Genomic and Bioinformatic Technologies

Jeremy Schmutz, HudsonAlpha Institute
Todd Mockler, Donald Danforth Plant Science Center

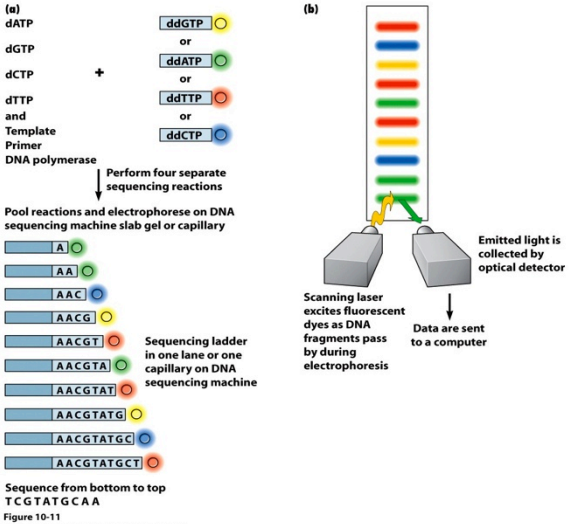
DOE ARPA-E "Advanced Plant Phenotyping for Increased Bioenergy Crop Yields"

Overview

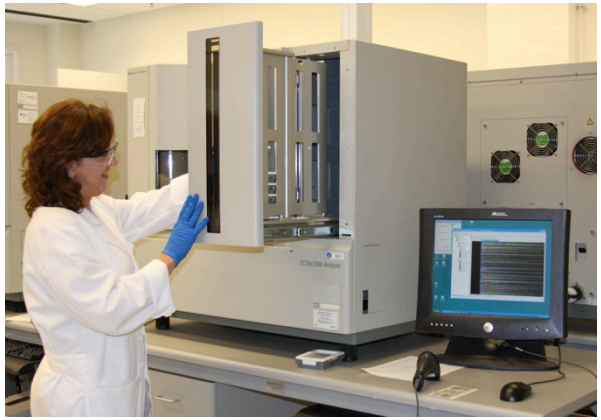
- Sequencing technology changes
- Plant reference genomes
- Sequence based tools for interrogation of reference genomes
- Plant databases/computational analysis platforms
- Bioinformatics landscape
- Controlled environment phenotyping and analytics
- Where we need to go

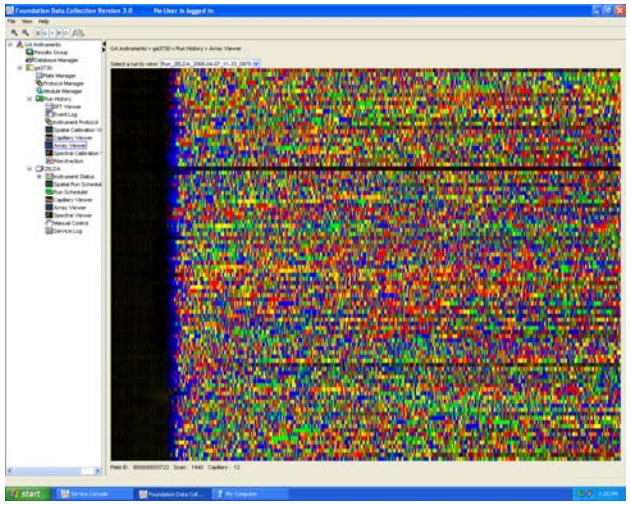
Sanger sequencing (3730xl 2002)

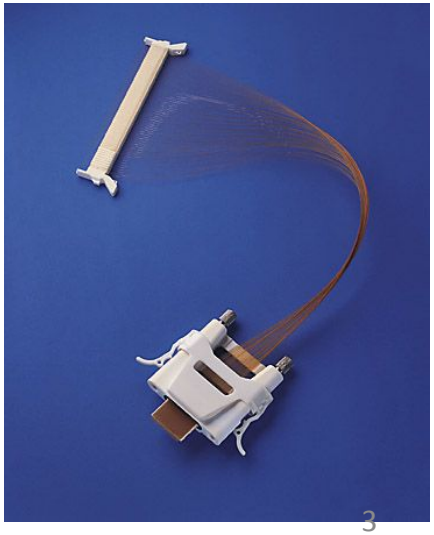
- Integrated robot loader and plate piercer –holds 24 plates
- 2 hour run time per 96 samples, 800 bp reads@ ~900k bps per day – ran unattended
- $937,500 per GB



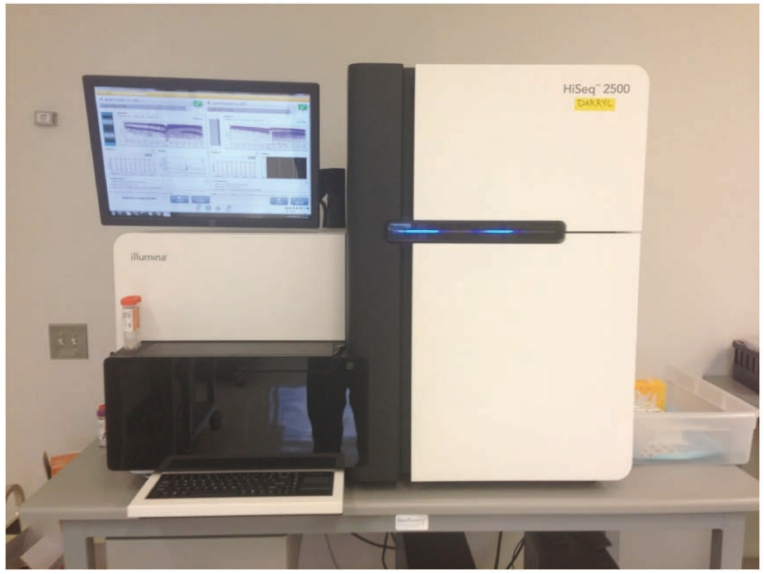
Recombinant DNA: Genes and Genomes (3e)
© 2007 W. H. Freeman and Company

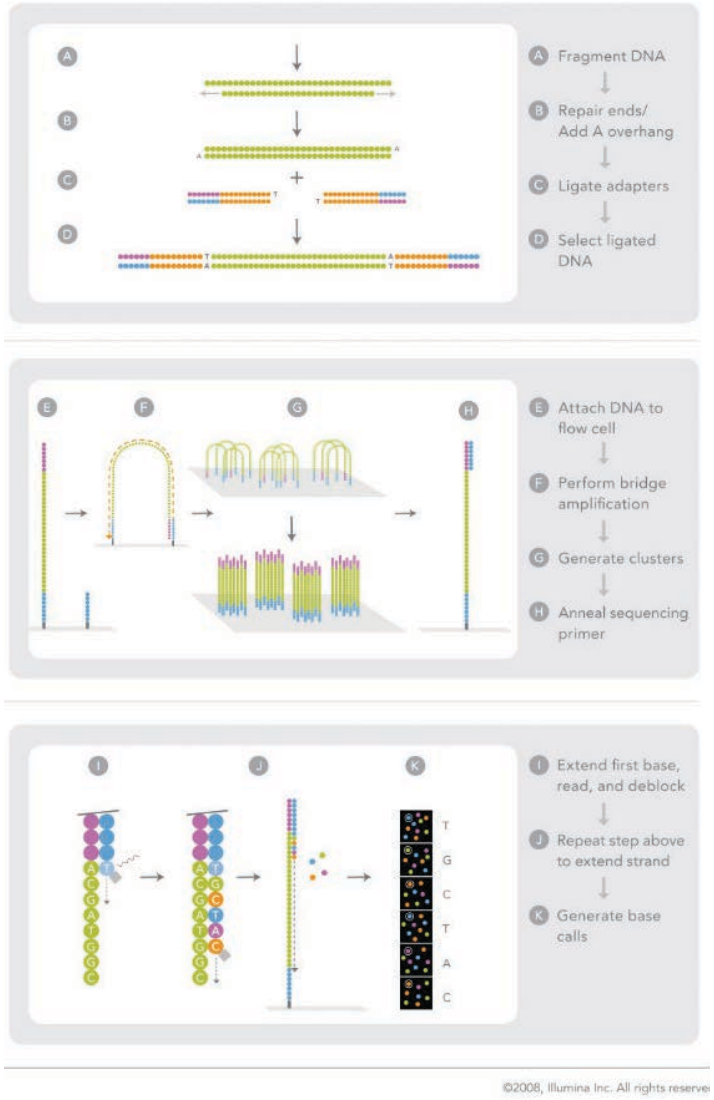






Modern data collection





- Produces ~900 Gb of data in 6 days (150 Gb per day)
- ~40k per run, or $44 per GB excluding library, machine cost, and data handling (1,500x Sanger & 23,000x lower cost)
- 2x125 reads currently, can also generate 2x200 reads at 2x the cost and ¼ the throughput

@97ZXTR1:359:H9E8NADXX:1:1101:5045:3839 1:N:0:CGATGT
GCCCTGCCTCCGAGCCAGATTTCCTTCCCGTTCGTTCCCGTGGCTTCTGGAGTCTTCTAGATGATAGAAAA
TTGCGCGACACGTTAATATCTCTATGTAAACCCGACGTGTGGGCCTTT

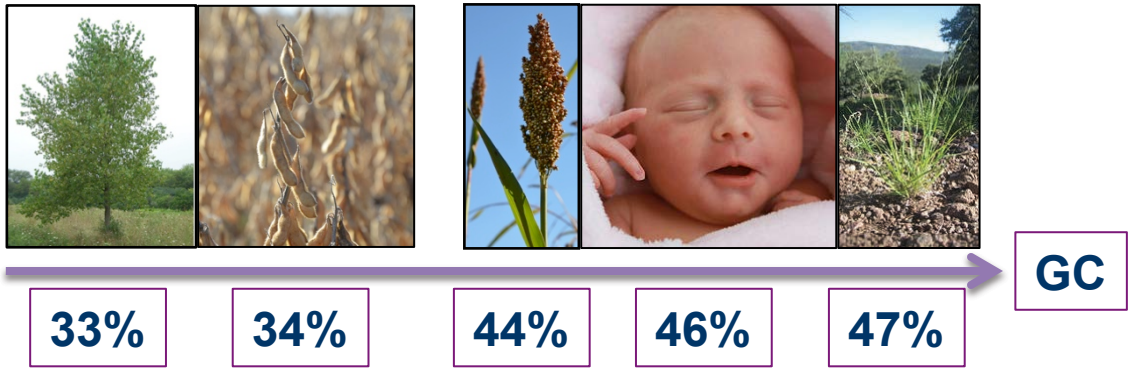
Plant reference genomes

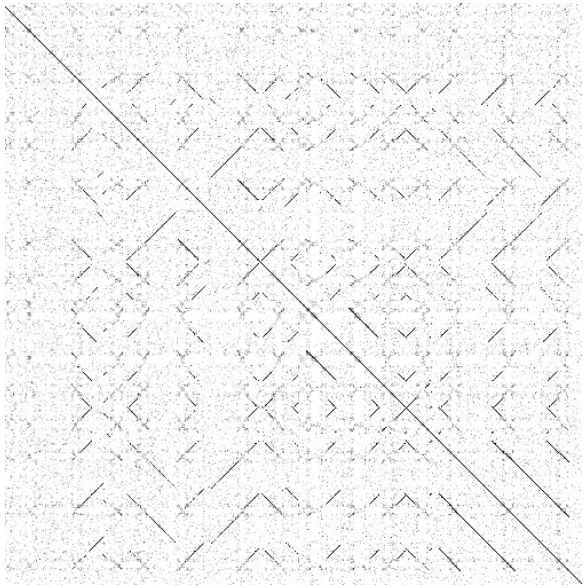
Why *de novo* plant genomes are hard

Polyploidy & Polymorphism

Repeat content & Transposons

MECHANISMS THAT DRIVE PLANT EVOLUTION!

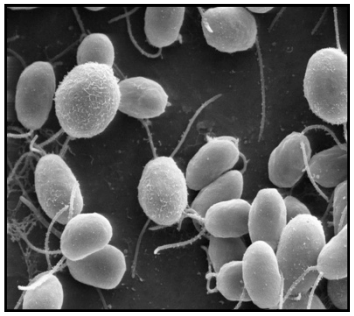




Reference plant genomes

- Come in different "flavors" – quality and completeness (finished/ improved vs. draft vs. drafty draft)

Sanger based BAC by BAC projects



Sanger based WGS genomes

NGS references

BAC reference genomes

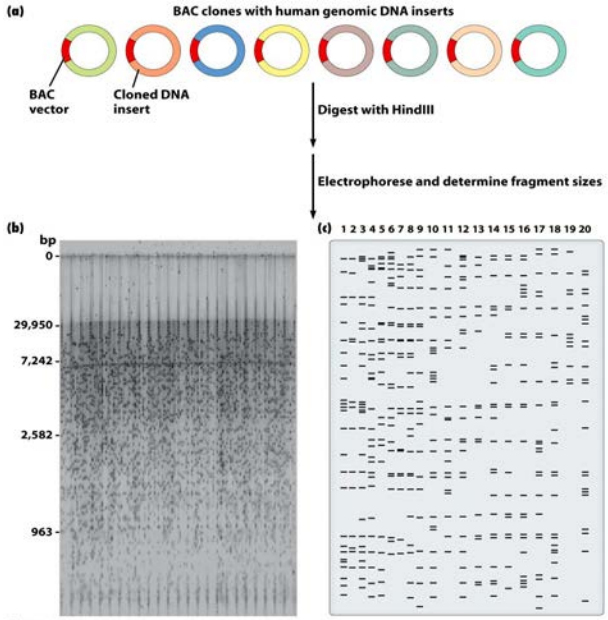
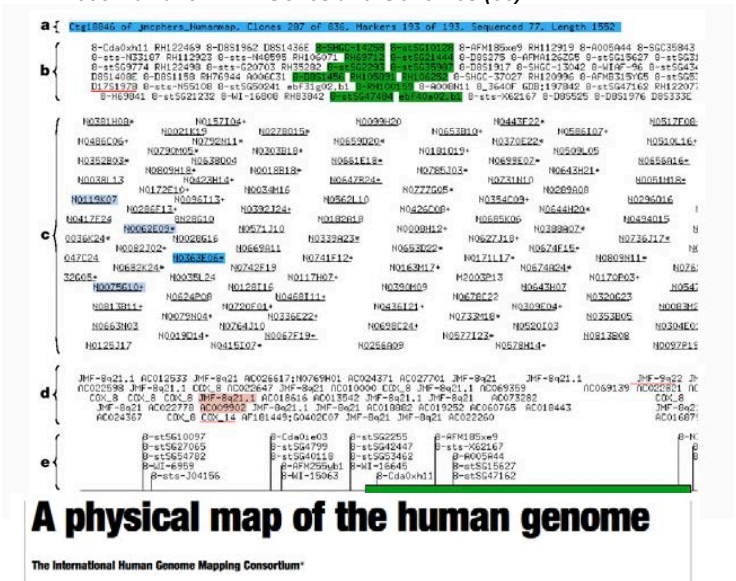
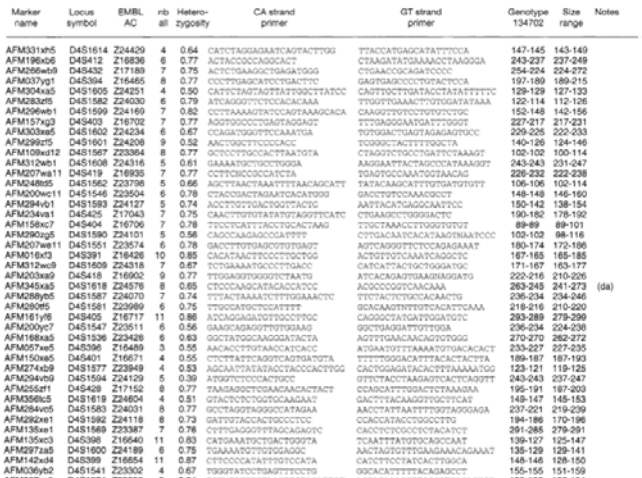


