Resequencing for SNP Identification and Expression Profiling in Crop Species

National Center for Genome Resources

Cornell University

Mizzou
University of Missouri

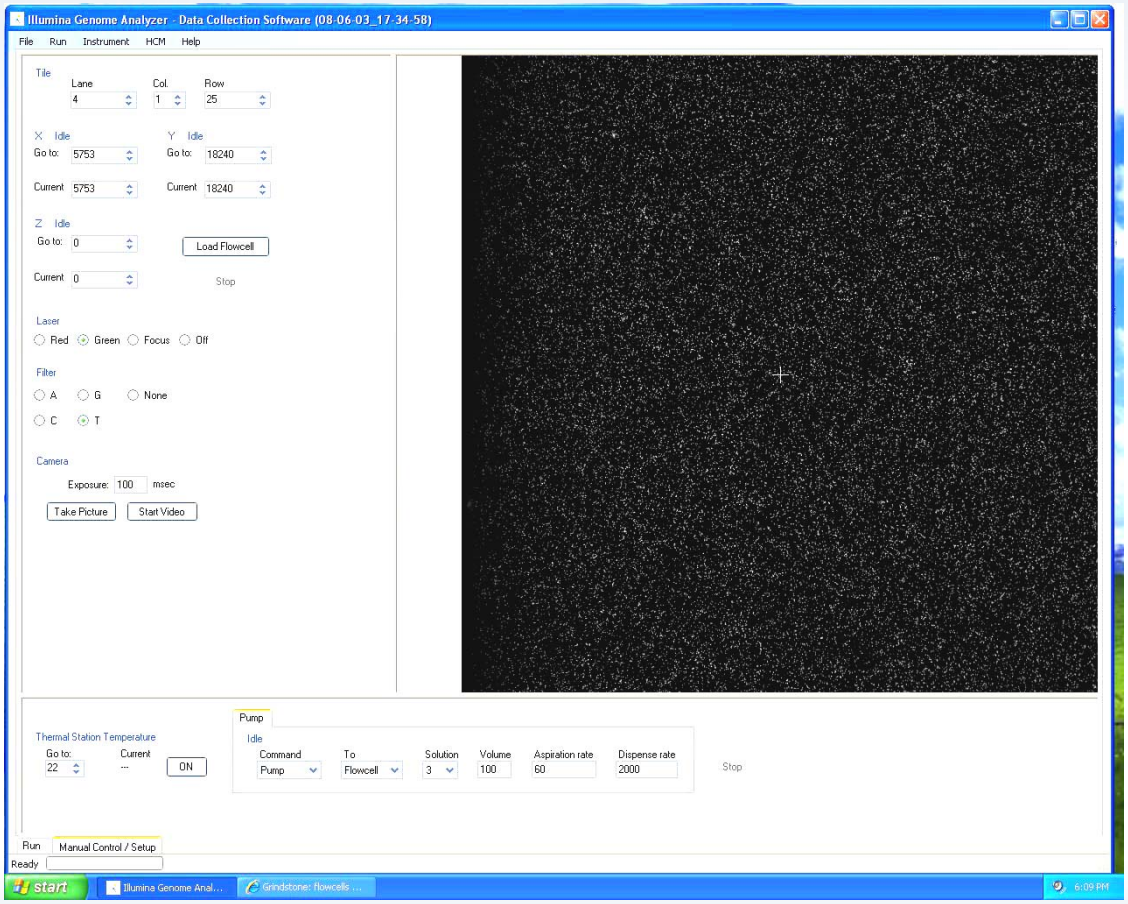
PURDUE
UNIVERSITY

Ryan W. Kim
rwk@ncgr.org

NCGR

Genome Center Instrumentation and Software Systems

- **Eight Illumina Genome Analyzer IIs (six IIxs)**
- **Six Genome Analyzer Paired-End Modules:**
- **Four GA Cluster Stations**
- **Data Management:** NCGR-developed LIMS and Alpheus variant and expression analyses system (Neil Miller et al)
- **Additional instrumentation:** As triggers met – Project dedicated instruments provide up to three short runs per week, two long run per month to that project.
- **Genotyping**



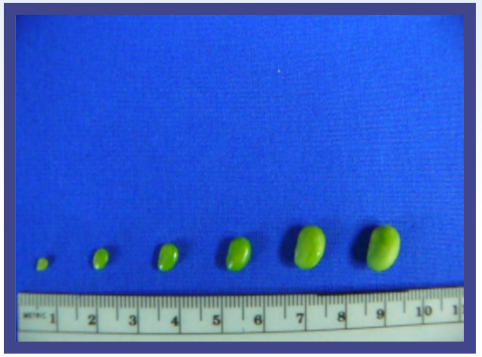
Genome Center Computational Infrastructure

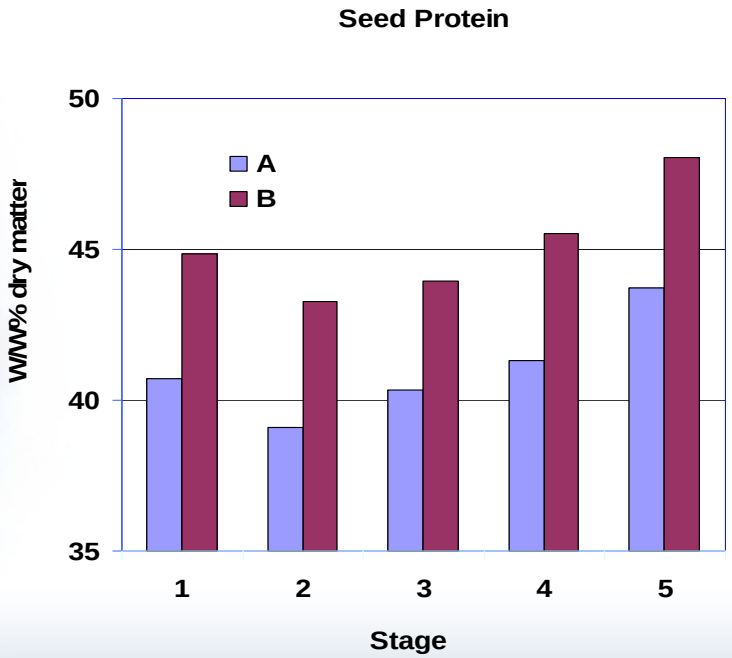
- **Raw data storage:** Sun X4500 - A Dual Processor, Dual Core 24 TB file Server.
- **Image processing/Base calling:** Sun X4150 – Two 4-core processors, 8G RAM, four 146G Hard drives
- **Variant detection:** Four Sun Blade 6000 chassis with ten 6220 Blades, each blade has two dual core processors with two 146G disks and 16G RAM. A total of 160 Cores, 640G of RAM and 11.6 TB disk.
- **DB storage:** Sun V490 DB Servers, 6140 SAN with 40 TB for DB disk space, 4 Gb/s Switch fabric

