THE UNIVERSITY of WISCONSIN MADISON

Grid Computing to Study the Functions of Plant Genes

Candace Moore, Logan Johnson, Nathan Miller & Edgar Spalding
Department of Botany

Phyto morph

www.botany.wisc.edu/spalding.htm

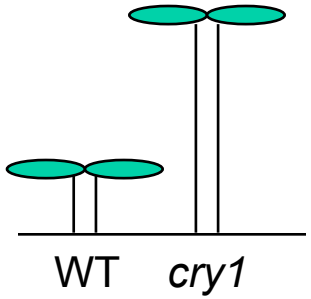
A Bit of Background

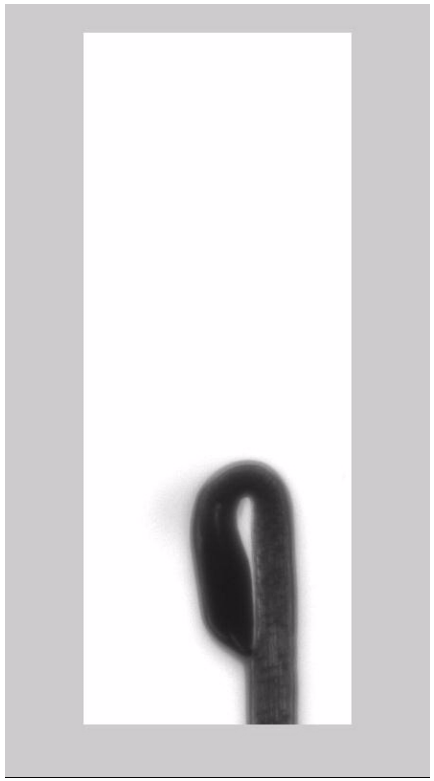
- A major goal in biology is to learn the function of each gene in an organism.
- A proven approach is to compare the behaviors of individuals possessing different versions of that gene.
- Organisms have on the order of 10^4 genes so that makes for a lot of comparisons.