Figure 11-9

Recombinant DNA: Genes and Genomes (3e)





1994 AD: 2,066 markers; Gyapay

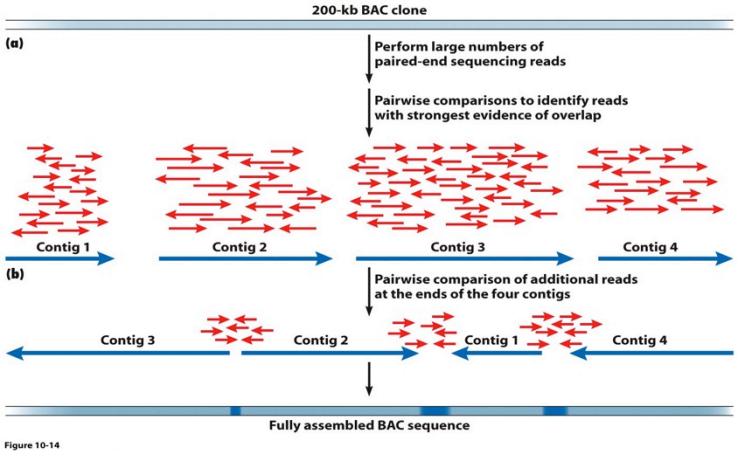


Figure 10-14

Recombinant DNA: Genes and Genomes (3e)

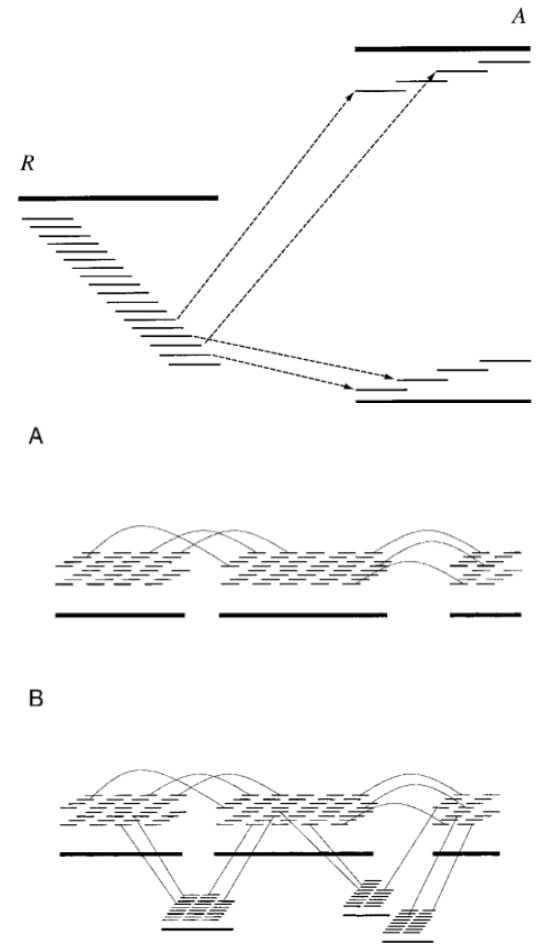
© 2007 W. H. Freeman and Company

WGS references

- Sequence the whole genome at once using multiple sized libraries
- Calculate all overlaps between reads and then work out a parsimonious path through the data set

- **WGS benefits**
  - Small number of good subclone libraries needed
    - Short Subclone (3-4kb)
    - Medium Subclone (6-9kb)
    - Long Fosmid (32-36kb)
    - Longer BAC (100-160kb)
  - No mapping phase
  - Less complex pipeline

- **WGS drawbacks**
  - Inconsistent genome coverage
  - Difficult to assess quality
  - Laborious post-production phase
  - Don't know what you have till you are done



ARACHNE: A Whole-Genome Shotgun Assembler

Serafim Batzoglou,[1,2,3] David B. Jaffe,[2,3,4] Ken Stanley,[2] Jonathan Butler,[2] Sante Gnerre,[2] Evan Mauceli,[2] Bonnie Berger,[1,5] Jill P. Mesirov,[2] and Eric S. Lander[2,6,7]

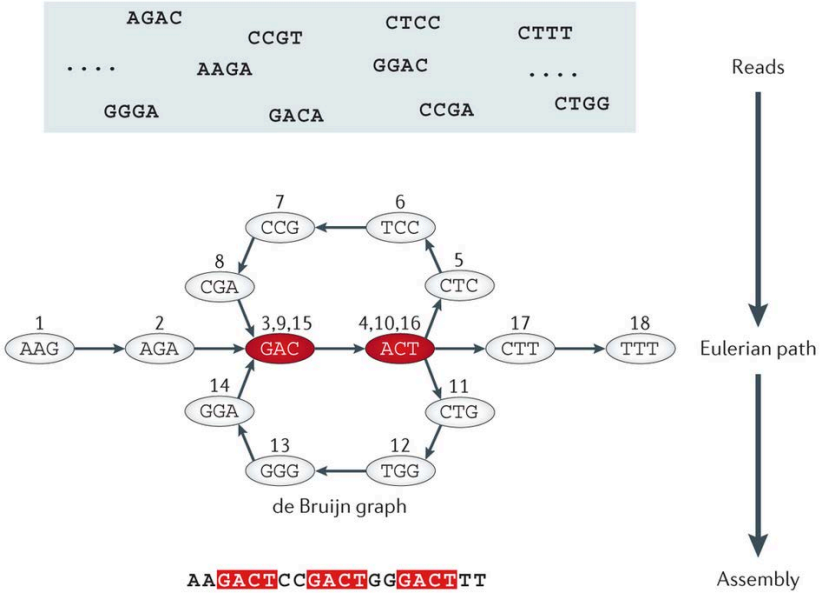
Genome Res. 2002

NGS “reference” genomes

- Sequence fragments of short inserts of varying sizes, recombination pairs, collapse very large data set into consensus

- **NGS benefits**
  - No cloning
  - Data collection is inexpensive
  - Sequence is high quality

- **NGS drawbacks**
  - Biases with whatever technology is used
  - Difficult to assess quality/completeness
  - Tend not to include long pairs, mapping information, validation
  - Lots of contigs, low repeat content, complex to use in practice
  - Difficult to improve with directed means
  - Fragmented



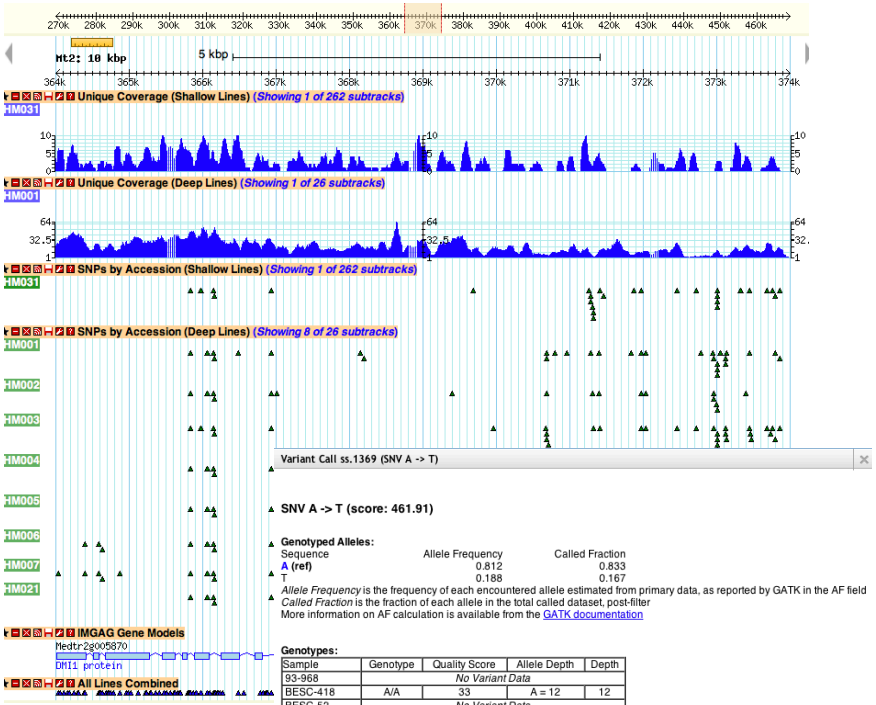
Computational solutions for omics data
Bonnie Berger, Jian Peng & Mona Singh
Nature Reviews Genetics 14, 333–346 (2013) doi: 10.1038/nrg3433

Sequence based tools for interrogation of reference genomes

Resequencing for SNP/indel identification

- Collect DNA from cultivars, breeding populations, make frag libraries, resequence
- 15 x of inbred sorghum = 14 Gb, $1,000 (detect homozygous changes)
- 30x of outbred sorghum = $1,800 (detect heterozygous changes)
- Lower coverage for pedigree typing 2x = $200

Medicago truncatula



Branca 2010 PN

Genotypes:

Sample	Genotype	Quality Score	Allele Depth	Depth
93-968	*No Variant Data*			
BESC-418	A/A	33	A = 12	12
BESC-52	*No Variant Data*			
BESC-79	A/A	63	A = 22	22
BESC-246	A/A	27	A = 10	10
BESC-313	A/T	50	A = 2; T = 4	6
CA-05-06	A/A	24	A = 10	10
BESC-460	*No Variant Data*			
GW-10958	A/A	24	A = 10	10
Nisqually	*No Variant Data*			
DENA-17-3	A/A	21	A = 7	7
CA-01-01	A/A	36	A = 13	13
VNDL-27-4	A/A	21	A = 8	8
BESC-15	A/T	99	A = 9; T = 6	15
BESC-105	A/T	23	A = 1; T = 4	5
BESC-366	A/T	99	A = 4; T = 4	8

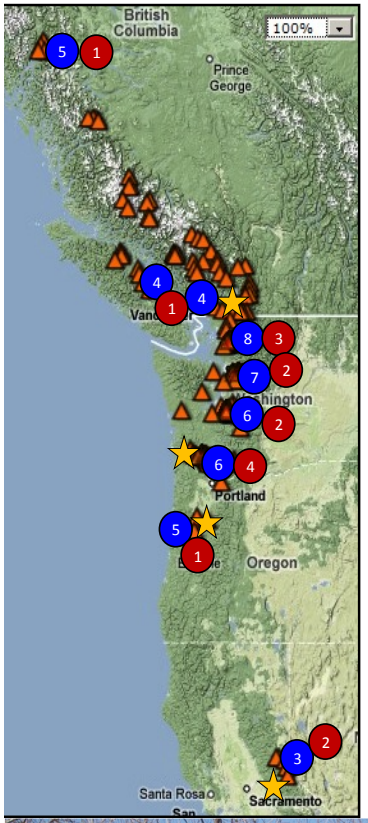
Allele Depth is the read count for each allele, pre-filter (AD subfield). *Depth* is the total read count, post-filter (DP subfield)
More information on these two fields is available from the GATK documentation in the link above

SnpEff Variant Annotation, where available:

Effect	Effect Impact	Codon Change	Gene Name	Transcript	Exon	Genotype
DOWNSTREAM	MODIFIER	1516	Potri.001G001900	Potri.001G001900.1		T
DOWNSTREAM	MODIFIER	2857	Potri.001G002000	Potri.001G002000.1		T
DOWNSTREAM	MODIFIER	2857	Potri.001G002000	Potri.001G002000.2		T
INTRON	MODIFIER		Potri.001G001800	Potri.001G001800.1	1	T
INTRON	MODIFIER		Potri.001G001800	Potri.001G001800.2	1	T
INTRON	MODIFIER		Potri.001G001800	Potri.001G001800.3	1	T
UPSTREAM	MODIFIER	1535	Potri.001G001700	Potri.001G001700.1		T

Poplar - Phytozome

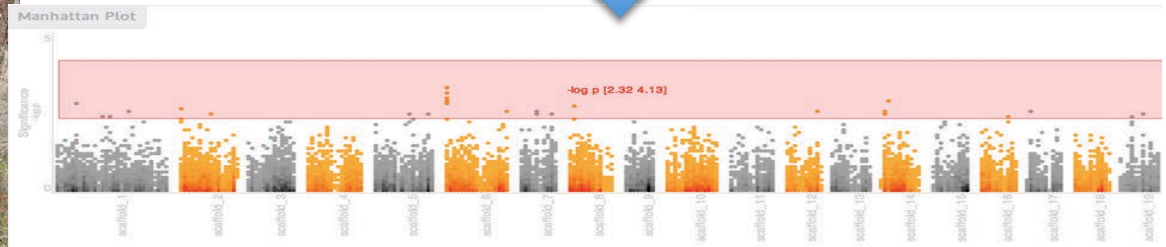
Real world example



Dense SNPs across population

```
#CHROM  POS     REF     201782_400122  201782_400194
        201782__400495 ...
Chr01   4       C       C/C     C/C    C/T
Chr01   5       A       A/A     A/A    A/A
Chr01   7       ACCCC   ACCCC/ACCCC    ACCCC/ACCCC    ACCCC/ACC
Chr01   18      A       A/A     A/A    A/A
Chr01   20      A       A/A     A/A    A/A
Chr01   21      C       C/C     C/C    C/C
Chr01   22      C       C/C     C/C    C/C
Chr01   23      C       C/C     C/C    C/C
Chr01   24      CAAA    CAAA/CAAA      CAAA/CAAA      CAAA/CAAA
Chr01   34      ACCCC   ACCCC/ACCCC    ACCCC/ACCCC    ACCCC/ACCCC
....
```