- **New hardware added in September:**
- Sun x4500 24 Terabyte file server, (4) 2.6Ghz cores, 16G RAM, (48) 500G Disks
- Sun x4540 48 Terabyte file server, (8) 2.3Ghz cores, 32G RAM, (48) 1000G Disks
- Sun x4600 Assembly server with (4) 3.0Ghz cores and 128G RAM, (4) 146G Disks
- Sun Blade 6000 chassis Giving us 50 total blades with 16G RAM, and (2) 146G Disks each
- Four Sun x4240 DB server with (16) 2.6Ghz cores, 64G RAM, (16) 146G disks
- Two Sun x4450 DB Servers with (16) 2.92Ghz cores 128G RAM and (8) 146G disks
- Two Sun x4450 with (16) 2.6 Ghz cores, 64G RAM, (8) 146G disks
- Two Sun x4150 with (8) 2.3 Ghz cores, 8G RAM, (4) 146G Disks

Whole Transcriptome Shotgun Sequencing of Soybean High vs Low Protein Lines

- Four seed developmental timepoints
- From each of two soybean NILs
- Each exhibiting either high or low protein content



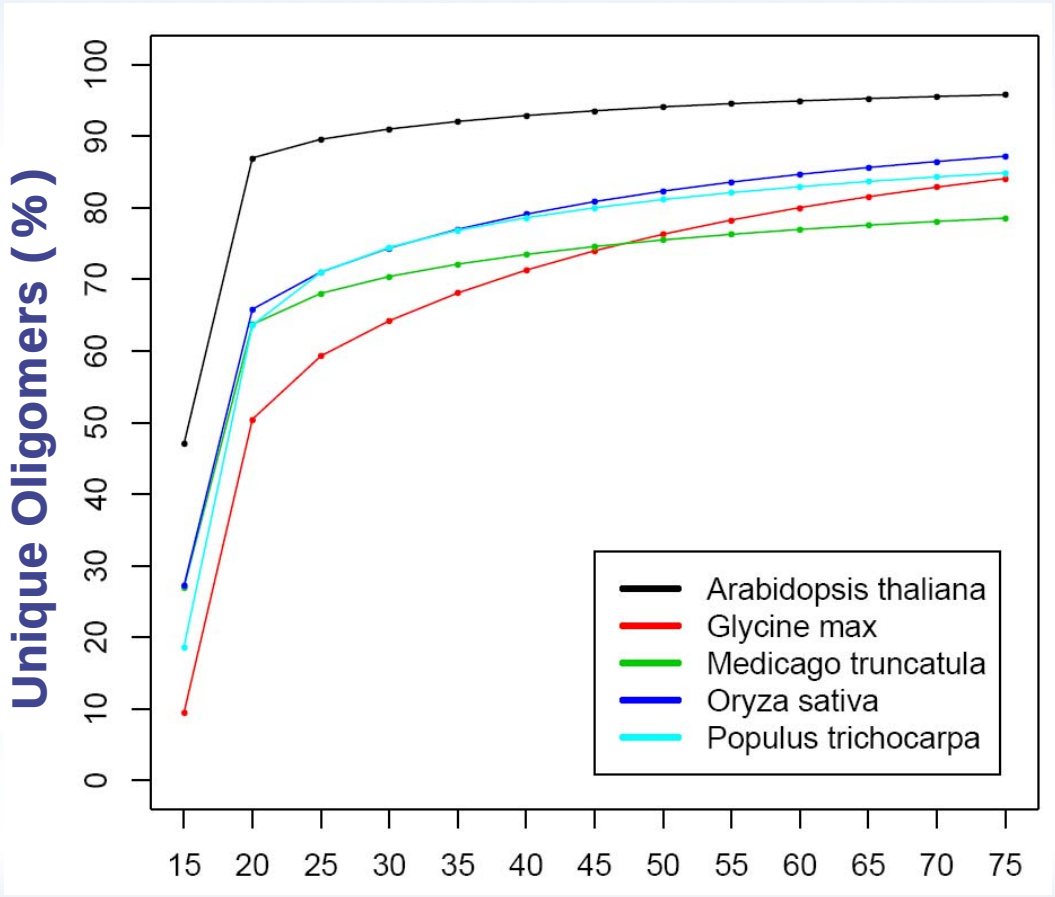


UNIVERSITY OF MINNESOTA

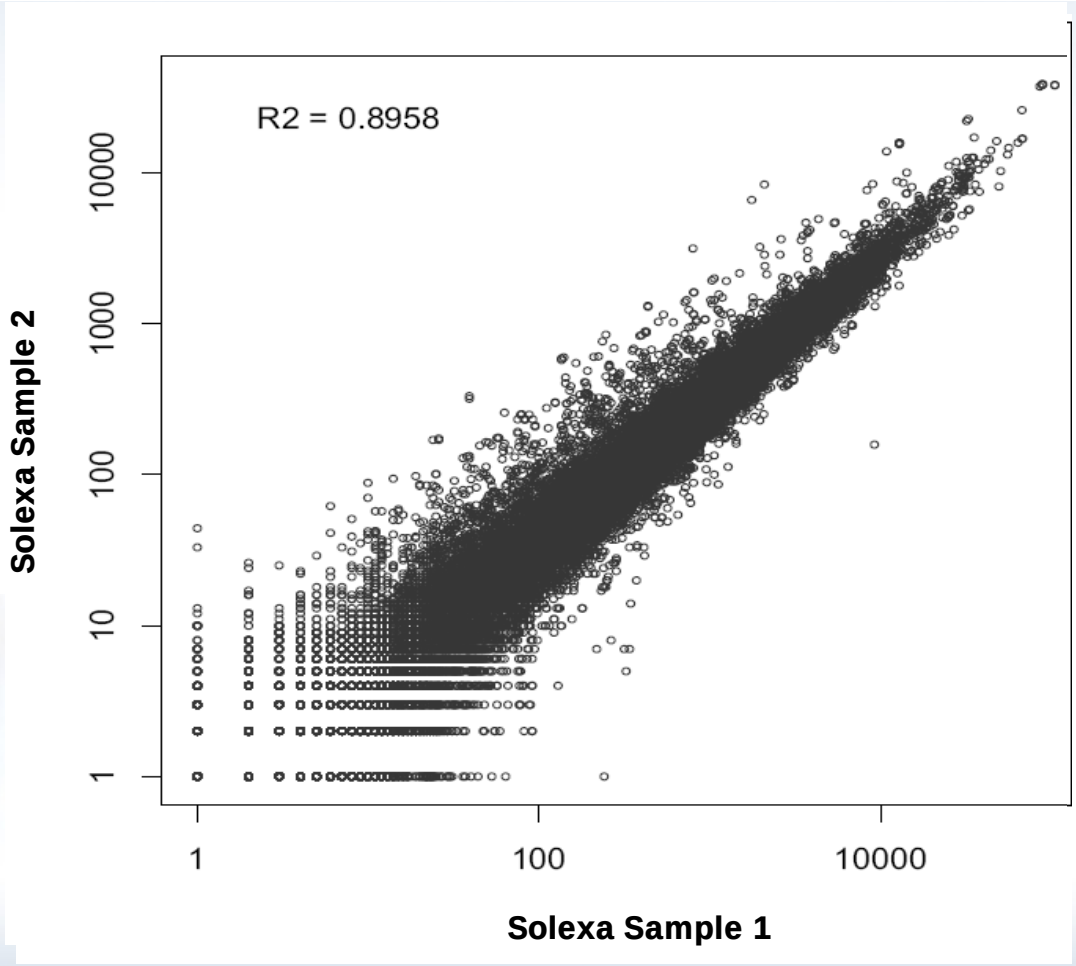
IOWA STATE UNIVERSITY

NCGR

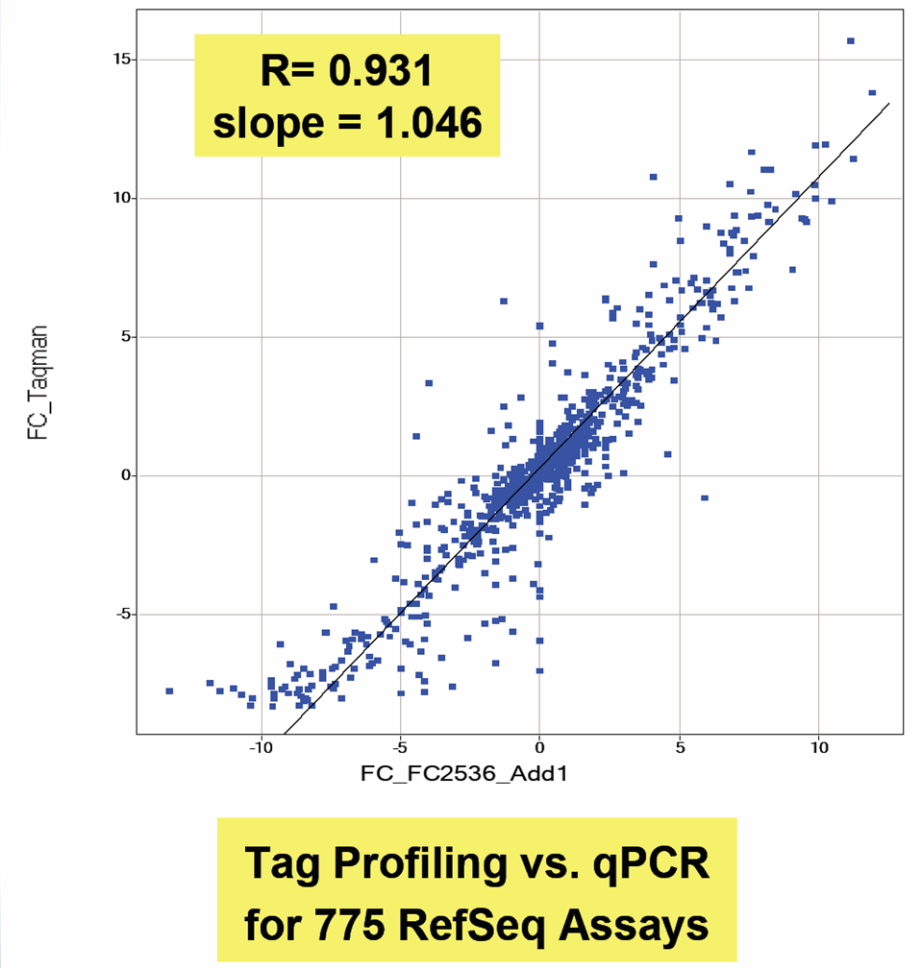
Uniqueness of Oligomers in Plant Genomes



The Number of Reads Aligned to Each Transcript Varies Little from Run-to-Run



Comparing qPCR with Illumina WTS



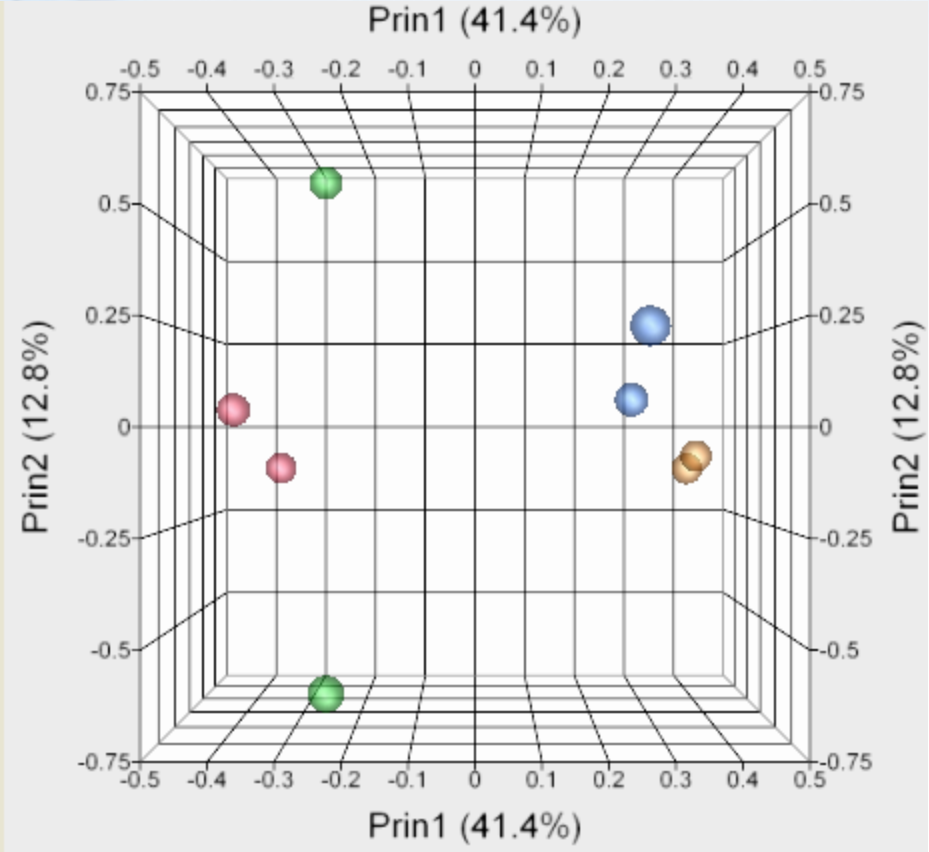
Tag Profiling vs. qPCR for 775 RefSeq Assays