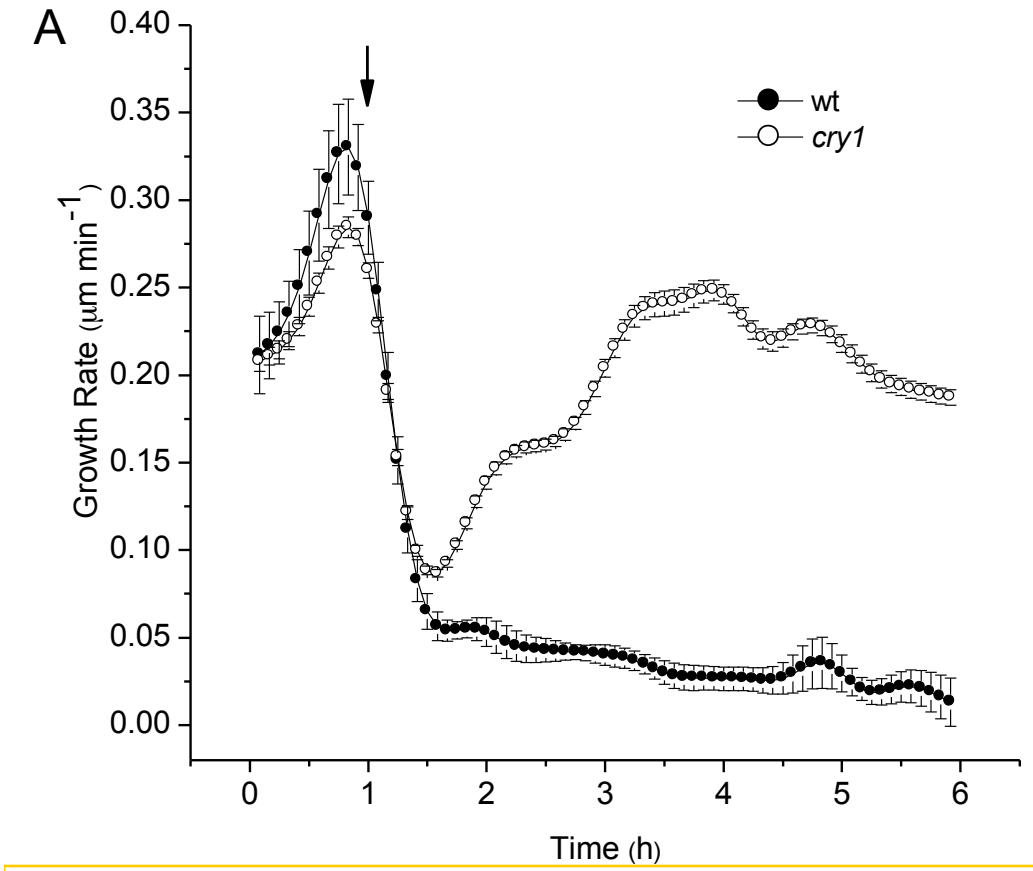
24,999 genes

25,000 genes

Hypocotyl Growth Inhibition Induced by Blue Light

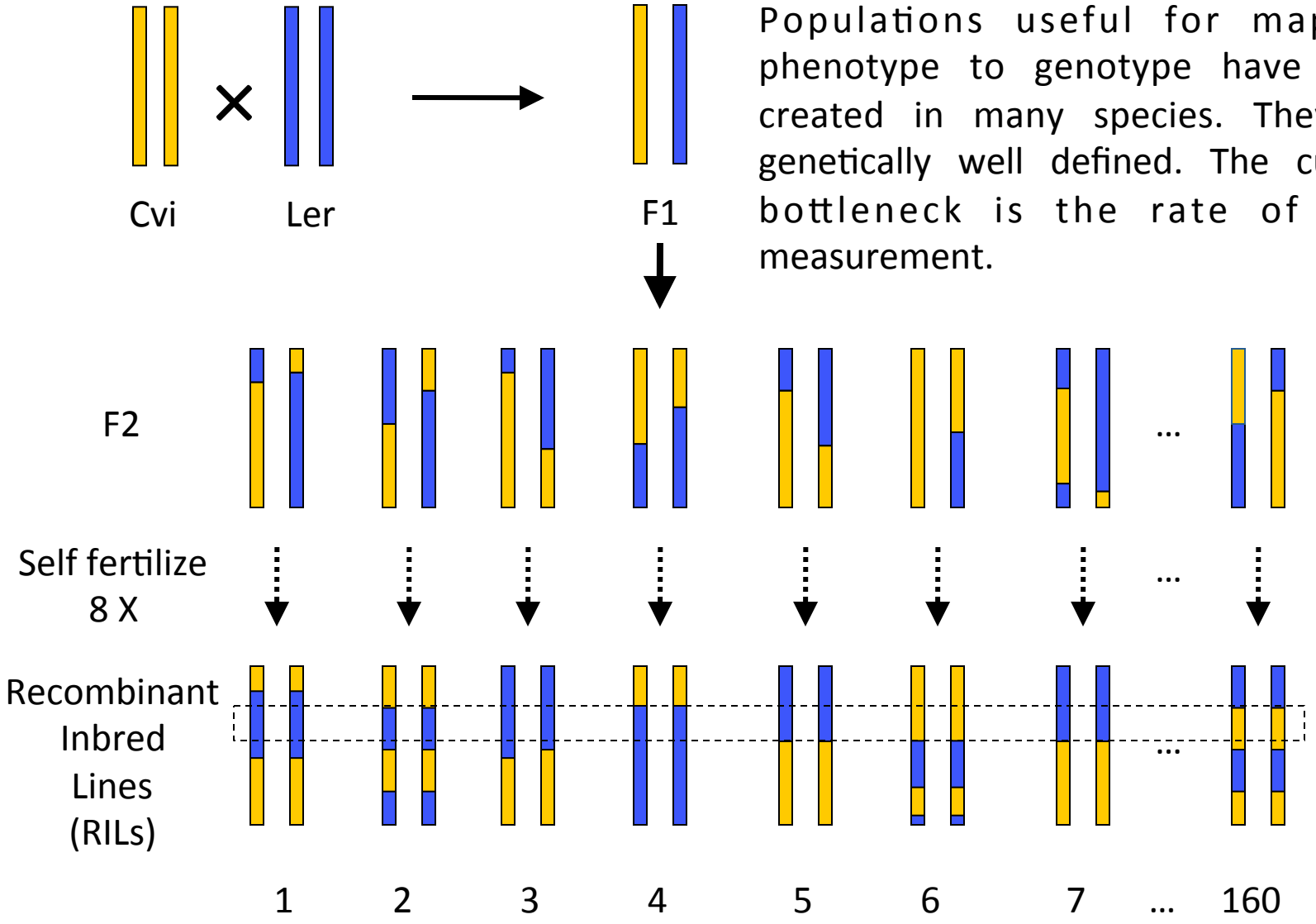






We developed image processing algorithms for measuring seedling growth and development in order to quantify effects of genetic differences (i.e. phenotypes of mutants) in space and time. Our purpose was to answer questions about one or two genes.