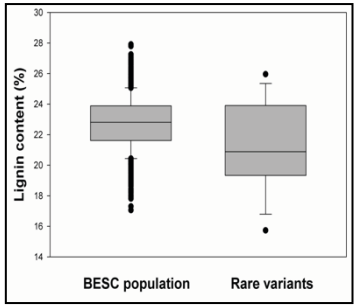
Detailed phenotyping

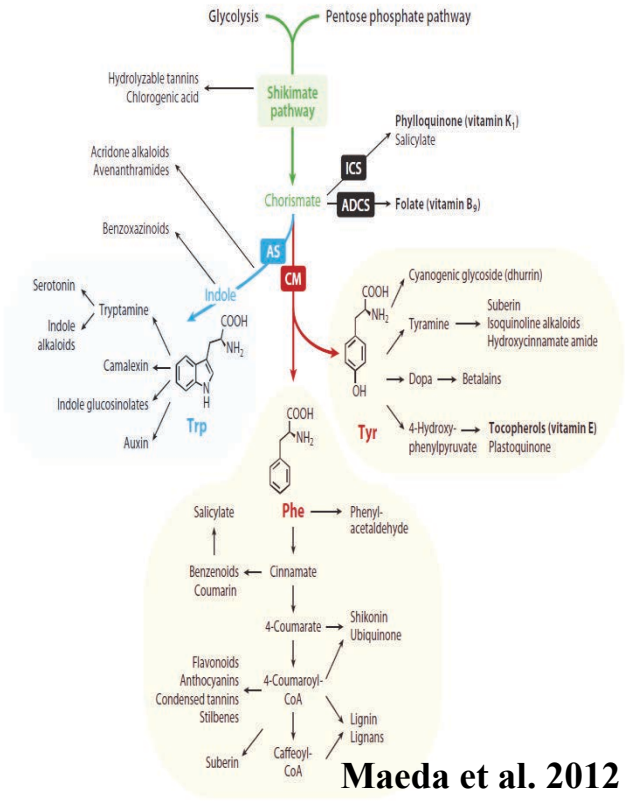


Jerry Tuskan: ORNL/BESC

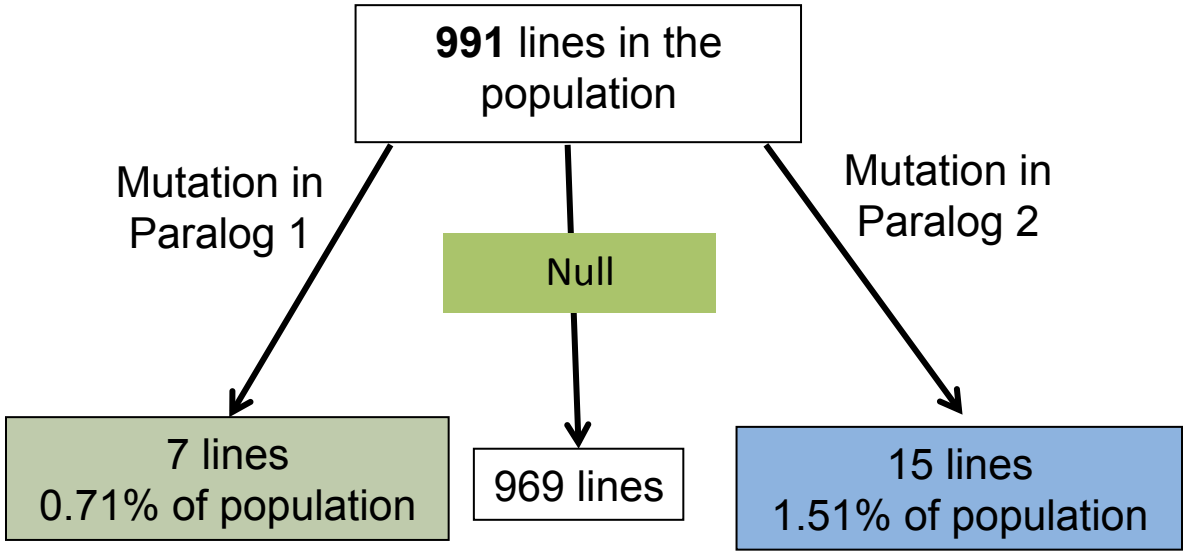
BioEnergy Science Center

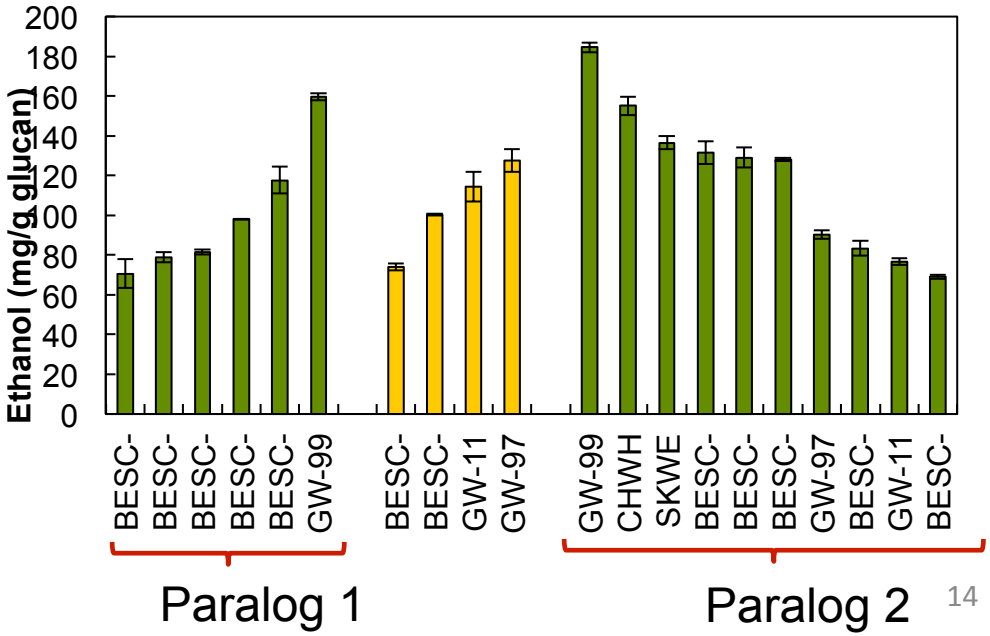
Lignin quality traits





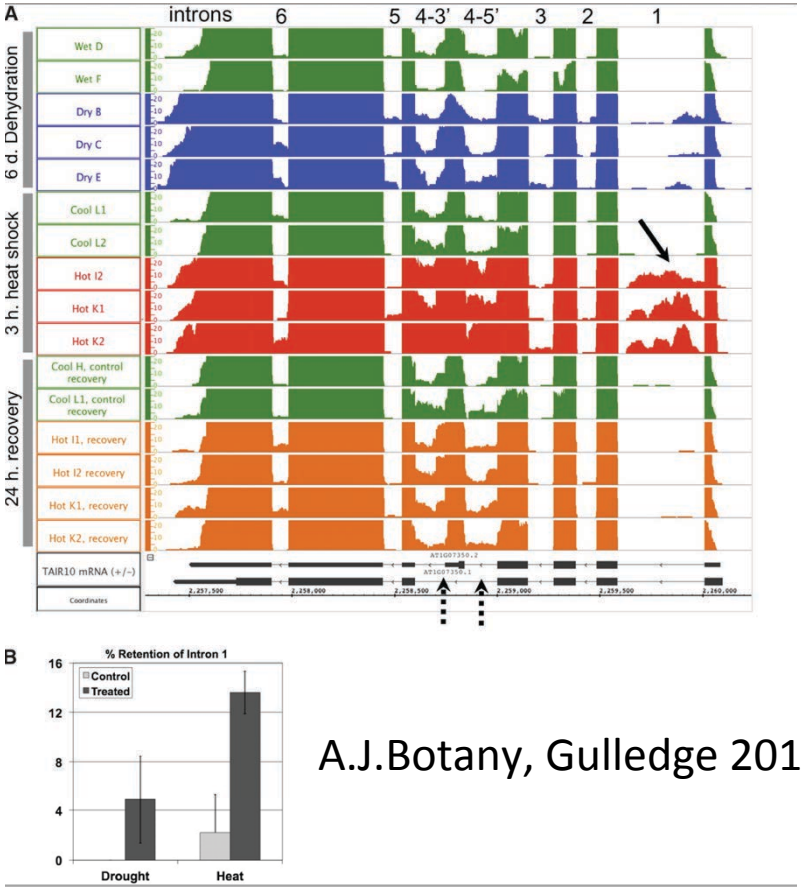
Maeda et al. 2012





RNA-seq

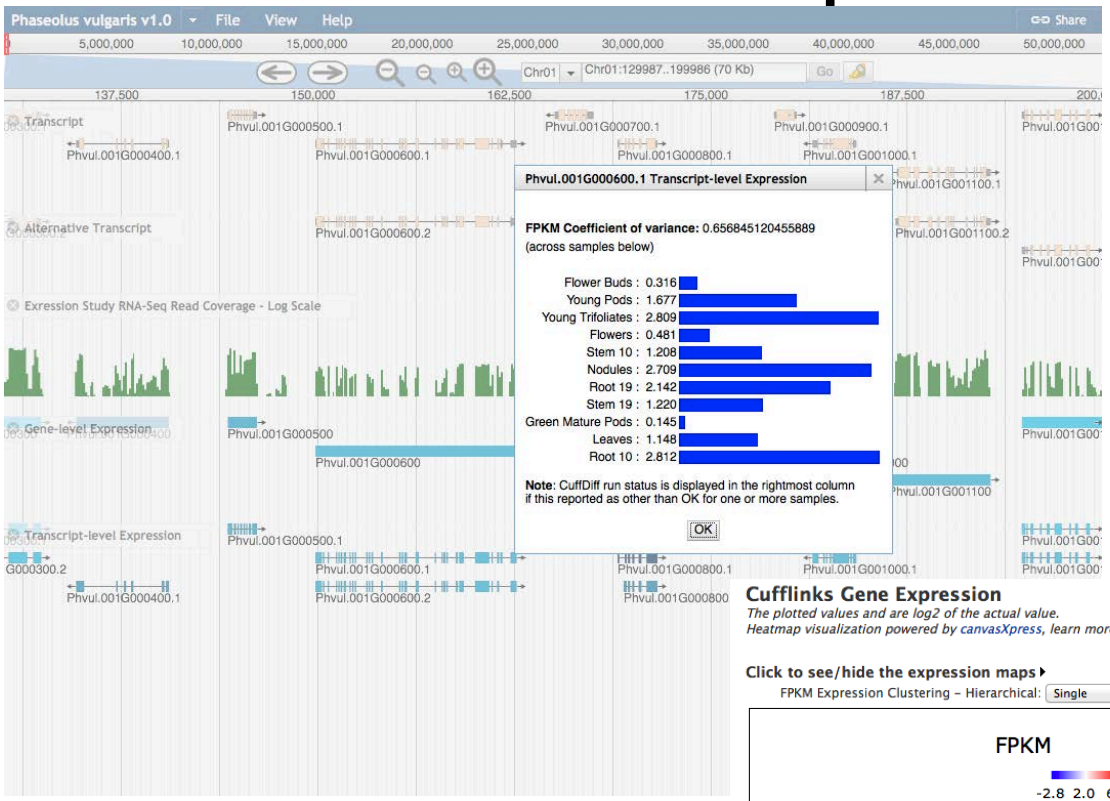
- Isolate RNA, make RNA-seq (RT) libs, collect data
- Expression analysis, tissue to tissue, condition to condition, or genotype to genotype comparisons
- Typically for full analysis run at 10-20M pairs per (or 10-20 samples per lane ~$325 per samples
- Variations such as 3' counting, non-coding RNA collection, microRNA collection