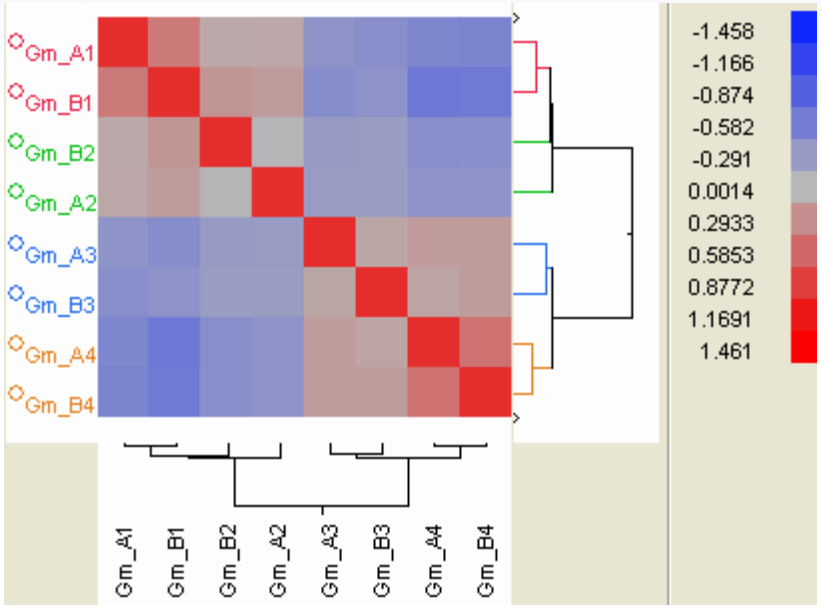
Courtesy G. Schroth, Illumina

Data Summary and Alignment to the JGI 7x Soybean Genome

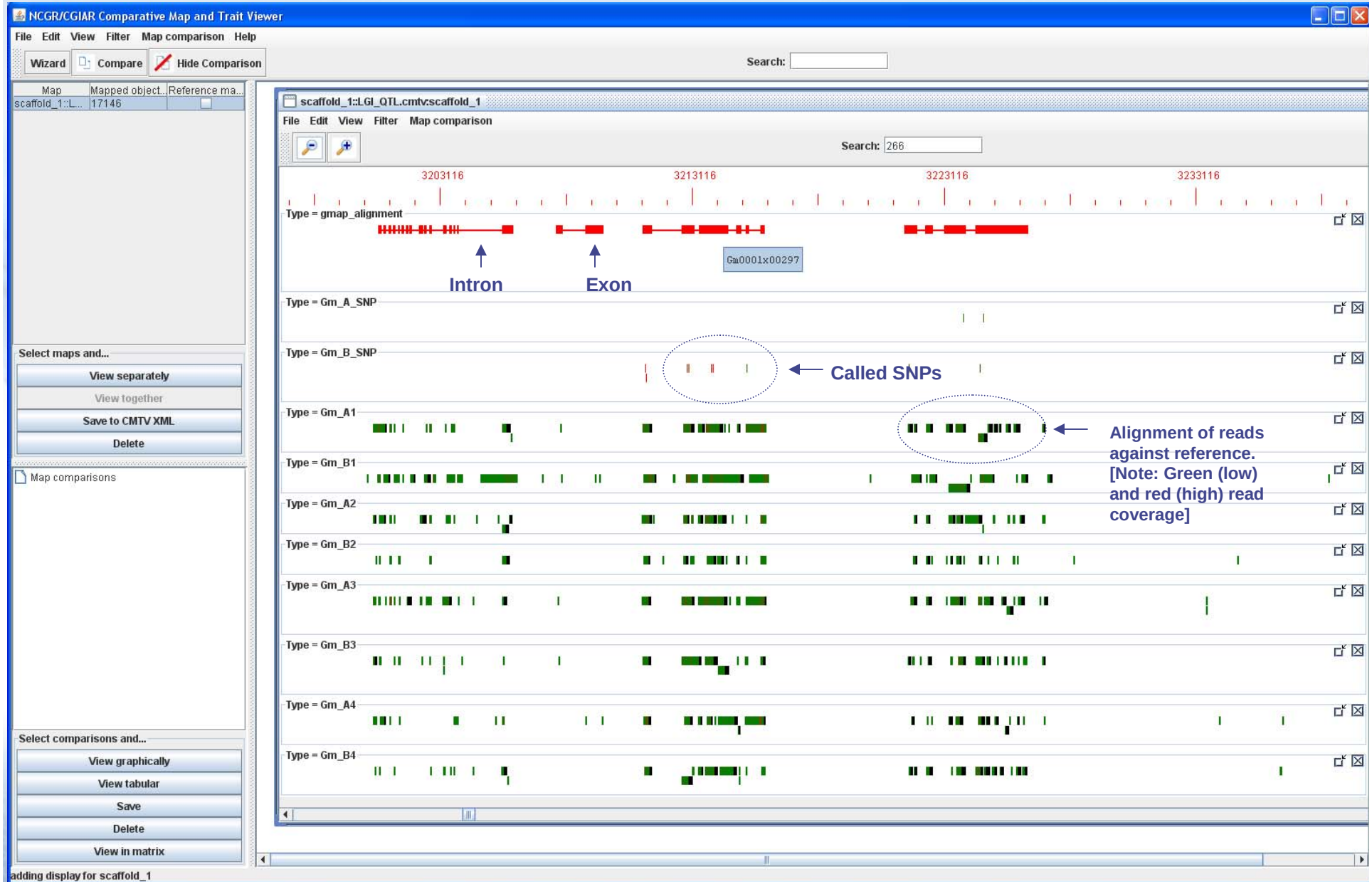
Sample	Reads	Reads Aligned	Reads Uniquely Aligned	Total SNPs	In/Dels
Gm_A1	7,835,063	5,808,687 (74%)	2,909,296 (37%)	1,374,602	4,429
Gm_A2	9,673,118	7,490,621 (77%)	2,662,020 (28%)	1,115,564	6,081
Gm_A3	9,102,649	6,700,125 (74%)	4,788,833 (53%)	1,502,187	4,418
Gm_A4	7,052,993	5,339,938 (76%)	3,519,076 (50%)	915,889	3,457
Gm_B1	16,988,687	13,394,686 (79%)	7,783,311 (46%)	2,610,326	8,545
Gm_B2	7,950,528	6,296,363 (79%)	2,181,665 (27%)	885,383	5,128
Gm_B3	9,201,789	7,400,210 (80%)	4,106,899 (45%)	1,014,933	5,677
Gm_B4	8,909,676	6,831,318 (77%)	4,161,016 (47%)	1,011,387	5,316