Why not / How to scale up to the whole genome level?



Populations useful for mapping phenotype to genotype have been created in many species. They are genetically well defined. The current bottleneck is the rate of trait measurement.

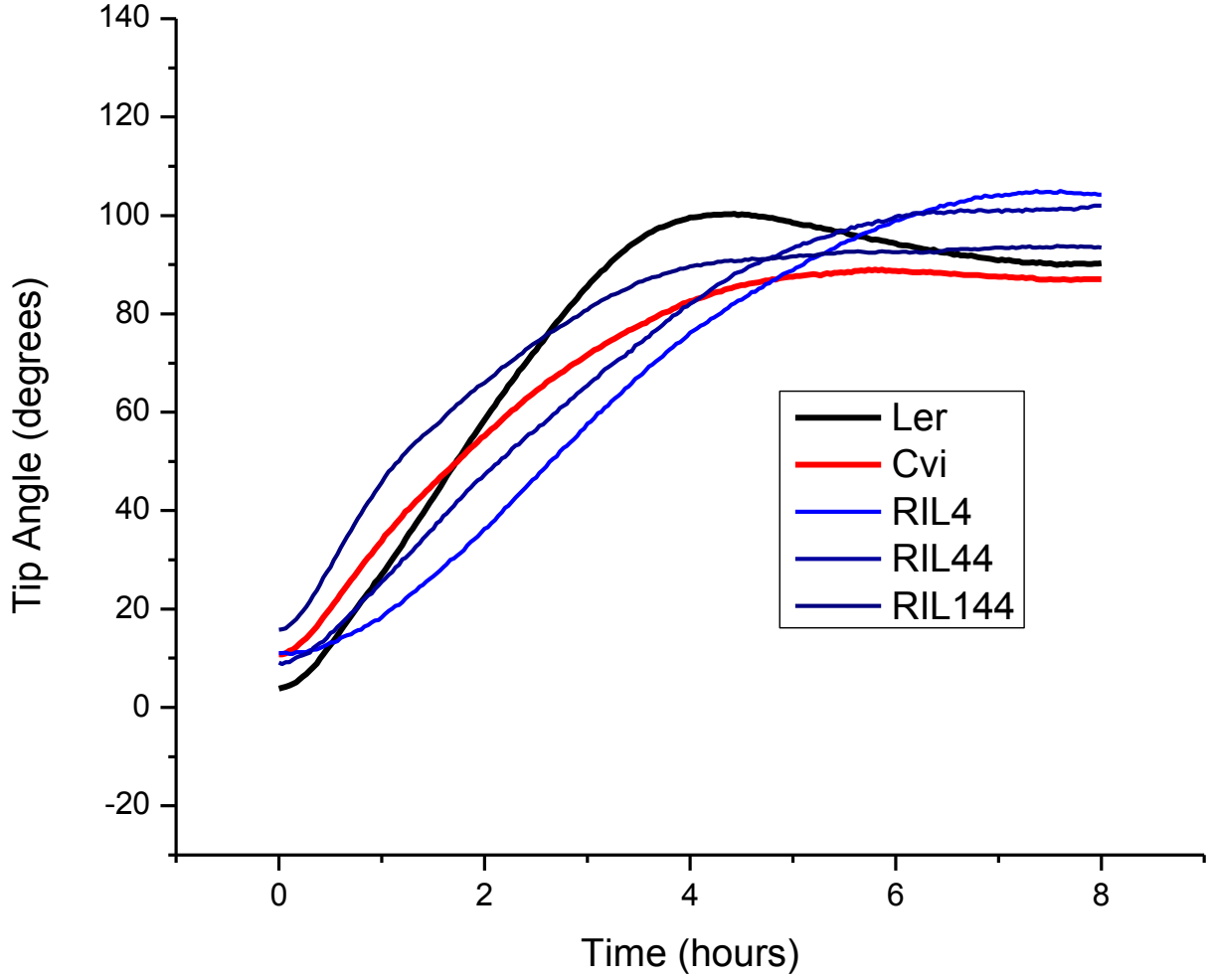
Alonso-Blanco *et al. The Plant Journal* **14**, 259-257 (1998)