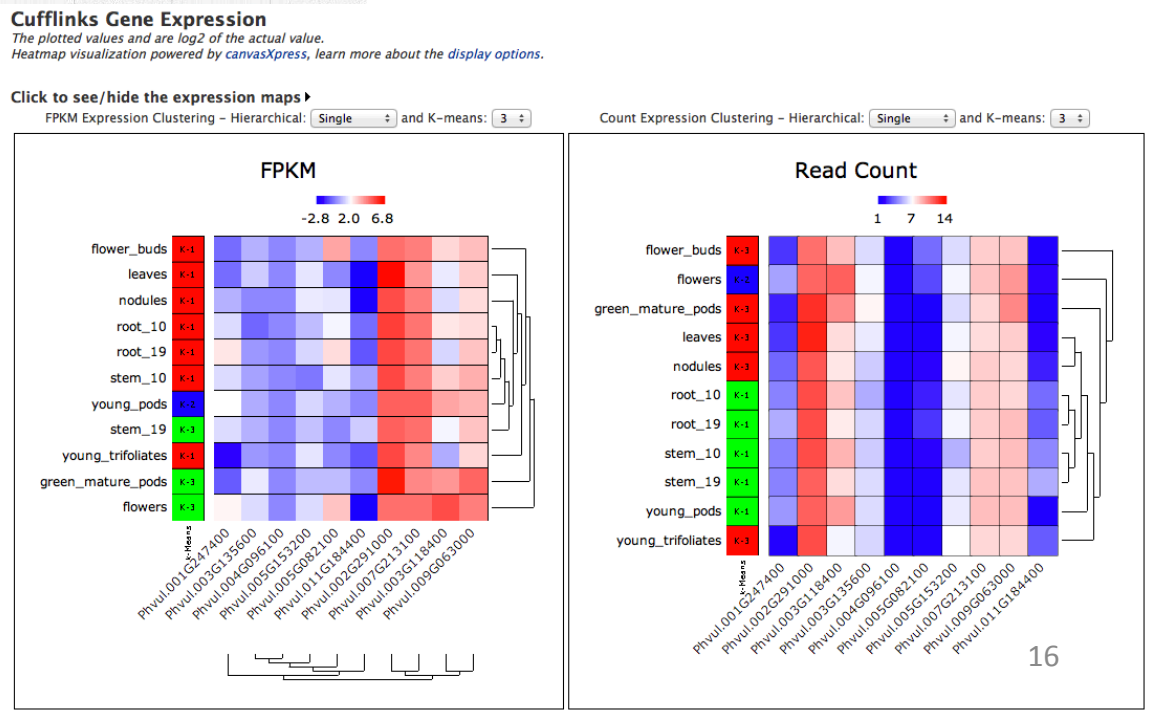


A.J.Botany, Gulledge 2012

Arabidopsis: alternative splicing in splicing regulator SR45a

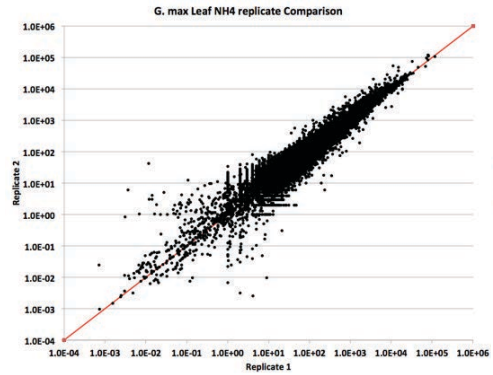
RNA-seq data displays

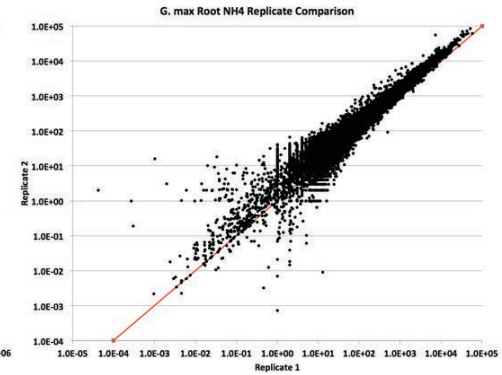


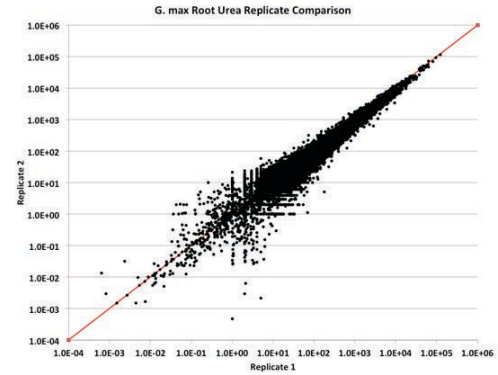


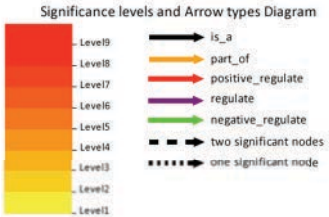
Phaseolus vulgaris

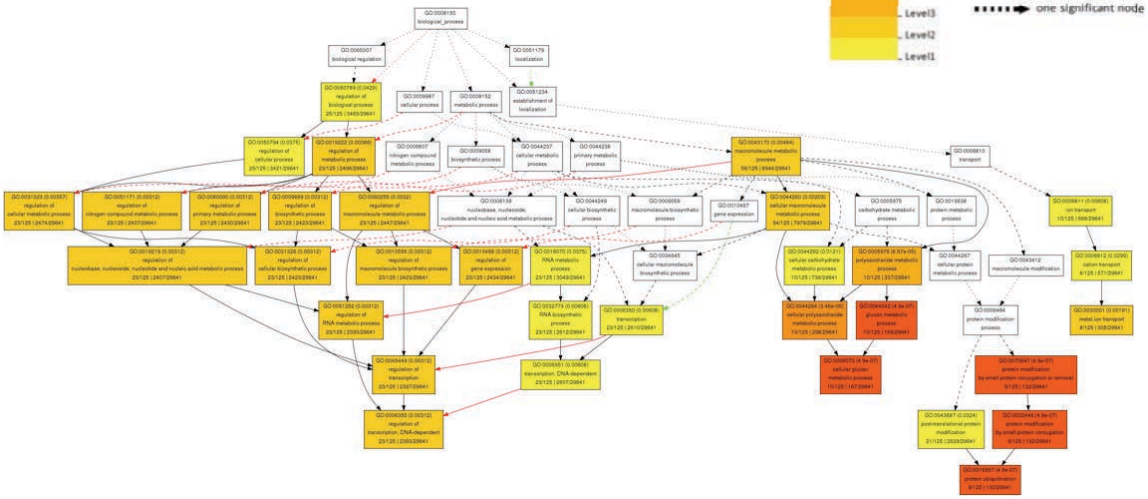
RNA-seq for expression comparisons





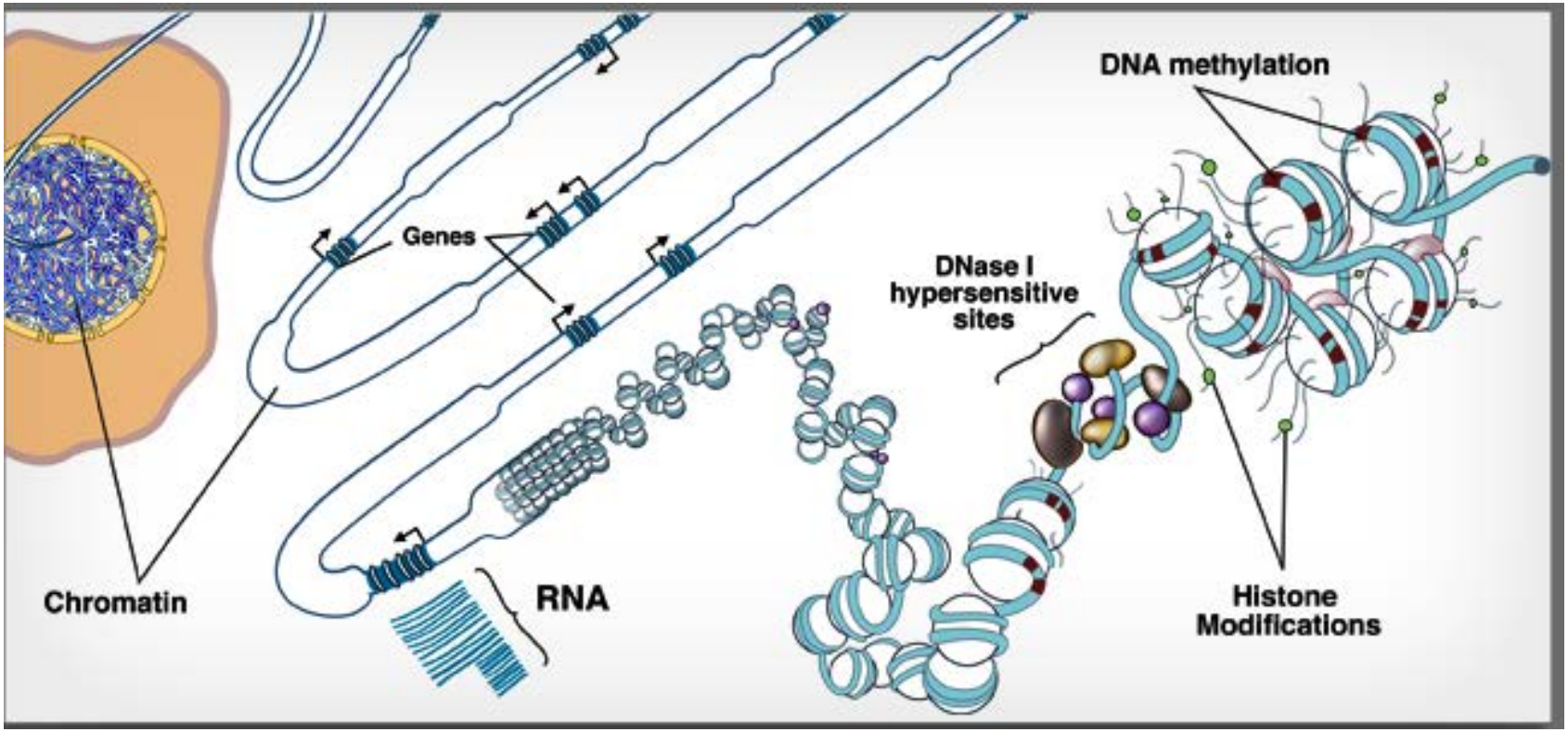






Glycine max (soybean)
w/ Gary Stacey : Mizzou

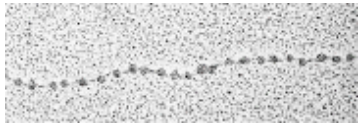
Histone modifications



http://www.roadmapepigenomics.org

Long range fiber-fiber interaction

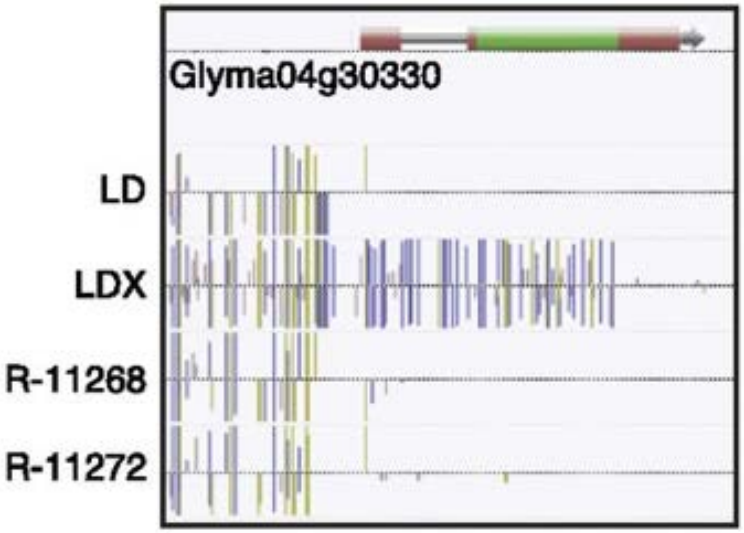
Short range internucleosomal interactions



Nucleosome

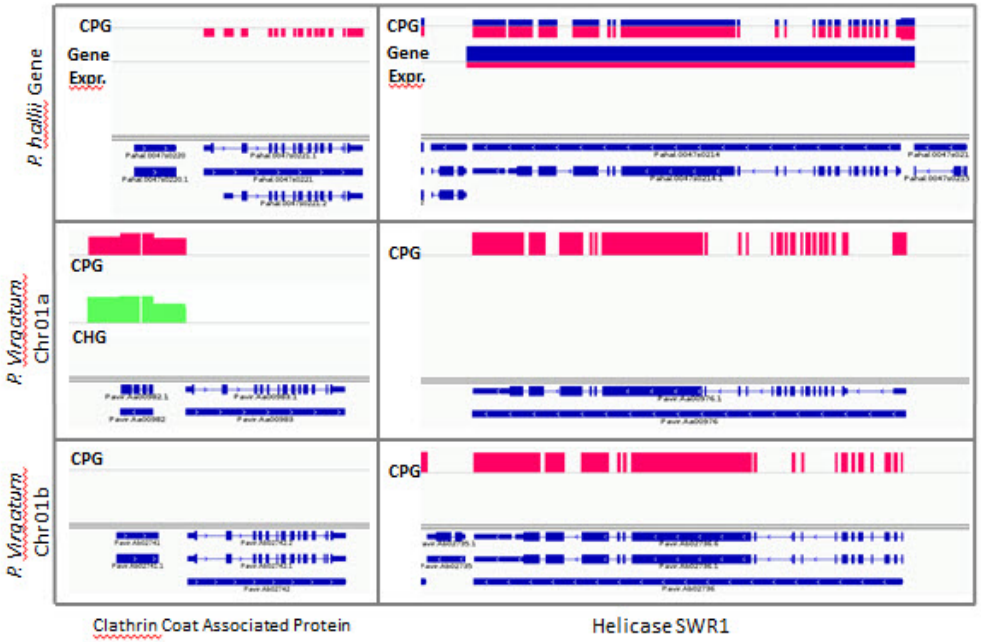
5mC methylation is likely important for crops!

- Bisulfate sequencing for bp resolution or Reduced Methylation Assay for broad picture
- Inherited methyl states affect transcription



Schmitz, GR, 2013

Glycine max (soybean)

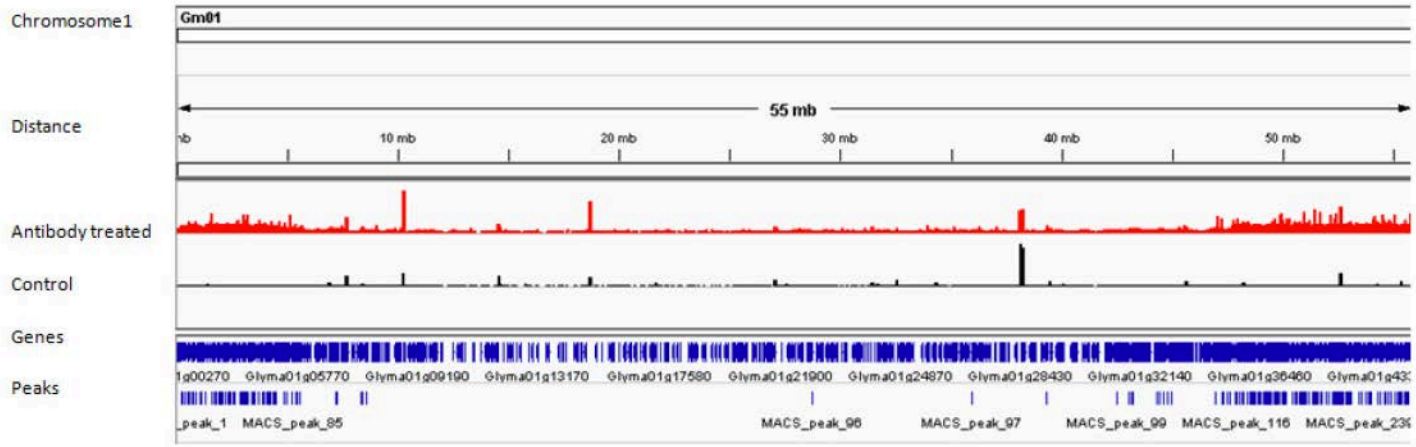


Methylation comp. between diploid *Panicum hallii* and tetraploid *Panicum virgatum*
Tom Juenger : UT-Austin

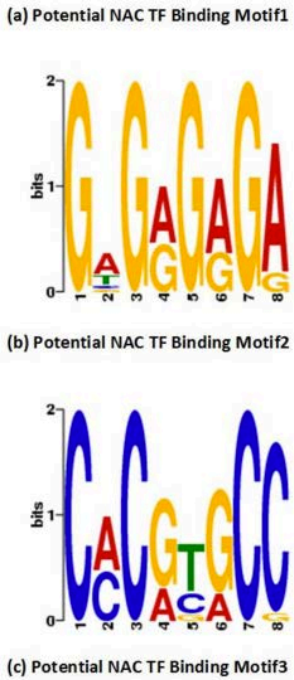
ChIP-Seq in plants

- Use antibodies to identify positions in a genome where a specific transcription factor binds