PCA of Loess Normalized, Log Transformed Unique Read Counts

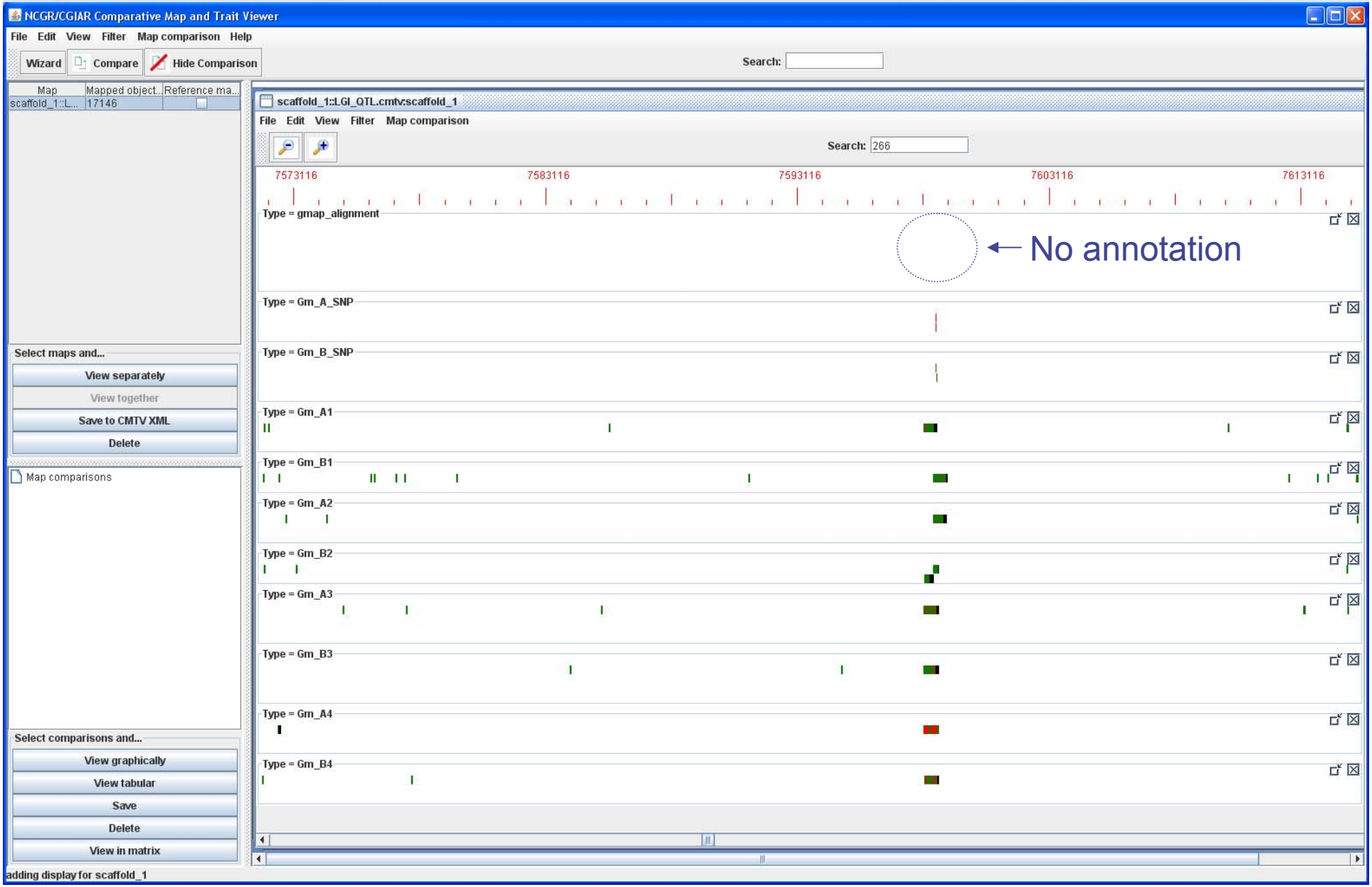




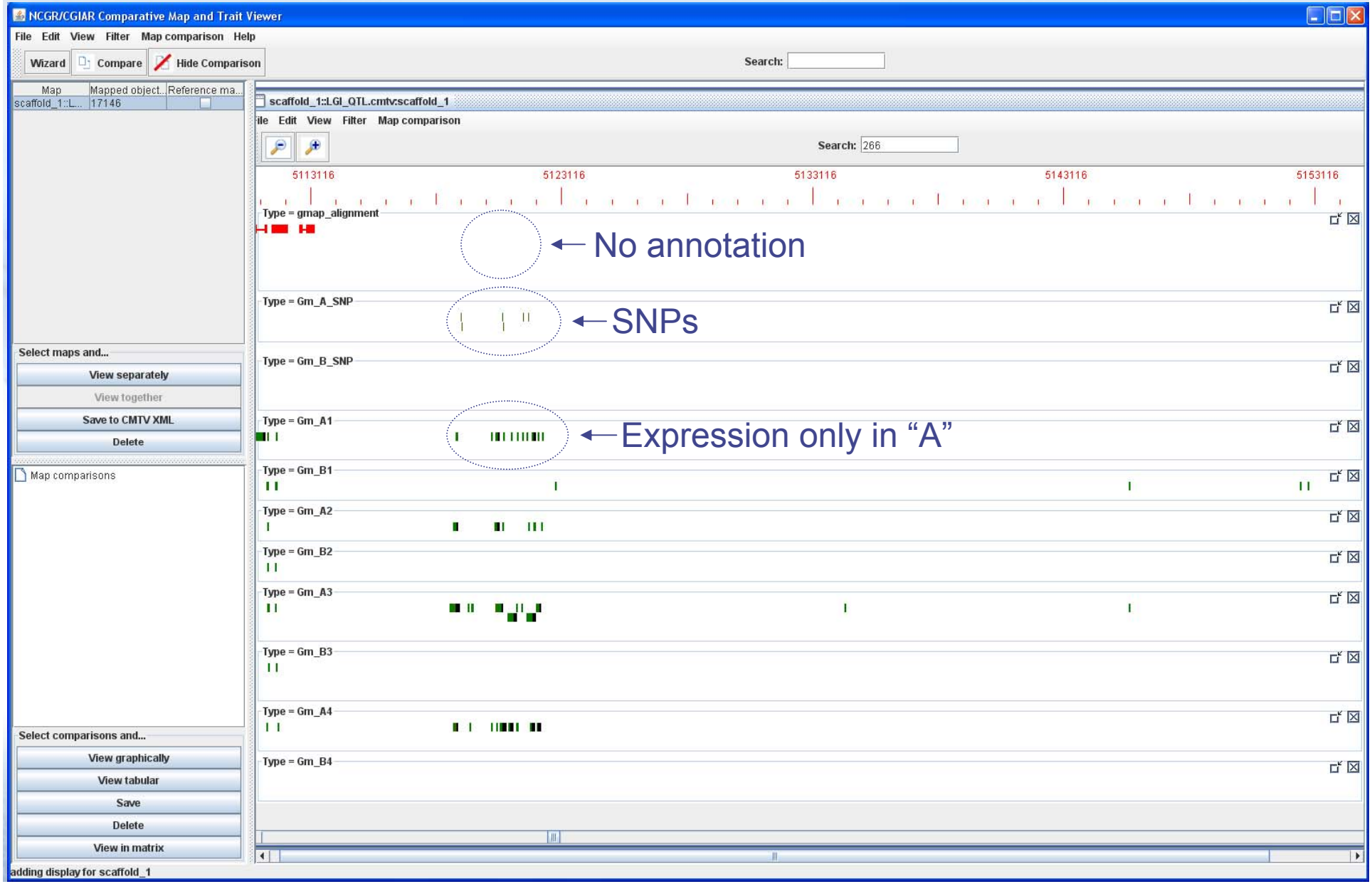
Alignment of Solexa Reads Against JGI 7x Soybean Genome



No gene annotated, but all libraries suggest transcribed region.