Let’s make the process High Throughput
Switch to Root Gravitropism…

because we could study with high resolution, high accuracy, and high throughput

Machine Vision to Study Natural Genetic Variation

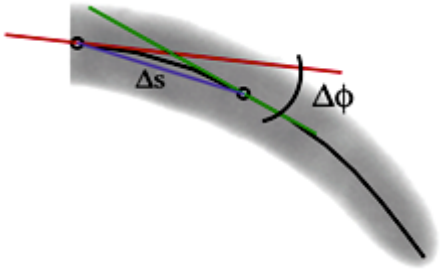
160 L*er* X CVI recombinant inbred lines for QTL mapping



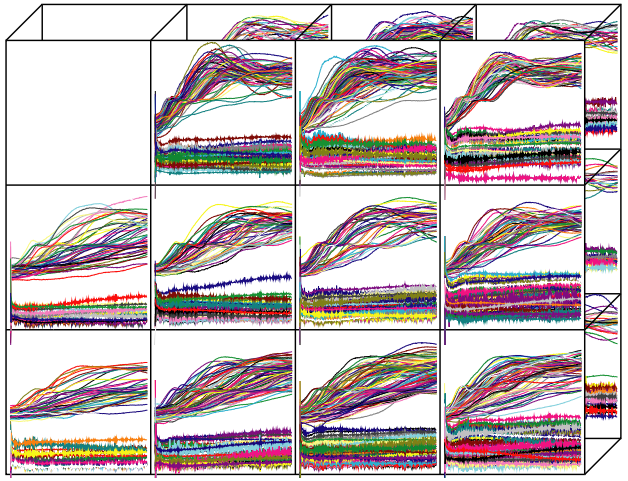
Our Automated Workflow

automated image acquisition generates movies

algorithmic feature extraction and quantification operates on each object…



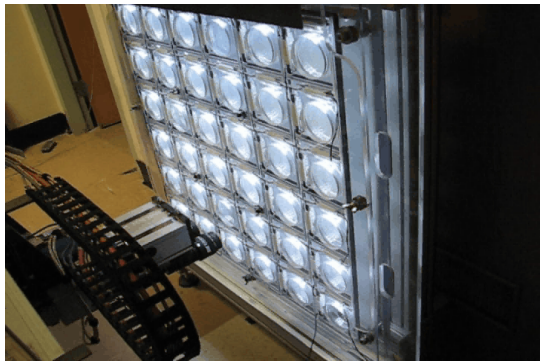
…resulting in large, high dimension data sets…