Glycine max (soybean) :NAC transcription factor

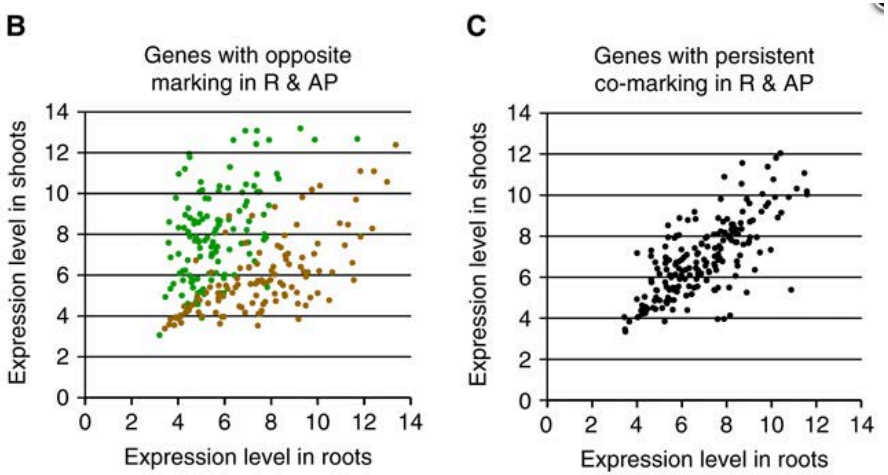


Shamimuzzaman, 2013 BMC Genomics

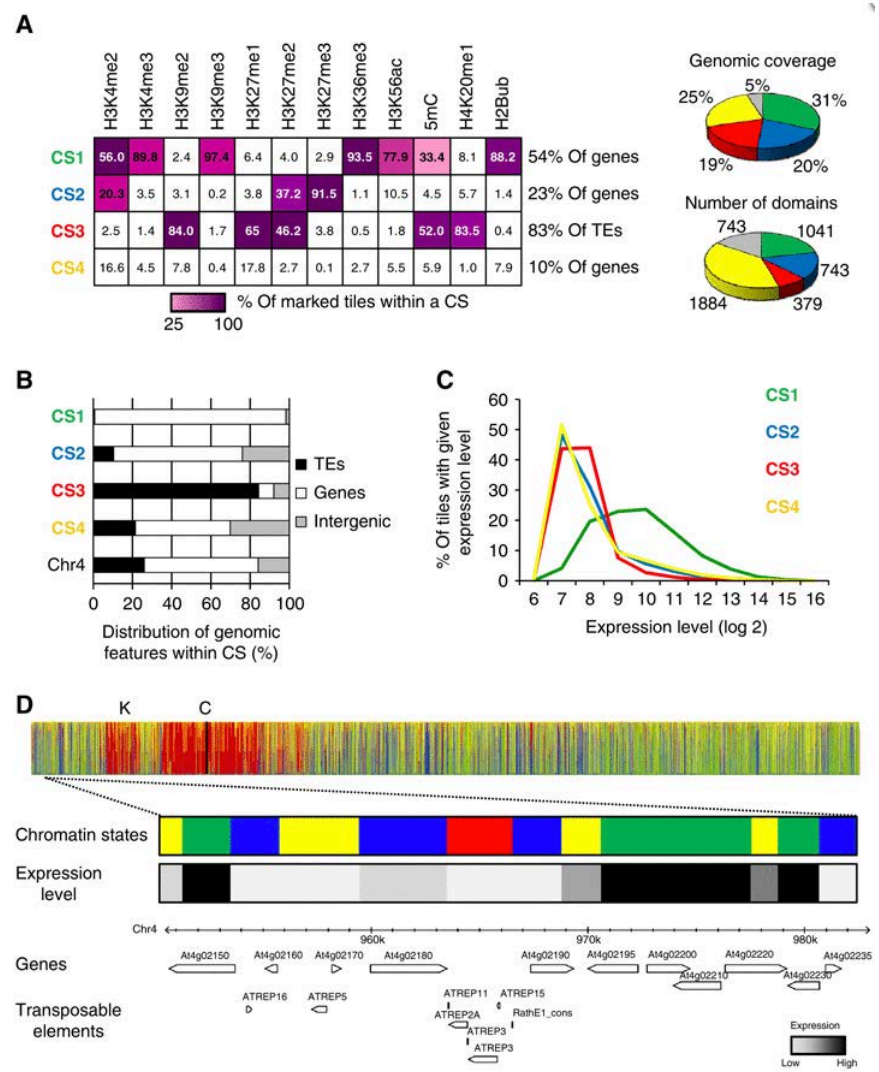


Histone modifications in plants

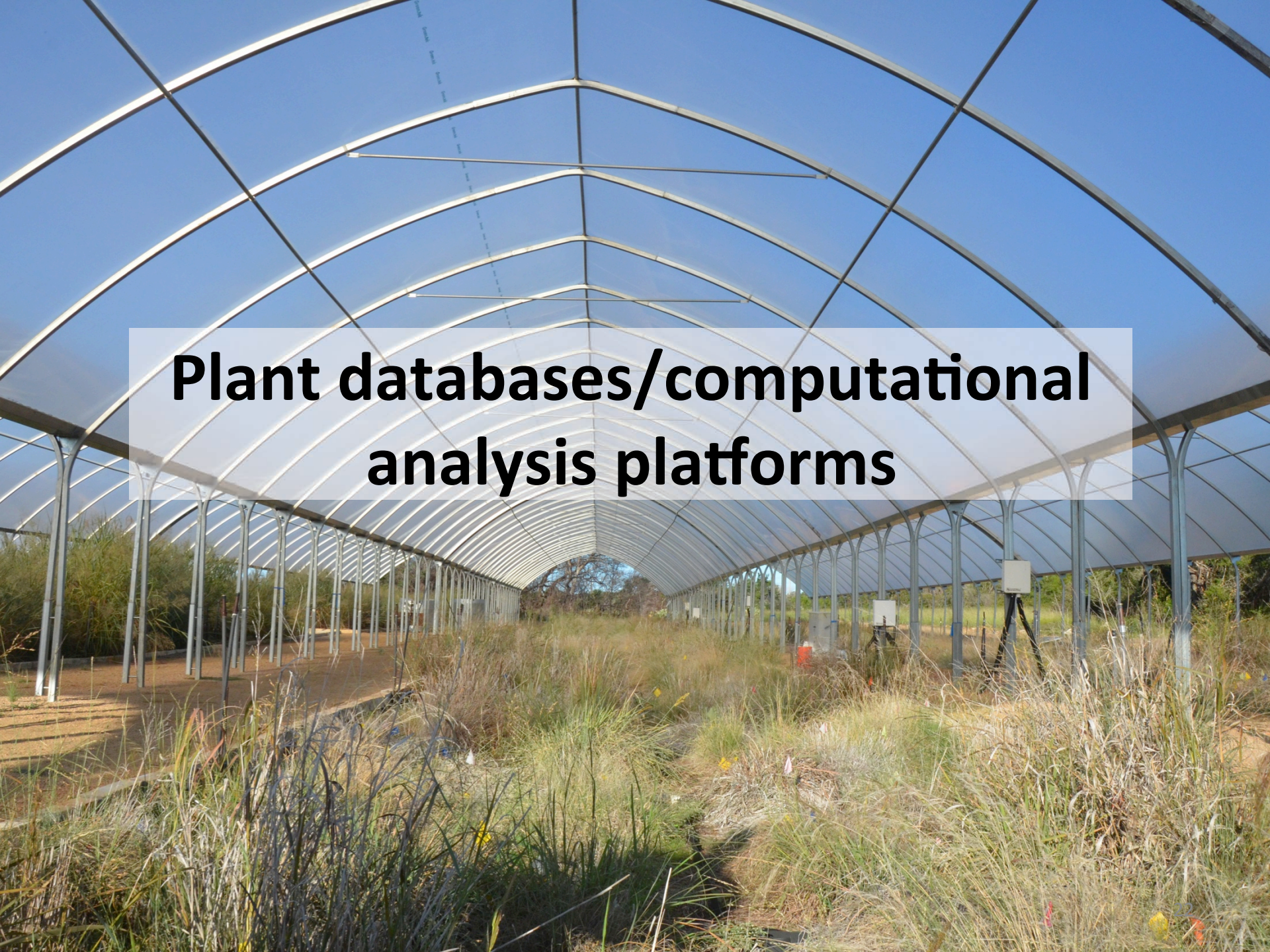
- Identify histone modifications using ChIP-Seq
- Index active genes, repressed genes, silent repeat elements and intergenic regions.



Arabidopsis thaliana

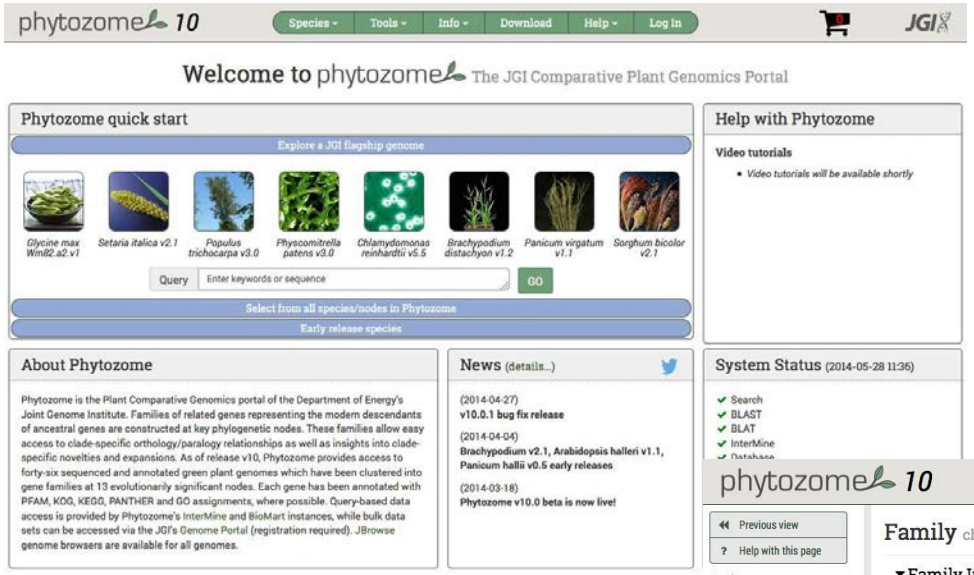


Roudier, EMBRO, 2011



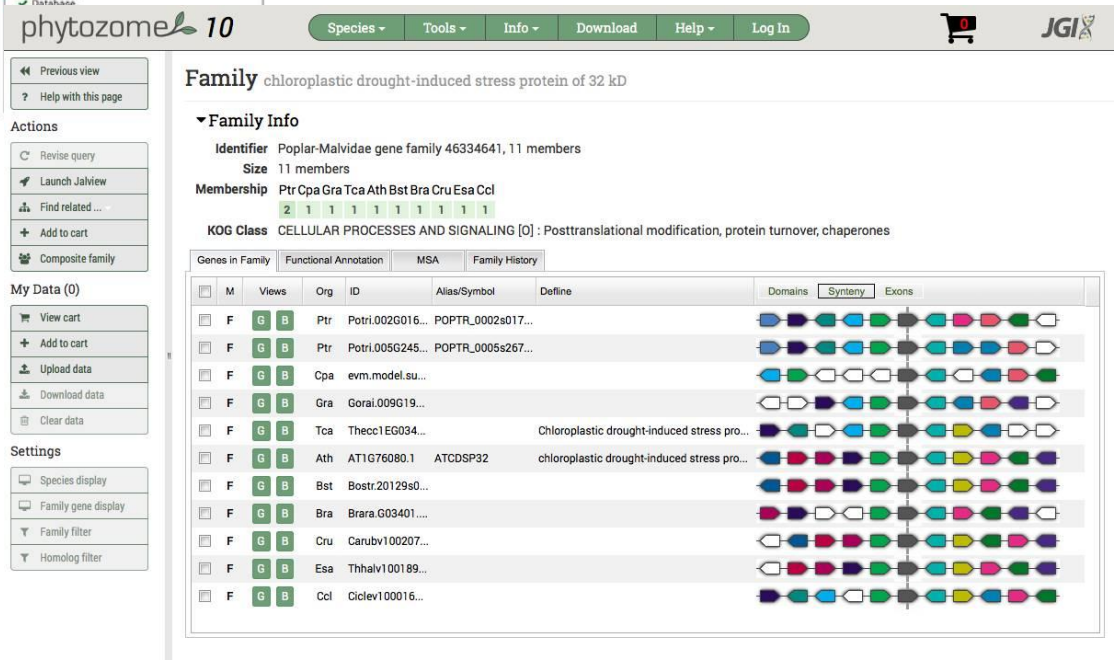
Plant databases/computational analysis platforms

JGI: Phytozome Portal



Data and Analyses available in v10

- 46 annotated plant and algal genomes
- Protein families constructed at
- InterPro/KEGG/KOG proteome annotation
- natural diversity data for *Ptr, Egr, Bdi*
- expression data for *Cre, Pvu, Gma*



Phytozome Features

- Genome-centric, gene-centric, family-centric search and visualization
- JBrowse genome browsers
- InterMine queryable data warehouse
- Searchable by sequence similarity, GO/ PFAM/KEGG/KOG/PANTHER,symbol, identifier, keyword,
- Custom user lists, data downloads
- Programmatic data access via web services API and multi-language client libraries

IPlant

iPlant Collaborative™ *Empowering A New Plant Biology*

Login or Register | About | Contact Us | Feedback

The iPlant Collaborative

The iPlant Collaborative develops cyberinfrastructure and computational tools to solve Grand Challenges in plant science

CHALLENGE

iPlant Genotype to Phenotype (iPG2P)

Mapping the links between genotypes and phenotypes

iPlant Tree of Life (iPToL)

Understanding the phylogenetic relationships between all plant life

Seed Projects

Supporting diverse cyberinfrastructure needs

DISCOVER

Discovery Environment

Access iPlant tools through a single user-friendly interface

MORE...

LOGIN

DNA Subway

An educator-tailored interface for bringing iPlant to the classroom

MORE...

LOGIN

Atmosphere

An integrative, private, self-service cloud computing platform

MORE...

LOGIN

Register for iPlant Tools

LEARN

Upcoming Events

- iPlant Tools and Services Workshop, UC Davis. March 12th and 13th 2012
  March 12 2012 - March 13 2012
- iPlant Workshop, UC Berkeley/ USDAARS Albany
  March 12 2012

MORE...

the iPlant Leaflet

Current Issue: December 05 2011

News and Announcements

- Trellis: Climb on for Faster Access to iPlant Resources
- 2.0 Release of Taxonomic Name Resolution Service

MORE...