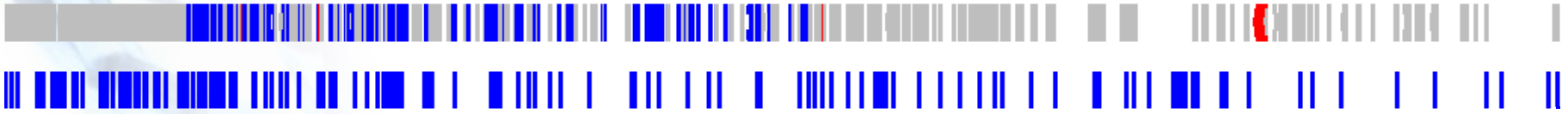
No gene annotated, Expression only in Line “A”



Gene expression within the linkage group I QTL region

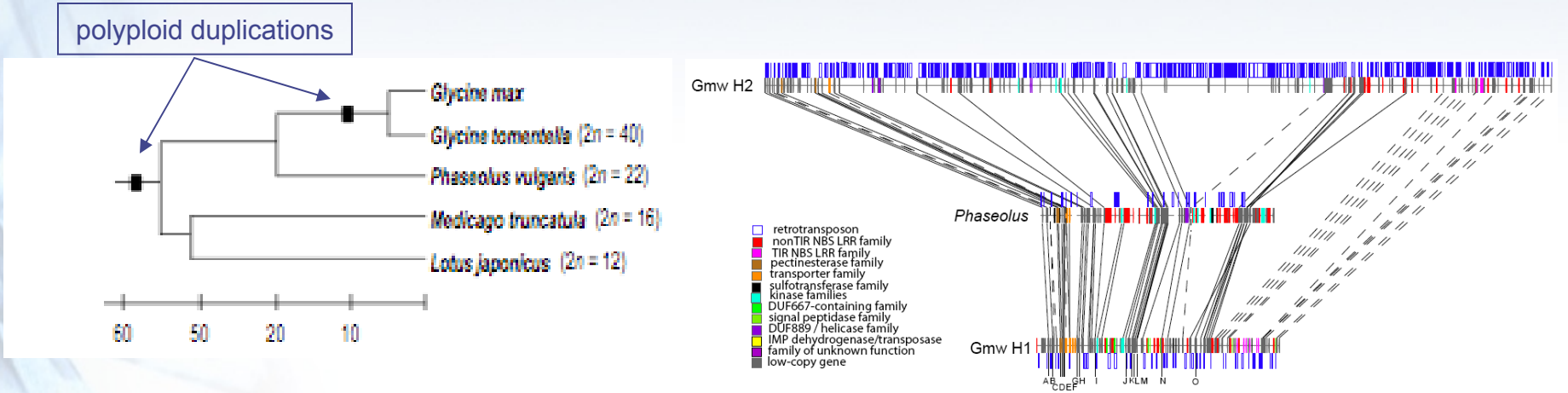
- Four differentially expressed genes
- 124 novel “exons” (An additional 13,300 outside of QTL region).