data mining

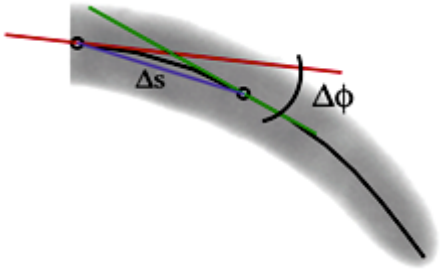
test new hypotheses

Our Automated Workflow

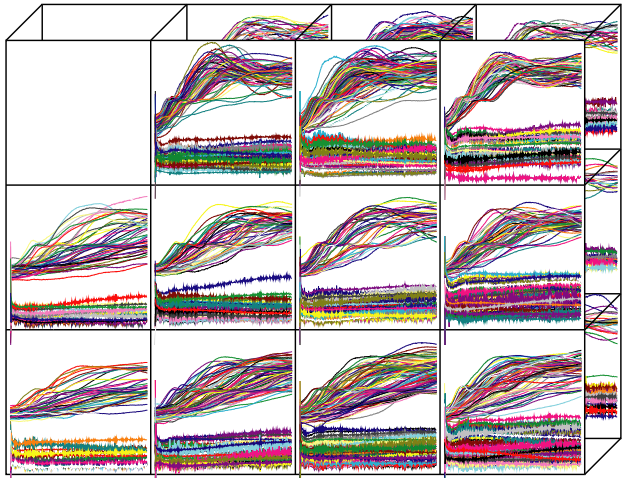


automated image acquisition generates movies

algorithmic feature extraction and quantification operates on each object…



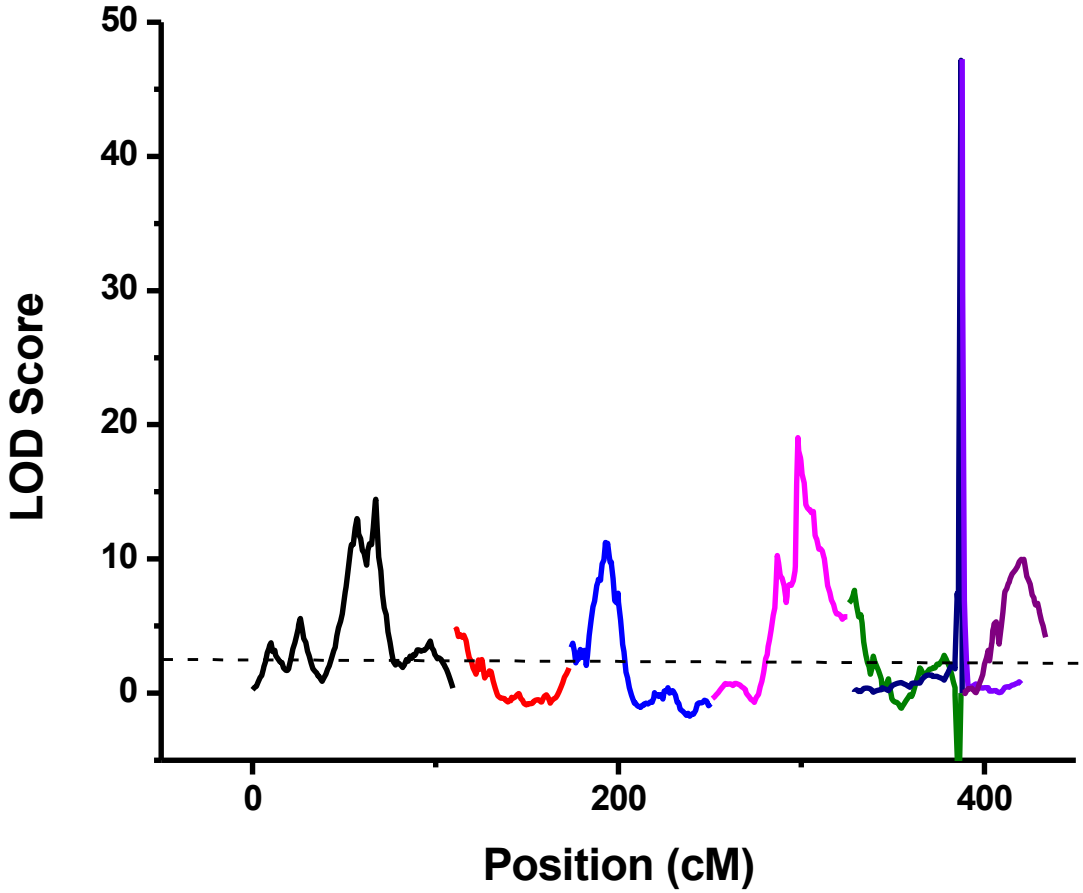
…resulting in large, high dimension data sets…