CONNECT

People at iPlant

Community driven science

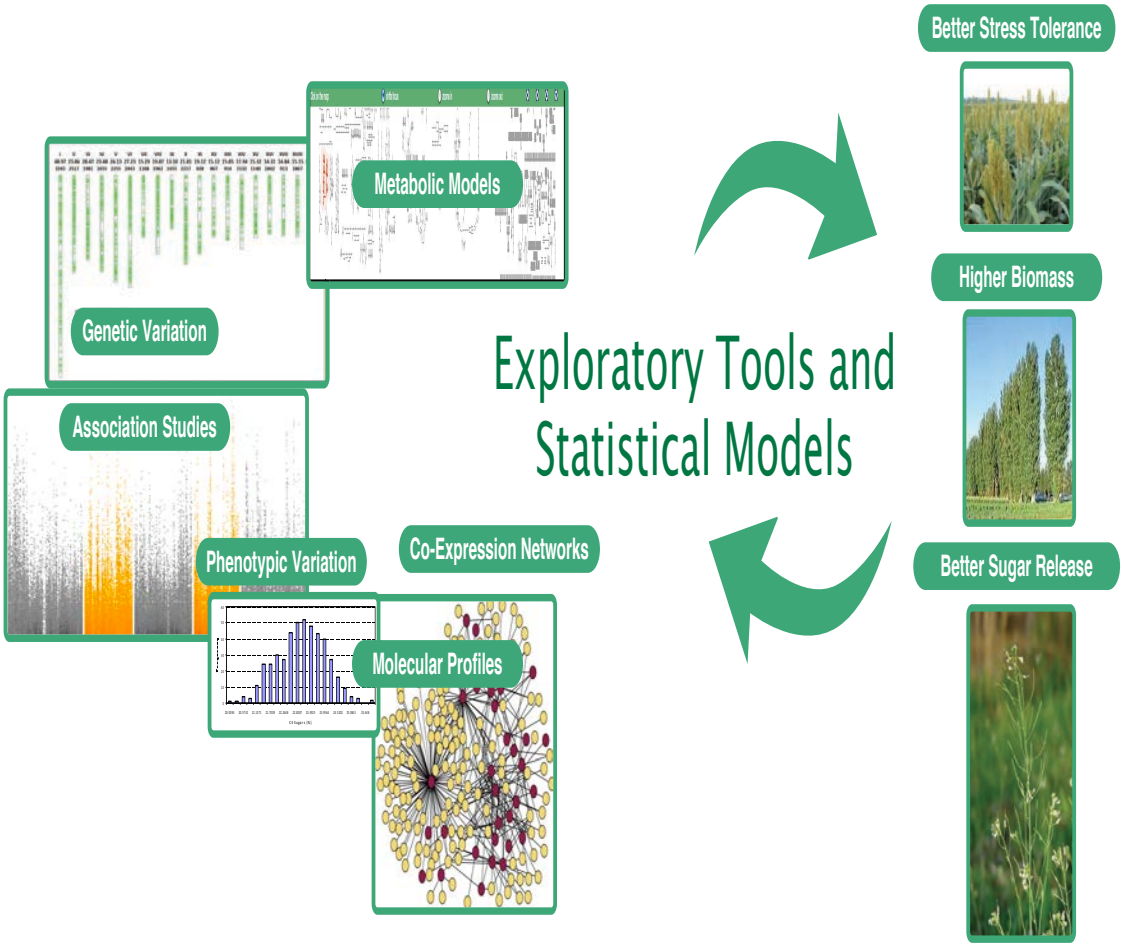
My-Plant.org

iPlant social networking

Find Us...

KBASE

- DOE-BER funded project to provide access to DOE computational resources
- Targeted towards metabolic modeling and statistical modeling
- Command line and GUI interfaces
- Under construction at http://kbase.us



KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Gramene: a Comparative Resource for Plants

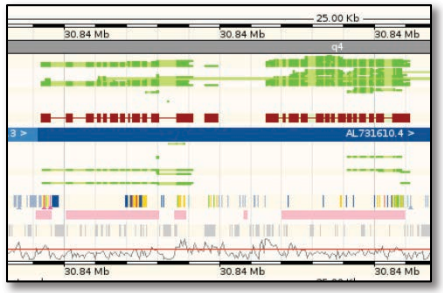
Gramene

Genome Browsers

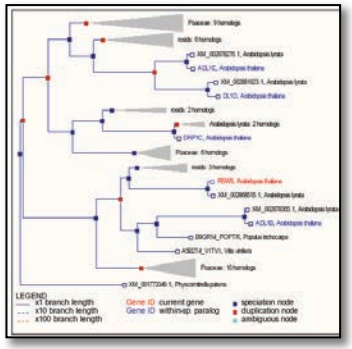
40 Species

Includes Rice, Maize & Arabidopsis

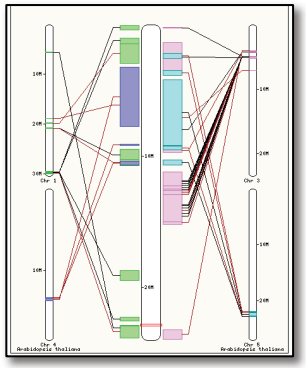
Annotation



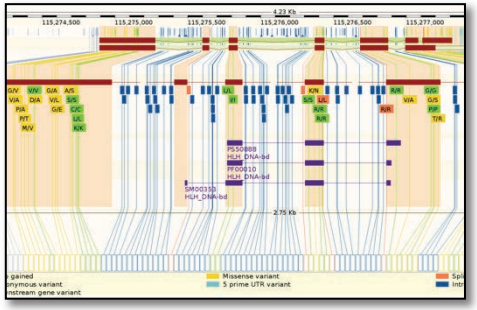
Gene trees & orthology



Comparative maps

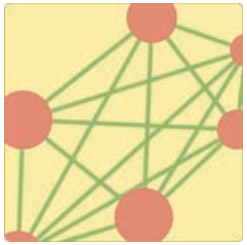


Variation & effect prediction



Collaborator: Paul Kersey (EBI) e!Ensembl

Pathways



Plant Reactome

Rice curated pathways

Maize & Arabidopsis projections

10 species "Cyc" metabolic pathways

Collaborator: Lincoln Stein (OICR)

Gramene services & programmatic access:
BLAST, BioMart, API, RESTFUL, MySQL

PI: Doreen Ware (CSHL) CSH

Co-PI: Pankaj Jaiswal (Oregon State U) OSU Oregon State University

NSF

Funded by NSF #0703908

Bioinformatics landscape

The Bioinformatics/Genome Analytics Landscape

Genome Assembly

- Allpaths-LG
- MaSuRCA

Gene Prediction

- MAKER-P
- Augustus

Variation Analysis

- GATK suite
- TASSEL

Network Prediction

- WGCNA
- Aracne

Transcriptome Assembly

- Trinity
- Abyss

Functional Annotation

- orthoMCL
- InterProScan

Transcriptome Profiling

- "Tuxedo" suite
- GeneCounter

Other HTS 'Omics Analysis

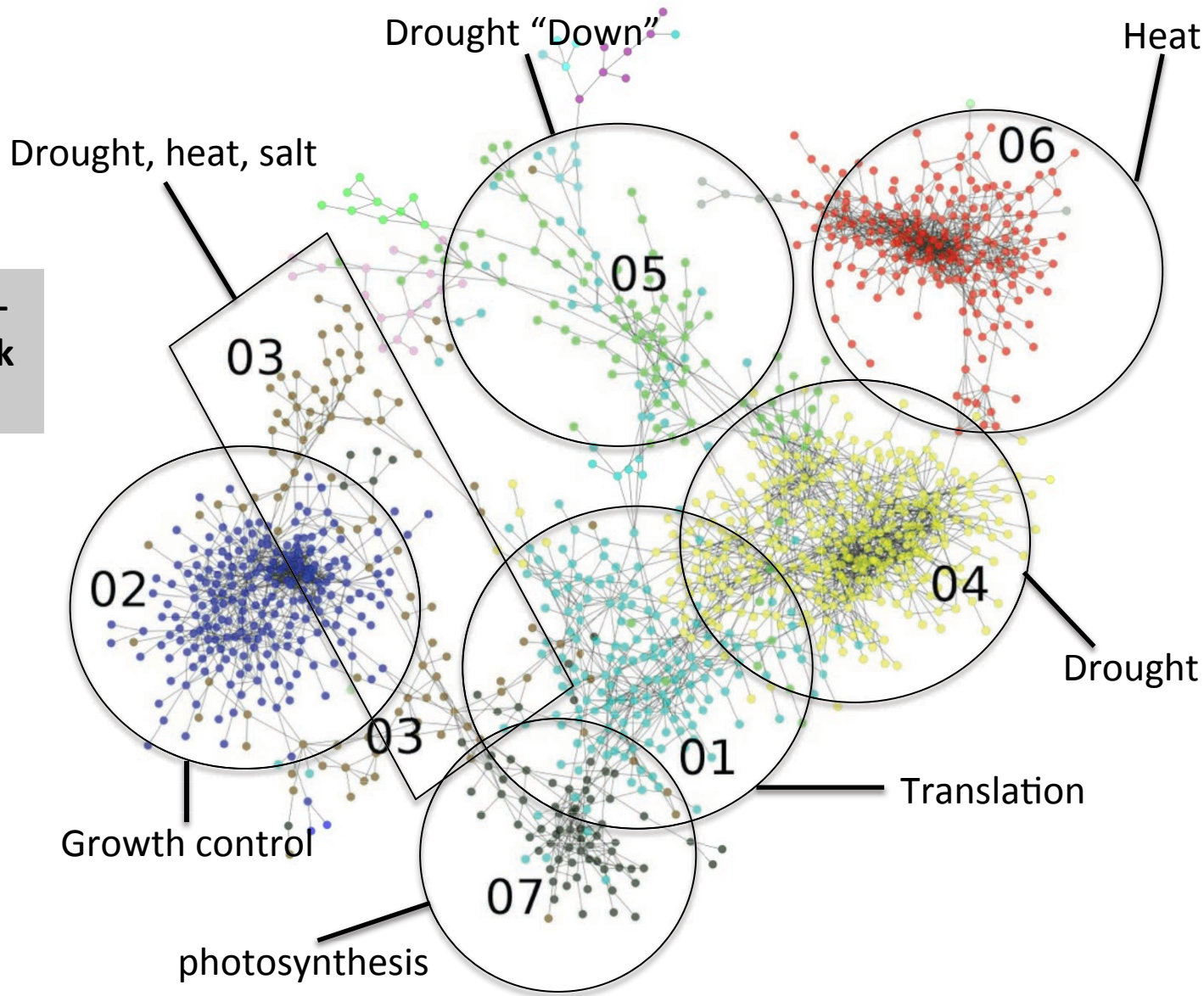
- MACS
- BS-Seeker2

Genome enabled plant systems

The landscape is mature with many options and developed user communities.

Example: Network analysis elucidates integration of stress responses, photosynthesis, and growth control (Brachypodium)

Weighted Gene Co-expression Network Analysis (WGCNA)



Priest et al., PLoS One, 2014

The Bottleneck Has Shifted

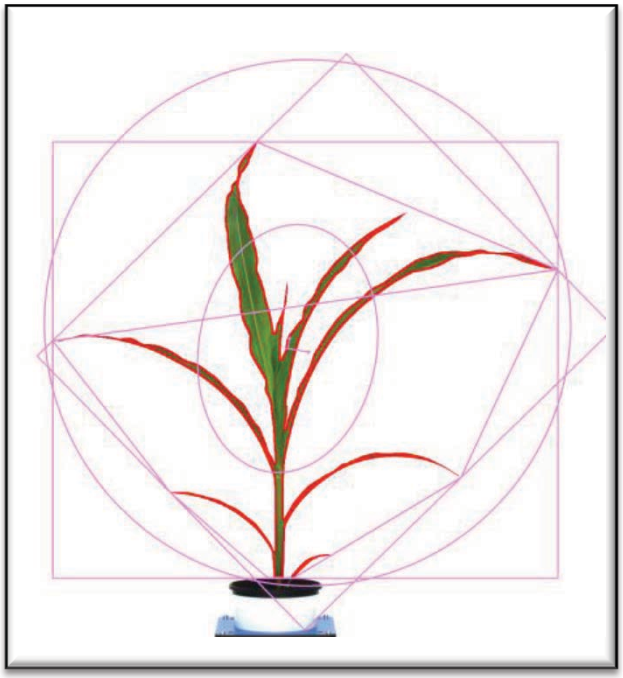
A systems-level analysis of drought and density response in the model C4 grass *Setaria viridis*

Tom Brutnell (PI); Co-PIs Andrew Leakey, Asaph Cousins, Ivan Baxter, Todd Mockler, Jose Dinneny, Sue Rhee, Dan Voytas, Hector Quemada

DOE Genomic Science Program - Award Number DE-SC0008769

Controlled environment phenotyping and analytics

High-throughput image-based controlled-environment phenotyping



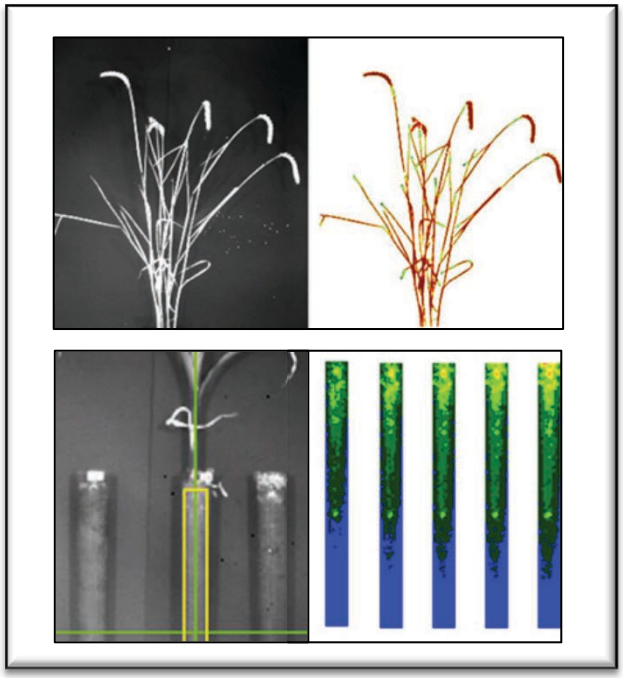
Visible Light Imaging

High-resolution color images for comprehensive morphological and growth phenotyping

Fluorescent Light Imaging

Blue light (< 500 nm) visualize any related fluorescence

- Chlorophyll (cont., flash)
- Green fluo. Protein (GFP)
- Phenolics
- Auto fluorescence



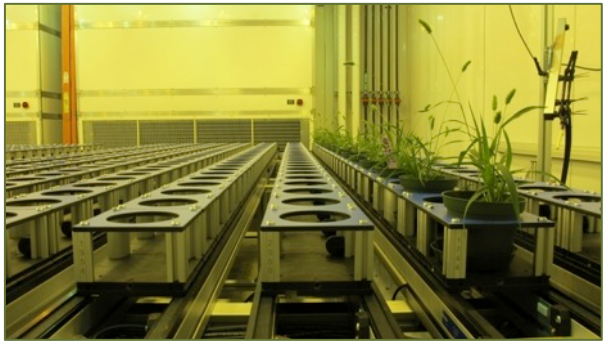
Near-Infrared Light (NIR) Imaging

Shoots: Measuring water distribution and dynamics

Roots: spatial distribution of water content in soil

Danforth's controlled-environment phenotyping facility

- 1200 plant capacity
- The first non-commercial high-throughput plant phenotyping facility in the US
- Imaging and software enable daily measurements of growth, 3D architecture, biomass accumulation, fluorescence and infrared signatures.