Known Genes



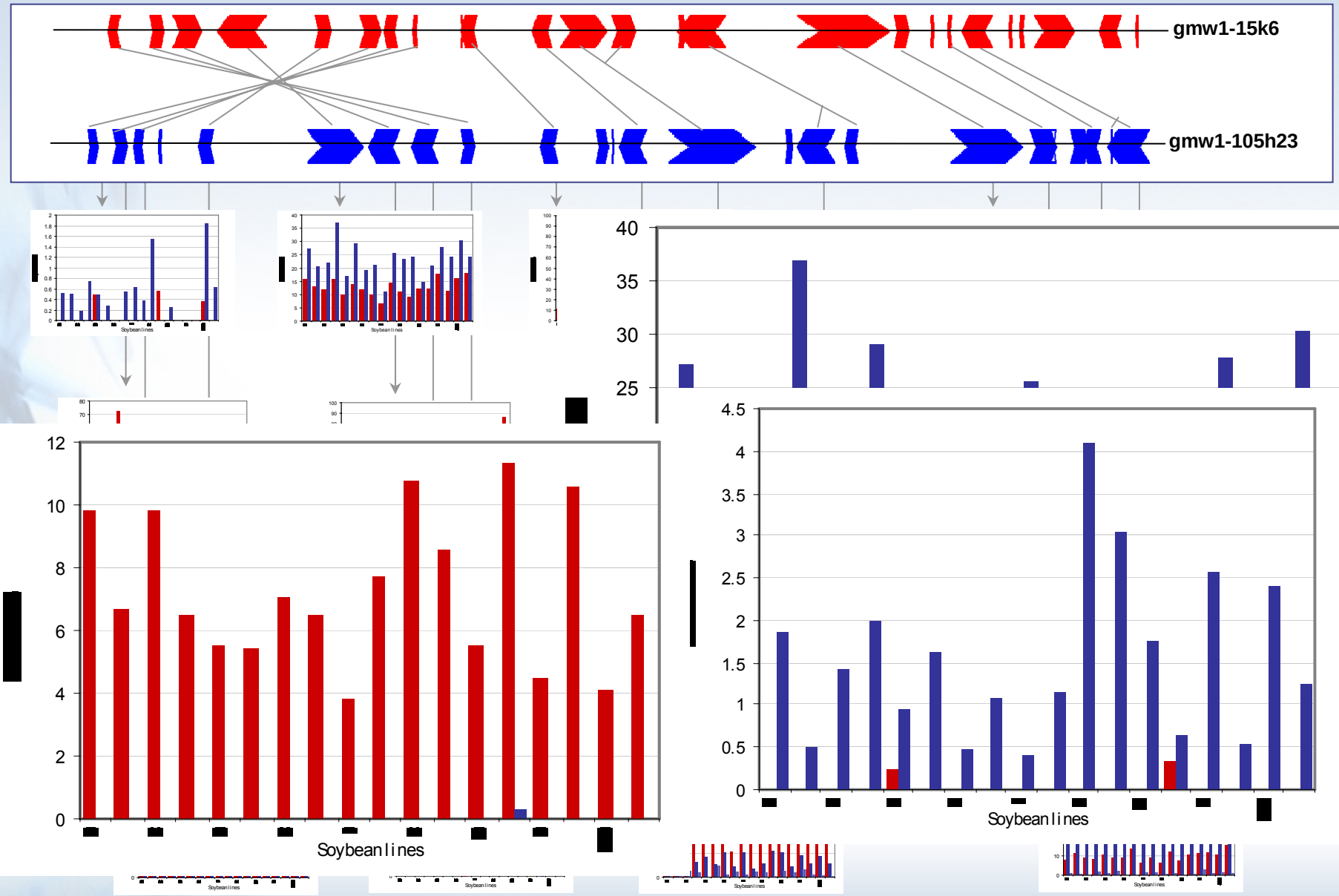
Novel Exons

Glycine, including soybean (*G. max*) is fundamentally polyploid

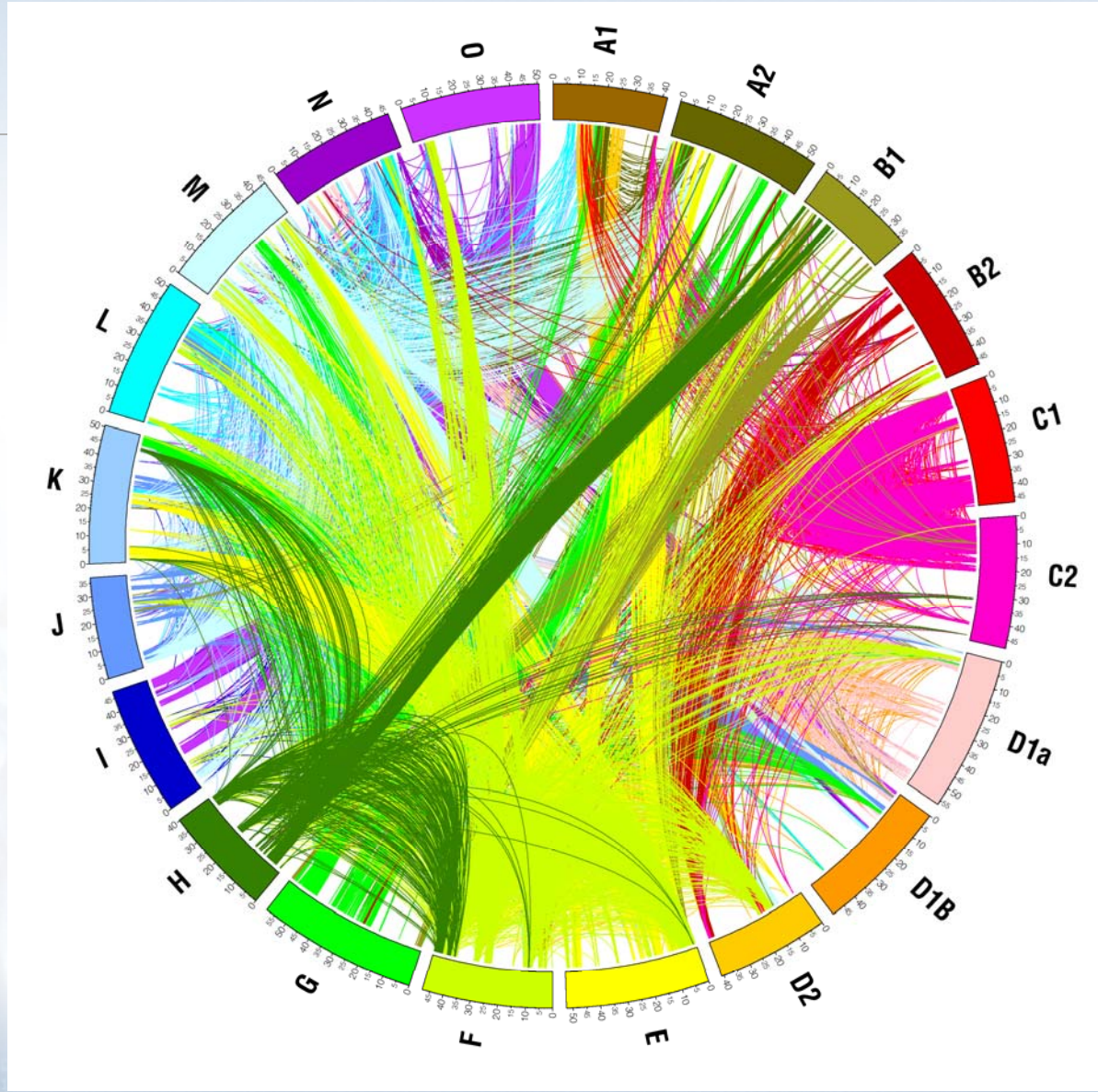


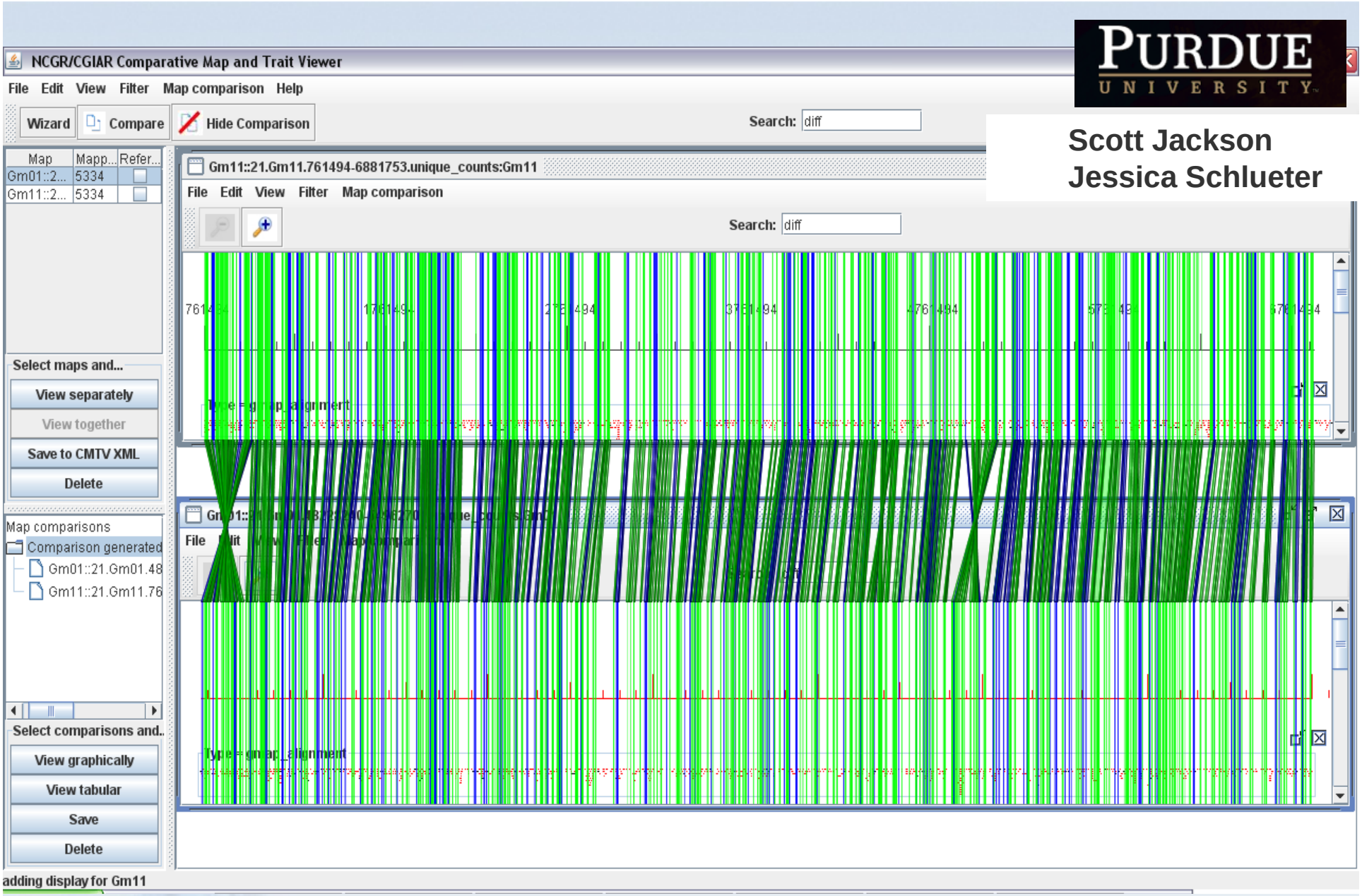
Innes, R., et al. 2008. Differential accumulation of retroelements and diversification of NB-LRR disease resistance genes in duplicated regions following polyploidy in the ancestor of soybean. *Plant Physiology* (in press).