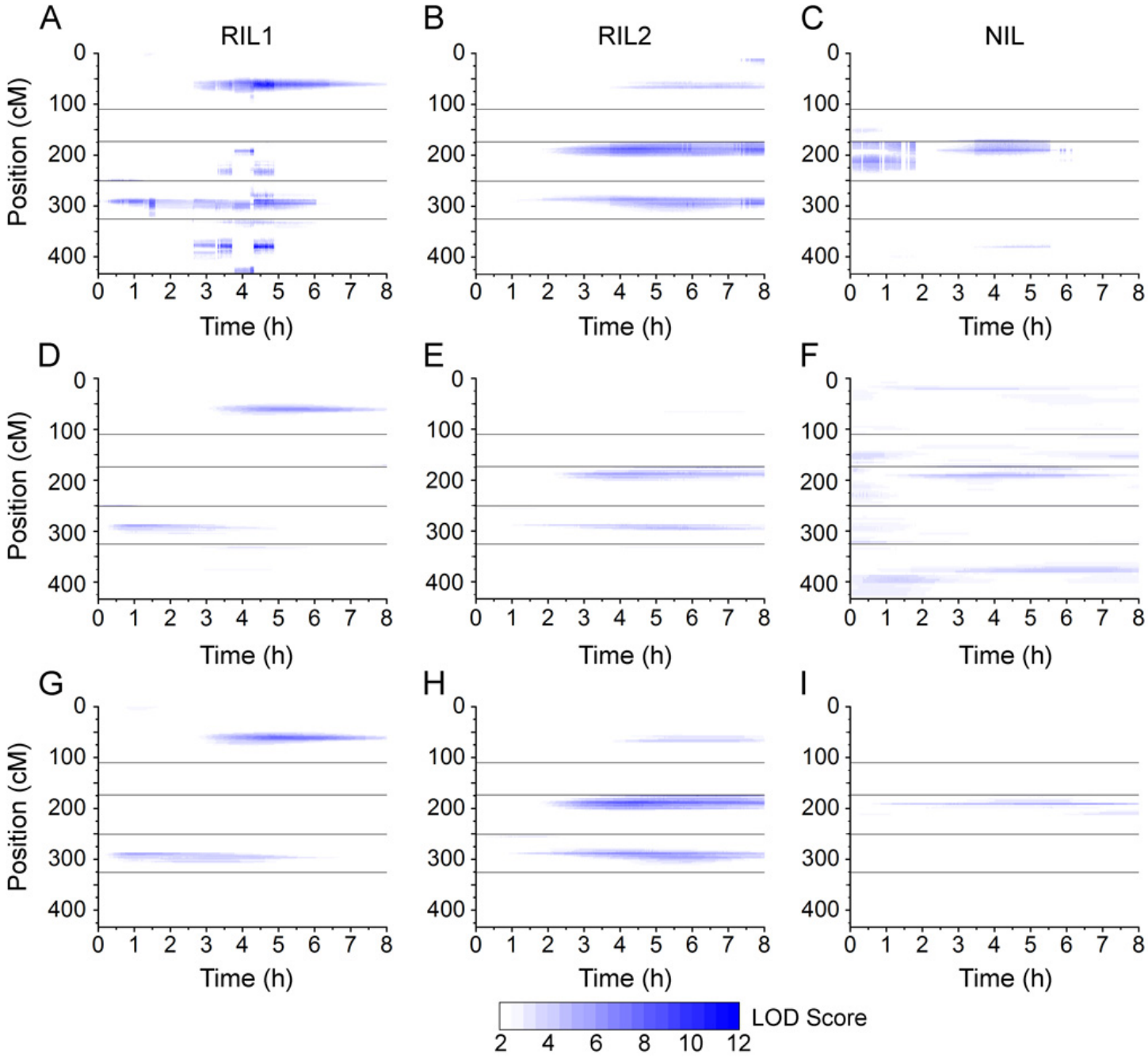


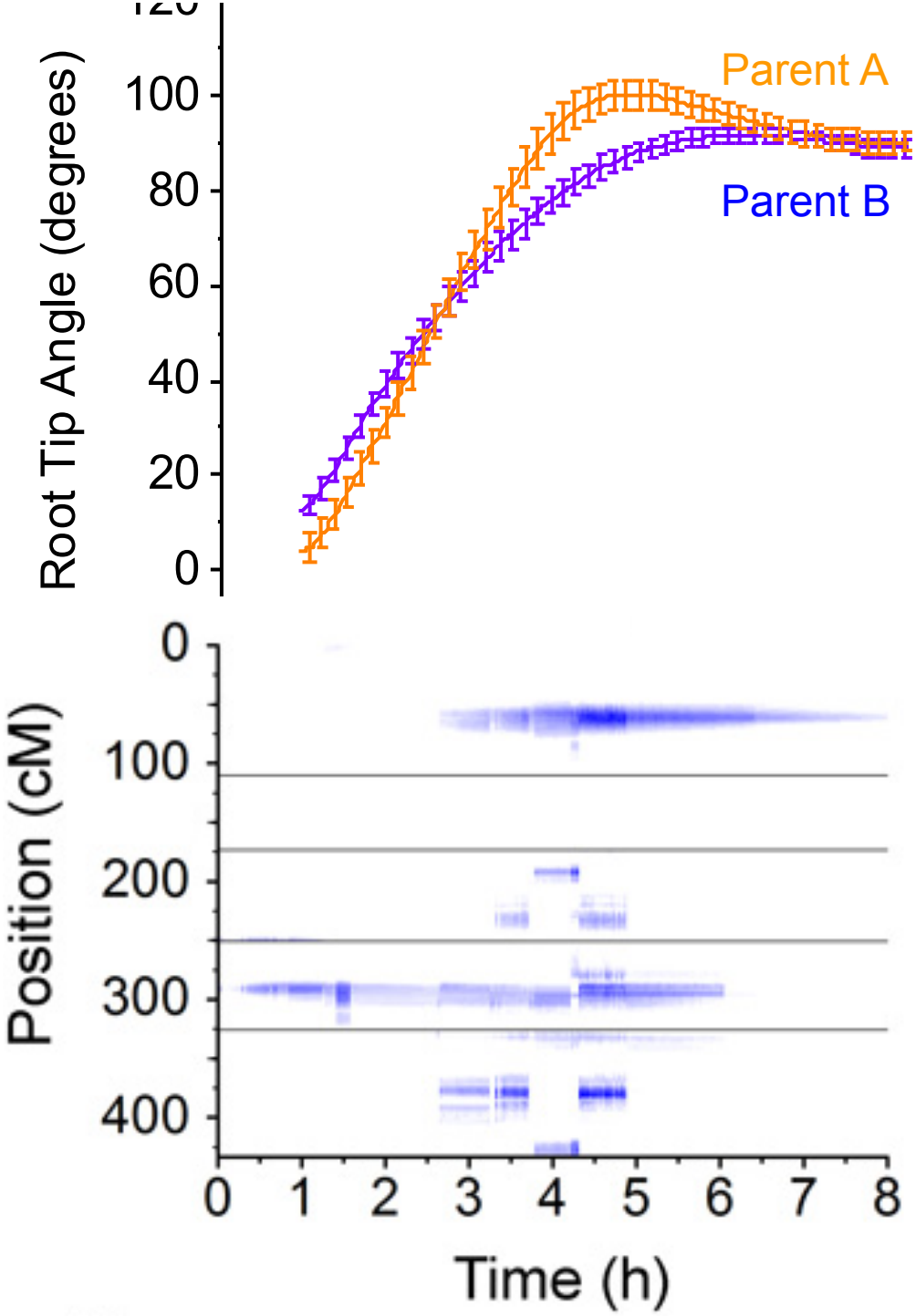
data mining

test new hypotheses

A single QTL plot
(root tip angle at one point in time is the phenotype)




A
RIL1
B
RIL2
C
NIL
D
E
F
G
H
I
Position (cM)
0
100
200
300
400
Time (h)
0 1 2 3 4 5 6 7 8
LOD Score
2 4 6 8 10 12


Root Tip Angle (degrees)
100
80
60
40
20
0
Parent A
Parent B
Position (cM)
0
100
200
300
400
0
1
2
3
4
5
6
7
8
Time (h)