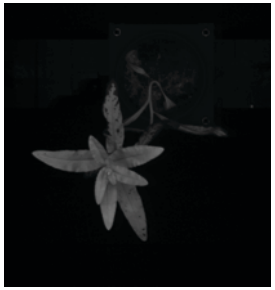
Conviron Growth House

Imaging Loop

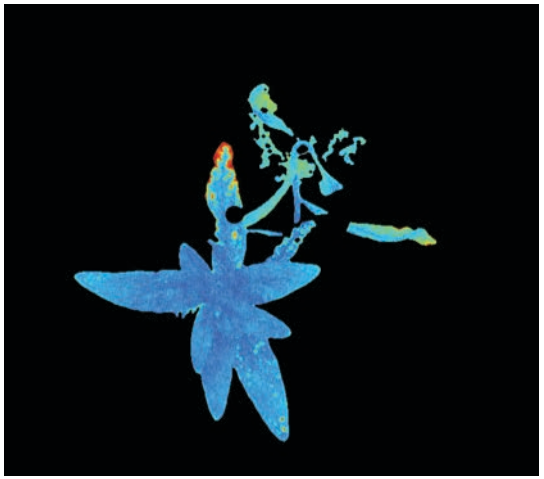
Three Imaging Stations: Fluor

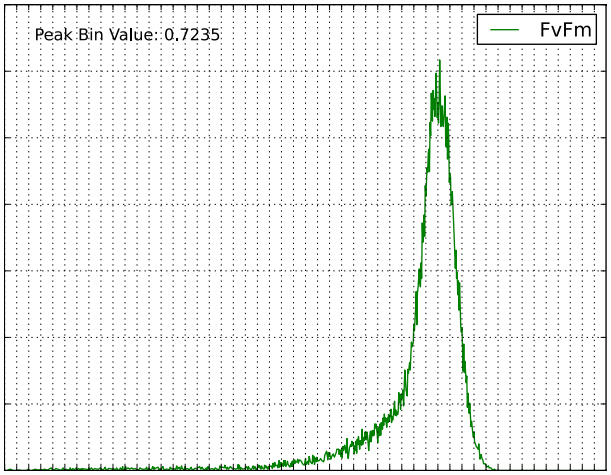
Fv

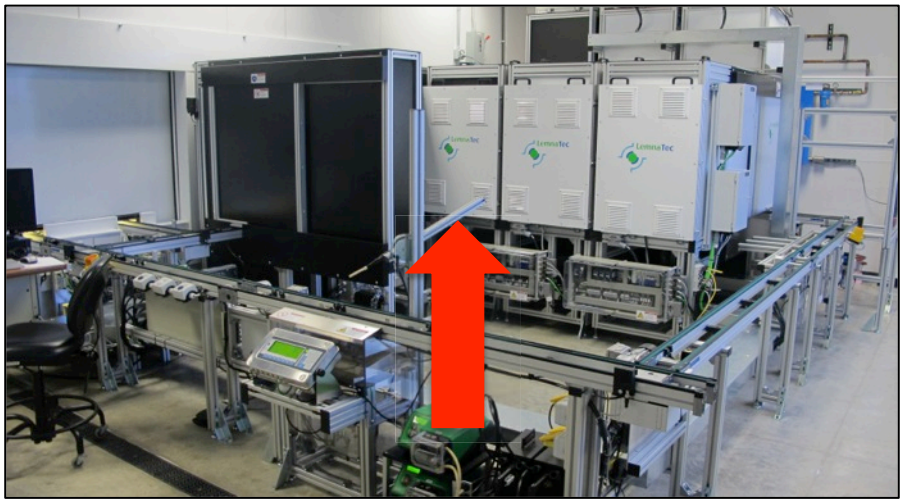
Fm



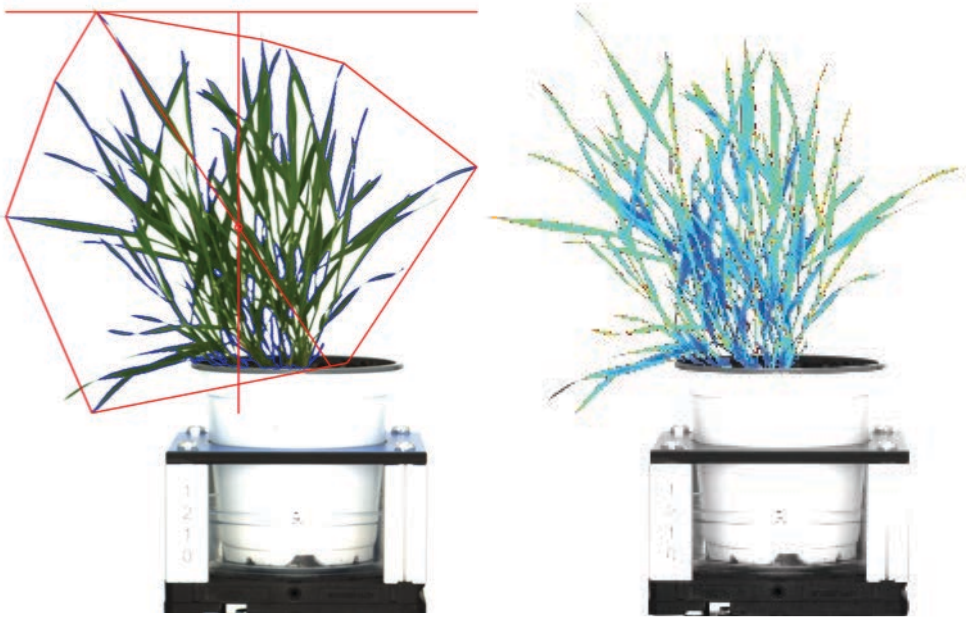
=



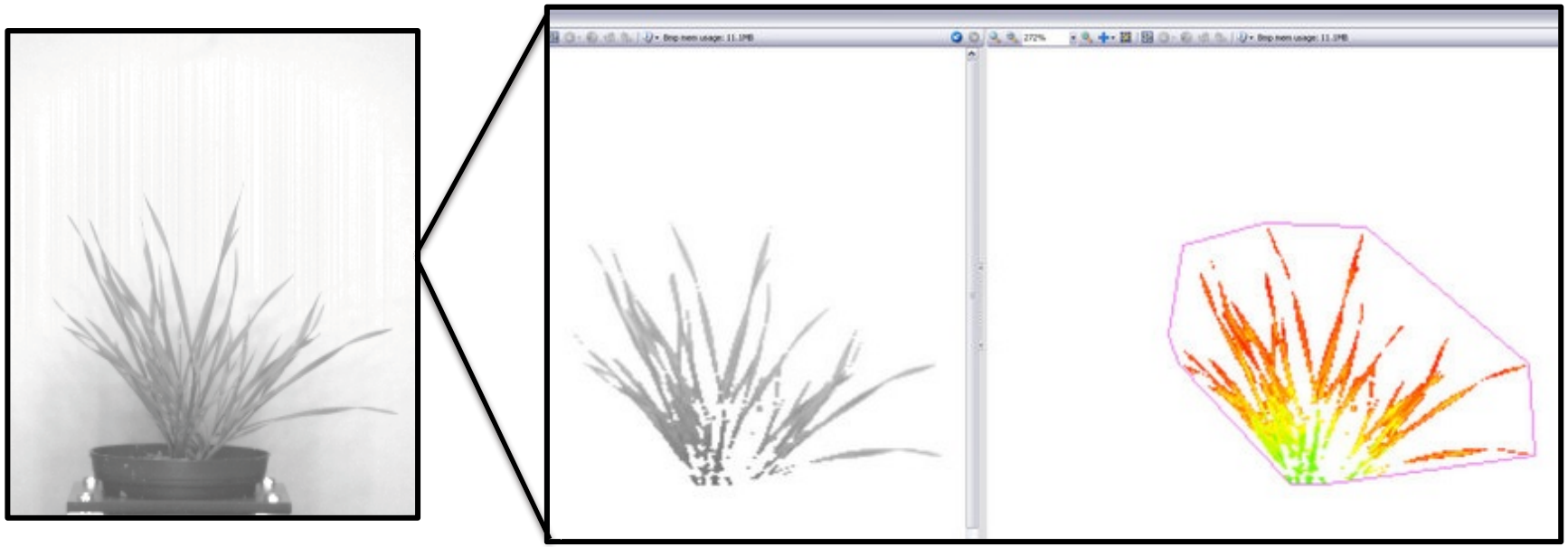




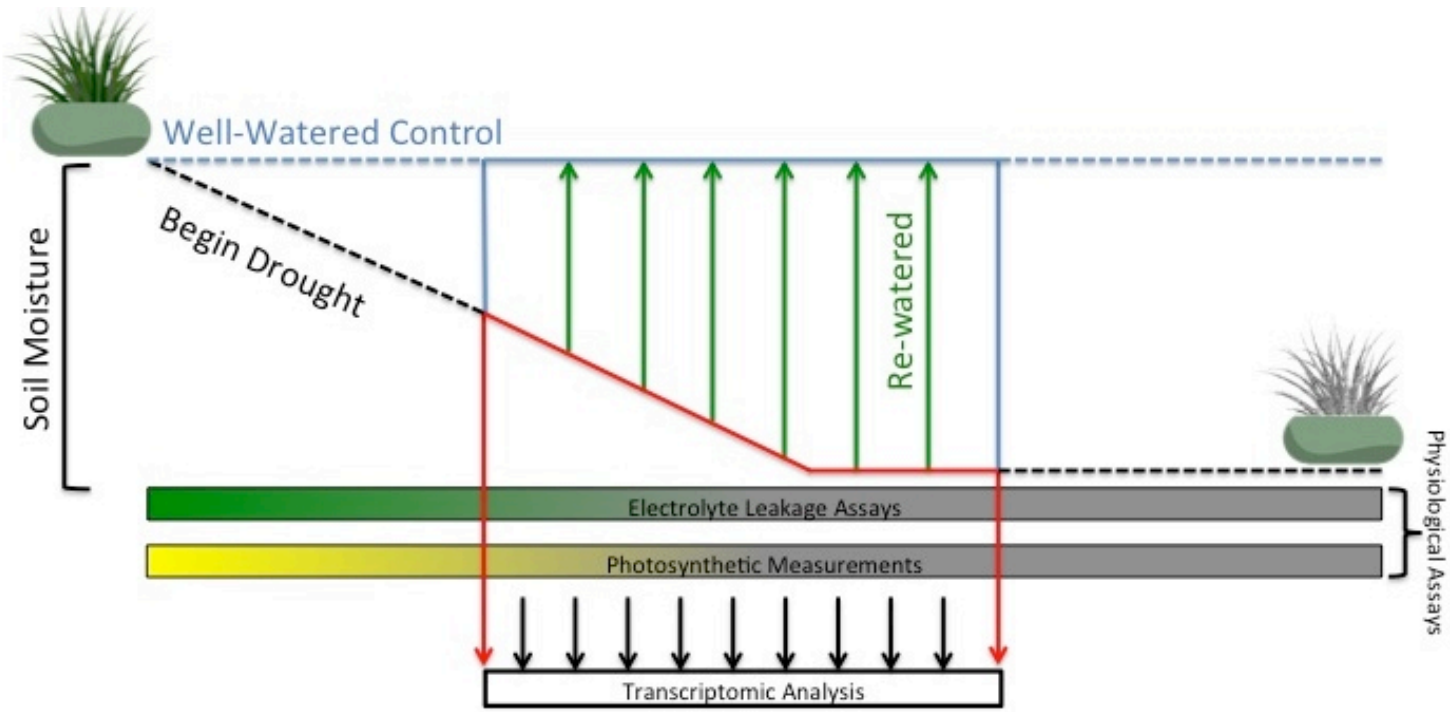
Three Imaging Stations: VIS



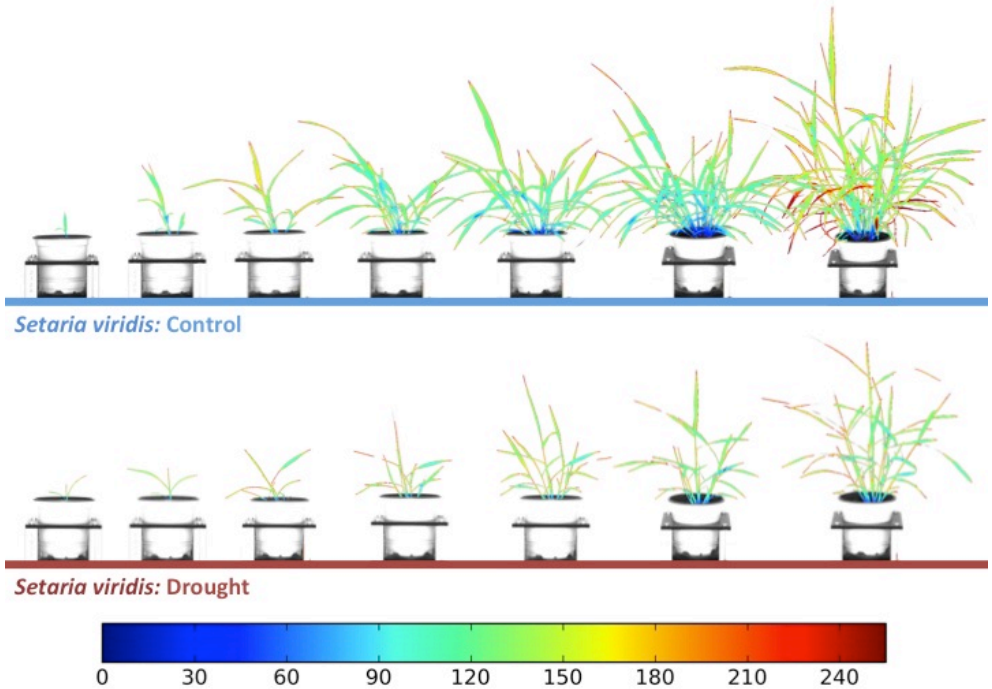
Three Imaging Stations: NIR



Controlled environment phenotyping enables coupling physiological analysis and 'omics profiling to understand plant-environment interactions and effects on growth and yield



Quantifying effects of drought on *Setaria viridis* growth, development, and physiology

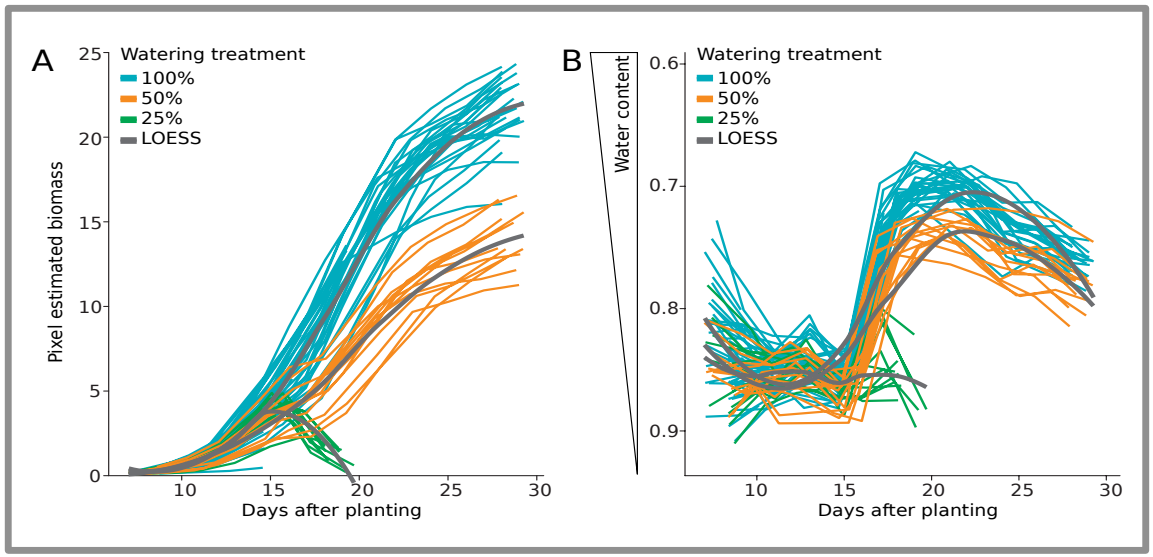


100% watering

50% watering

Visible Imaging:

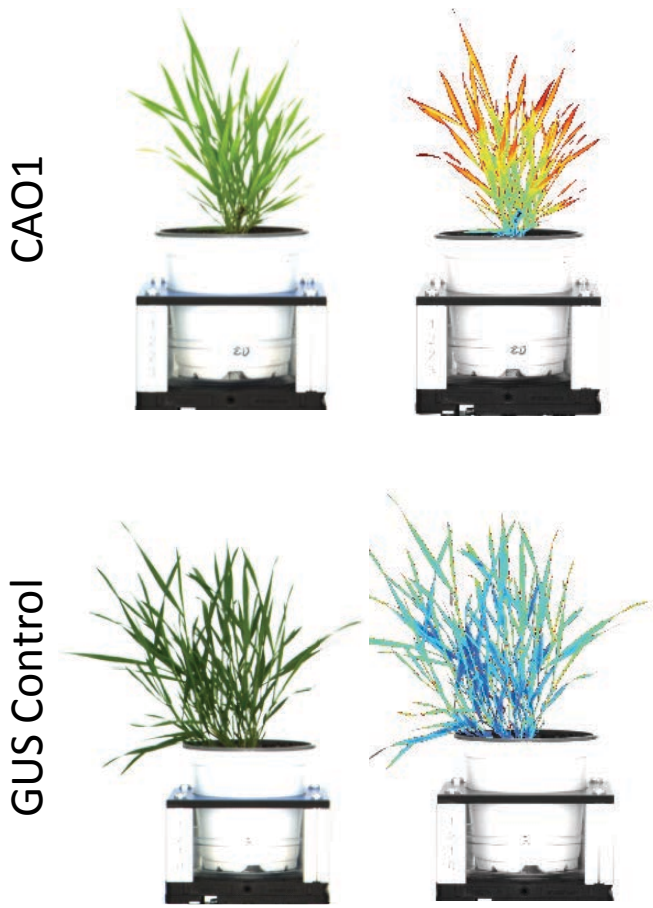
biomass

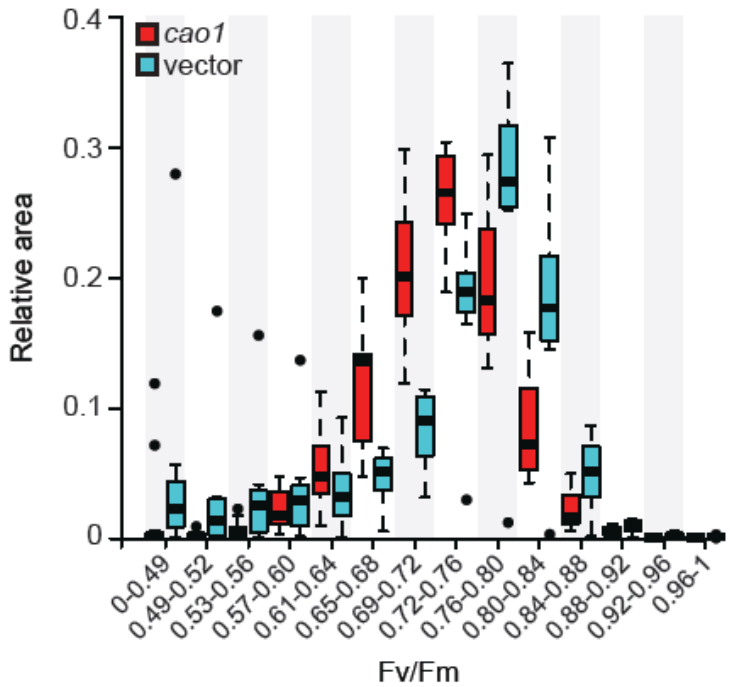


NIR Imaging:

water content

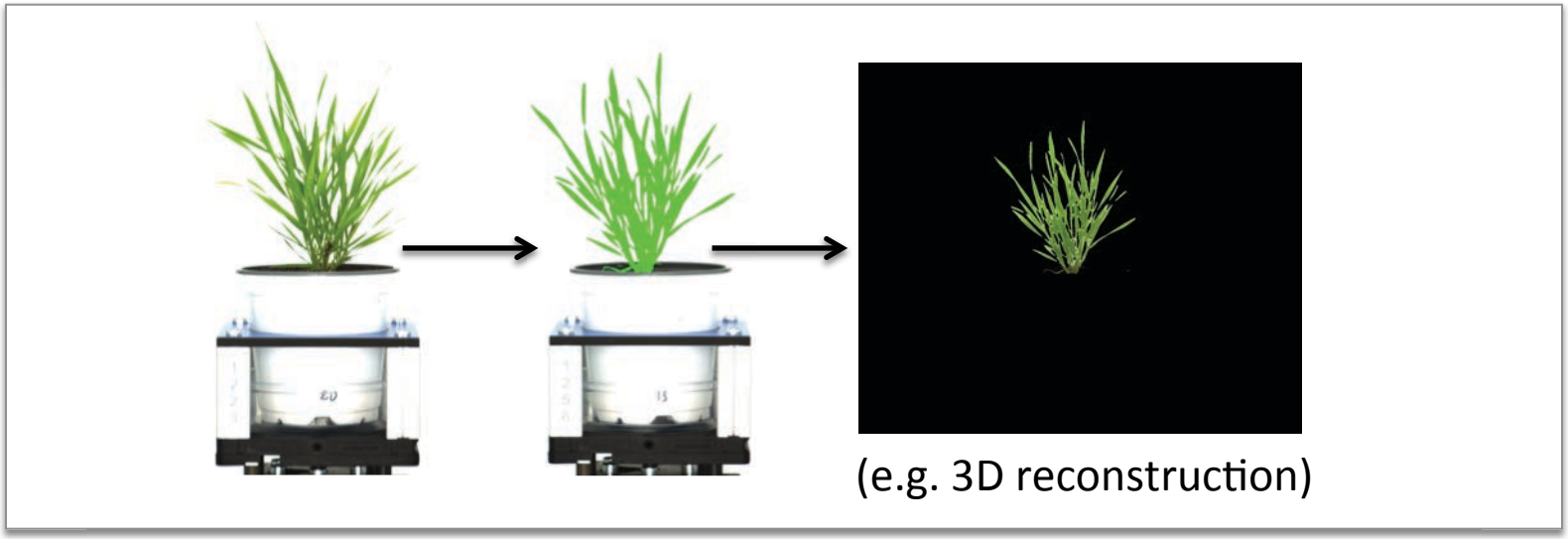
Fluorescence imaging-based photosynthetic analysis (benchmarking)





Lowered Fv/Fm ratio in *cao1* (chlorophyllide a oxygenase) is consistent with fewer open reaction centers available in this photosynthetic mutant.

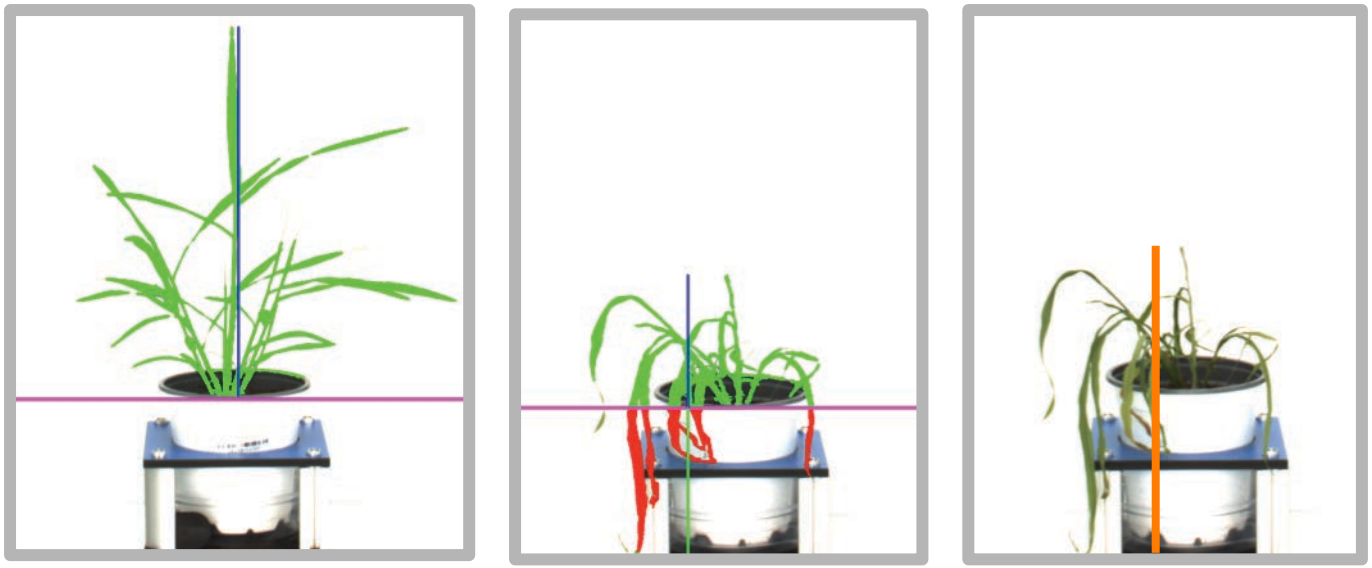
PlantCV (Plant Computer Vision)



- Built with opensource libraries OpenCV, Numpy and MatPlotLib
- Integration of plant identification, feature identification, and feature analysis
- Works on other types of images, not LemnaTec-specific
- Will be published and released publicly - soon

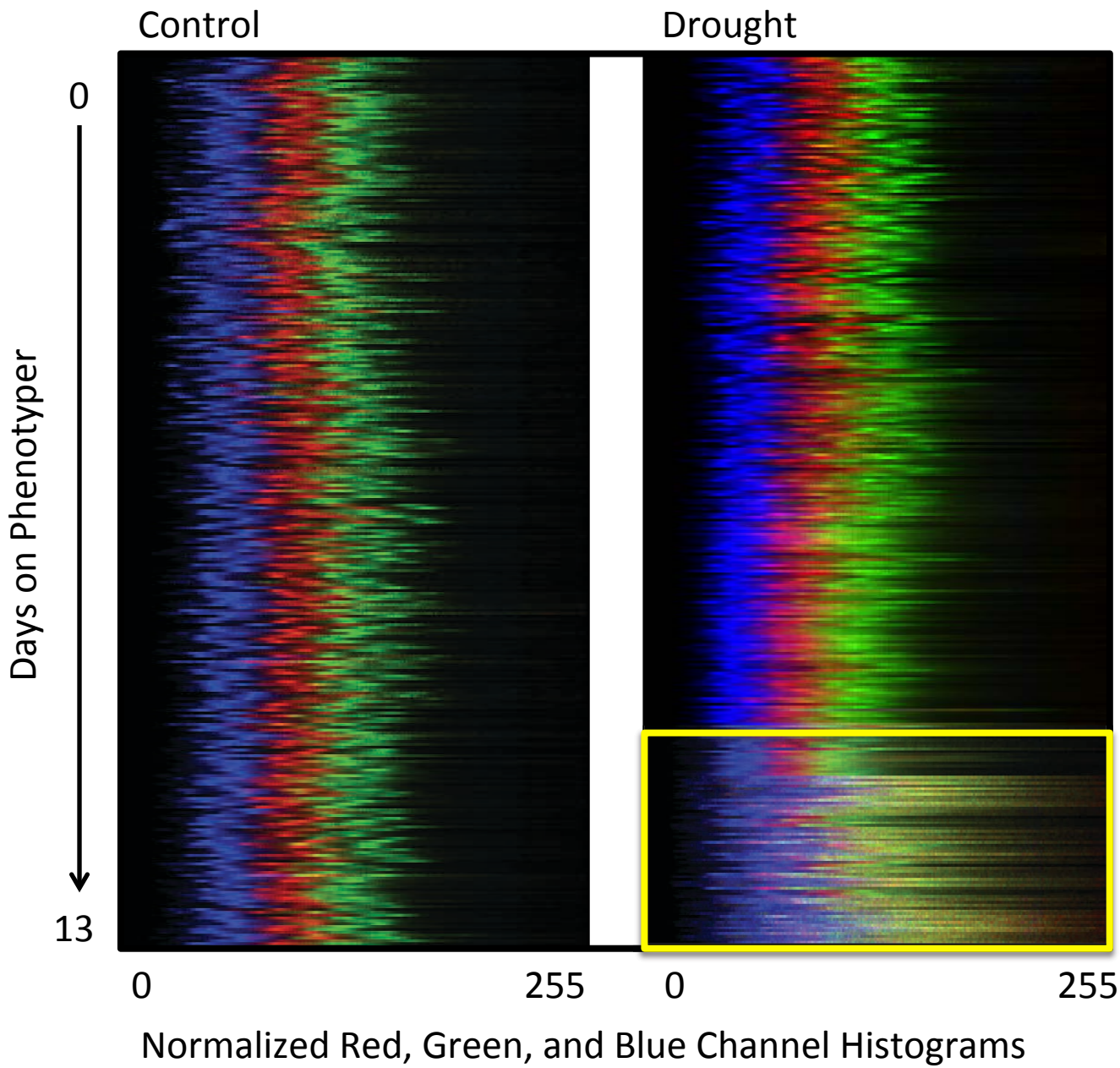
Being developed jointly by Mockler, Carrington, Baxter, and Brutnell groups at Danforth

Proxy measurements for more complex traits of interest



Automatic wilting assessment as a proxy for drought tolerance

Setaria viridis: Visualization of color spectrum histograms to develop drought signatures



Where do we need to go?

Where do we need to go – genomics/genetics?

- Sequencing technology and bioinformatics are well established – and not a real limiting factor now (but should be tightly integrated into improvement efforts).

- For some potential target species, existing 'omics resources are significant but insufficient. Sequence data acquisition is still needed (diversity collections, understanding epigenomic impacts).

- For example, existing gene expression profiling resources in sorghum limit the quality and utility of gene network models that can drive hypothesis-driven allele-specific breeding strategies.

- Investment in data organization/coordination and integration is needed in specific crops – generic databases don't address this issue for specific crops

Where do we need to go – phenomics/analytics?

- More plant phenotyping data needs to be acquired in order to inform the development of improved analytics.
- It's not all about image-based phenotyping - molecular phenotypes matter (e.g. gene expression, metabolomics).
- Algorithms and software are needed to automatically extract and/or derive digital plant phenotypes from large, typically image-based datasets.
- Hyperspectral imaging seems to have great promise – but we don't have off-the-shelf hyperspectral profile models that correlate with key plant traits.
- Robust packages/pipelines and communities of users are needed to drive standardization.
- Investment in data organization/coordination/curation and ongoing integration.