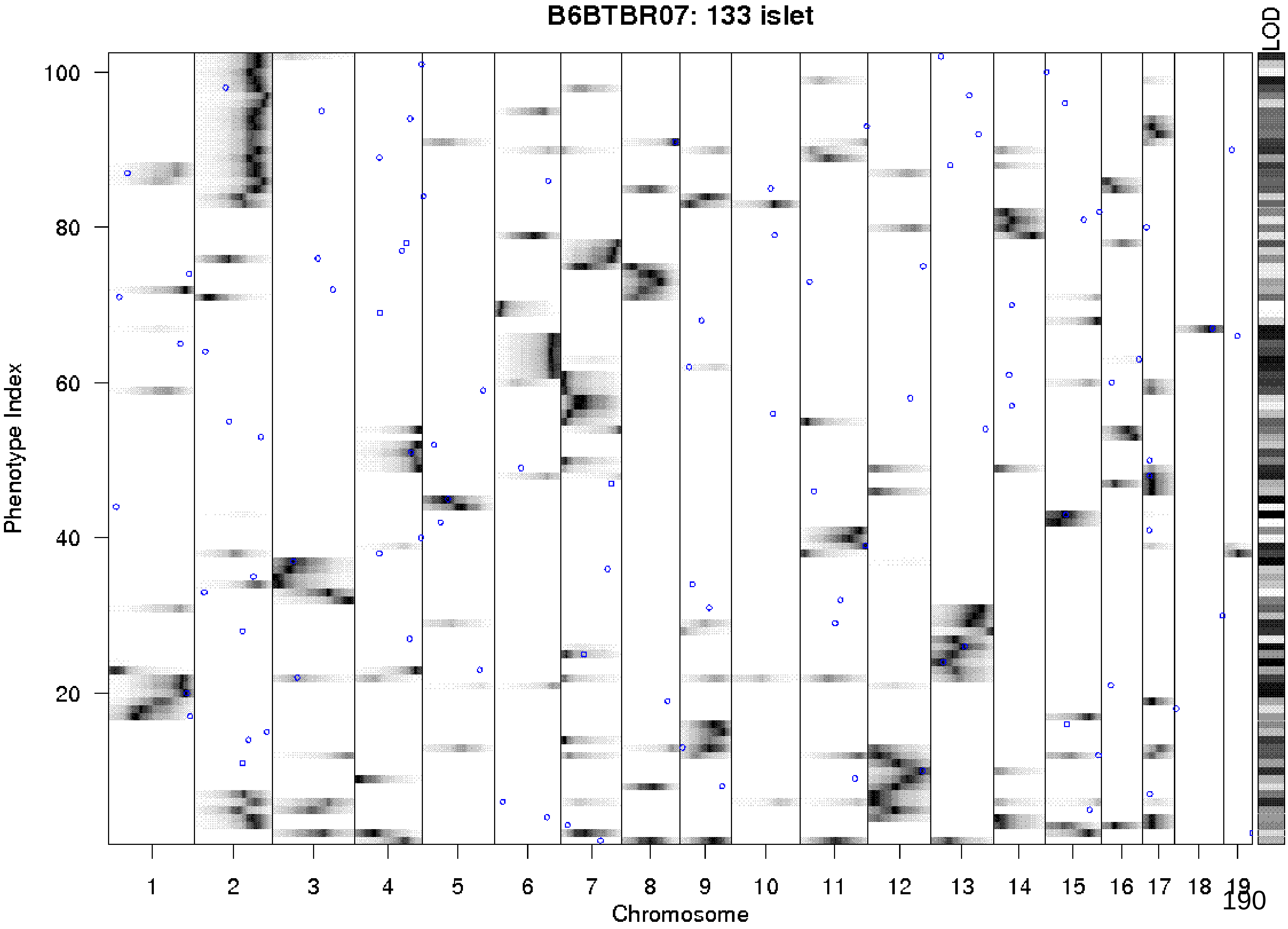


Download PDF Image





Download PDF Image


B6BTBR07: 133 islet
LOD
Phenotype Index
Chromosome