The Key People

Logan Johnson

Candace Moore

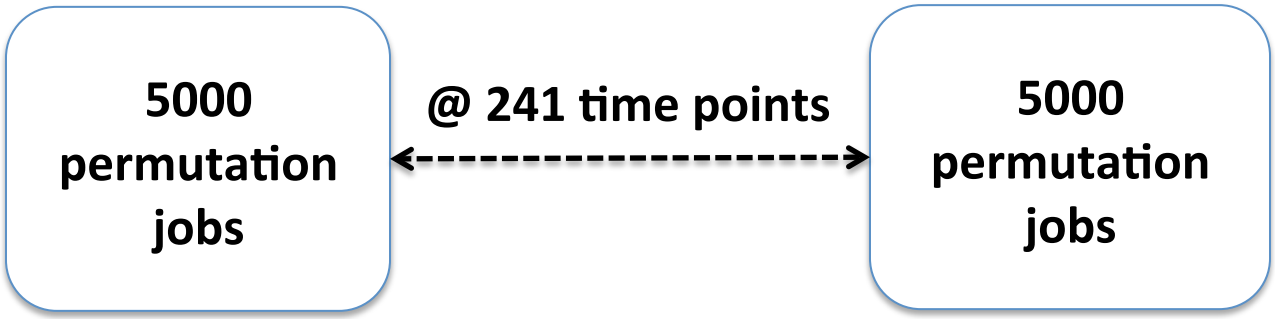
Logan + the Grid

To set the LOD score significance threshold you permute the phenotype against the genotype lots of times, to see what kind of signal 'chance' produces.

We chose to perform 25,000 permutations. To distribute this work on a grid, we bundled 5 permutations per job and ran 5000 jobs.

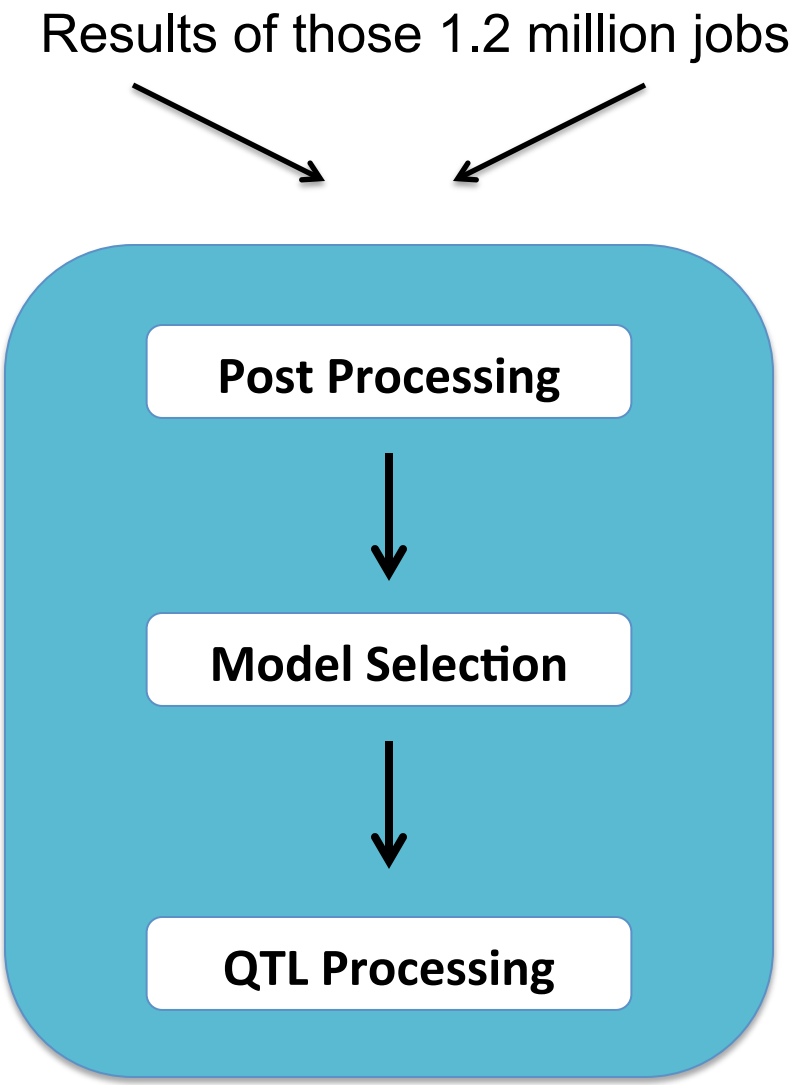
5 permutations per job **X 5000 jobs = 25,0000 permutations**

But we made measurements every 2 min for 8 h, essentially 241 separate expts



5000 jobs @ 241 time points = 1,205,000 jobs

This work was performed by HTCondor *Glide In* from CHTC to OSG


Results of those 1.2 million jobs
Post Processing
Model Selection
QTL Processing

Biology is becoming a high-throughput science

New insights will come from Big Data

People who are comfortable with HTC will win!