Homoeologous Soybean BACs



Duplicated regions in the soybean genome





Lines indicate homoeolog pairs.
Blue lines indicate inter-homoeolog differential gene expression

Scott Jackson
Jessica Schlueter

Soybean Illumina Data Overview

Mizzou
University of Missouri

Gary Stacey

Sample	Reads	Reads Aligned	Reads Uniquely Aligned
Apical Meristem	6,477,456	5,669,285 (88%)	3,885,504 (60%)
Flower	5,176,140	4,275,733 (83%)	3,322,166 (64%)
Green Pods	4,183,024	3,694,647 (88%)	1,442,311 (34%)
Leaves	5,290,196	4,740,401 (90%)	2,772,753 (52%)
Nodule	6,335,851	5,265,596 (83%)	3,386,046 (53%)
Root	6,107,312	5,154,176 (84%)	3,740,765 (61%)
Root tip	5,154,499	4,212,283 (82%)	3,187,659 (62%)

Read length: 36
Average read quality: 37-39
Number of gene matches: 43,455-49,992 in JGI 8X Genome

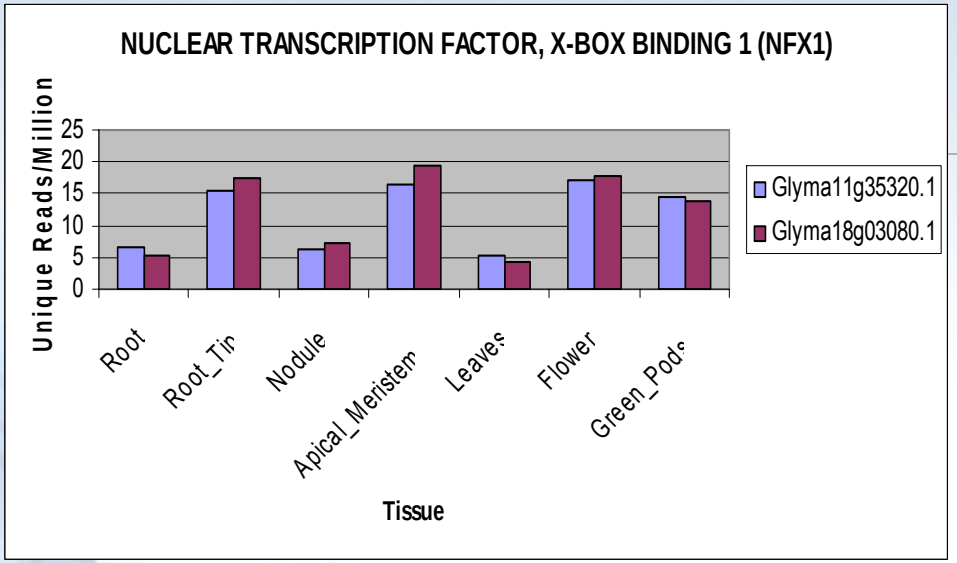
NCGR

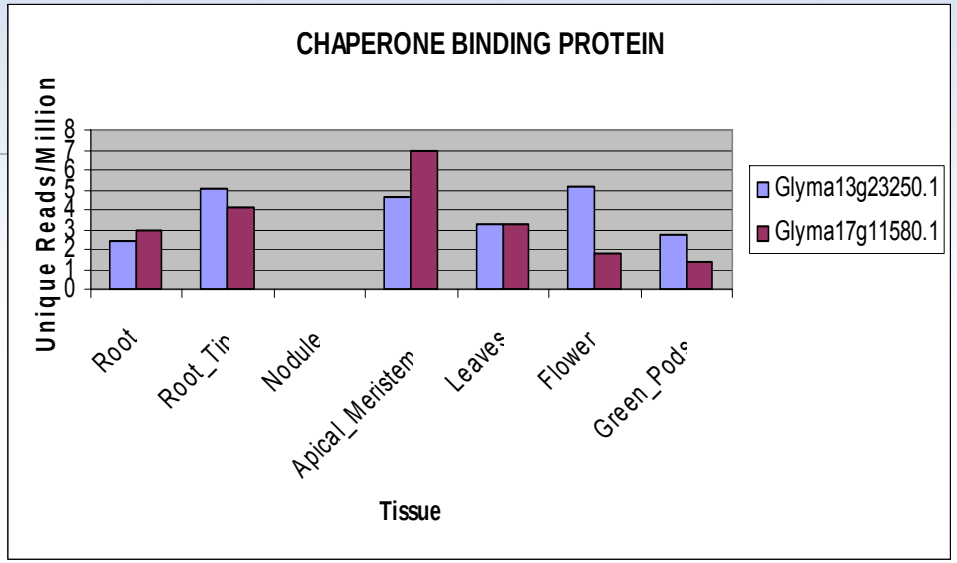
Inter-Homoeolog Differential Expression

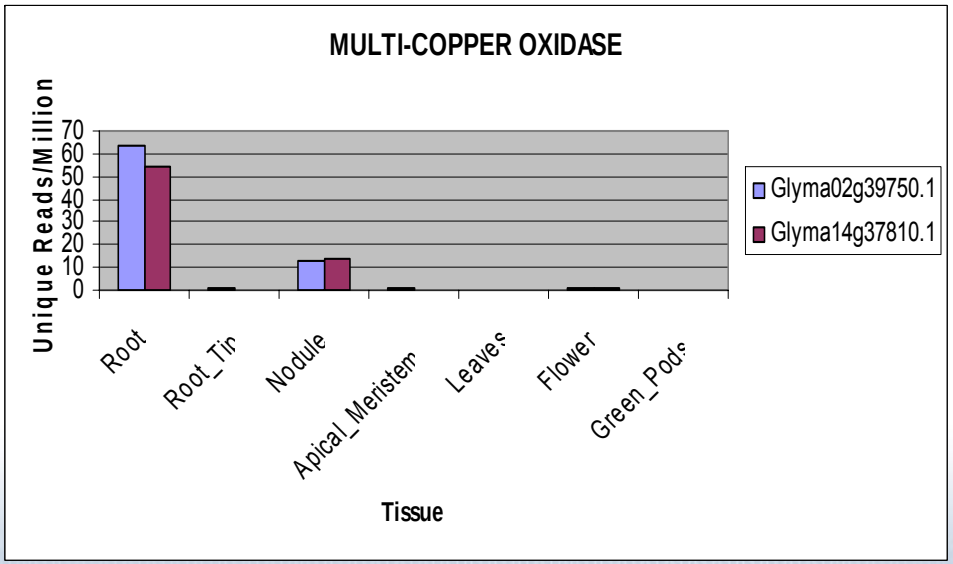
Tissue	p=0.05	p=1e-10
Root	7,340	2,168
Root Tip	6,570	1,729
Nodule	6,675	2,031
Apical Meristem	7,462	2,175
Leaves	6,600	1,741
Flower	7,245	2,266
Green Pods	5,017	878

Of approximately 14,000 homoeolog pairs

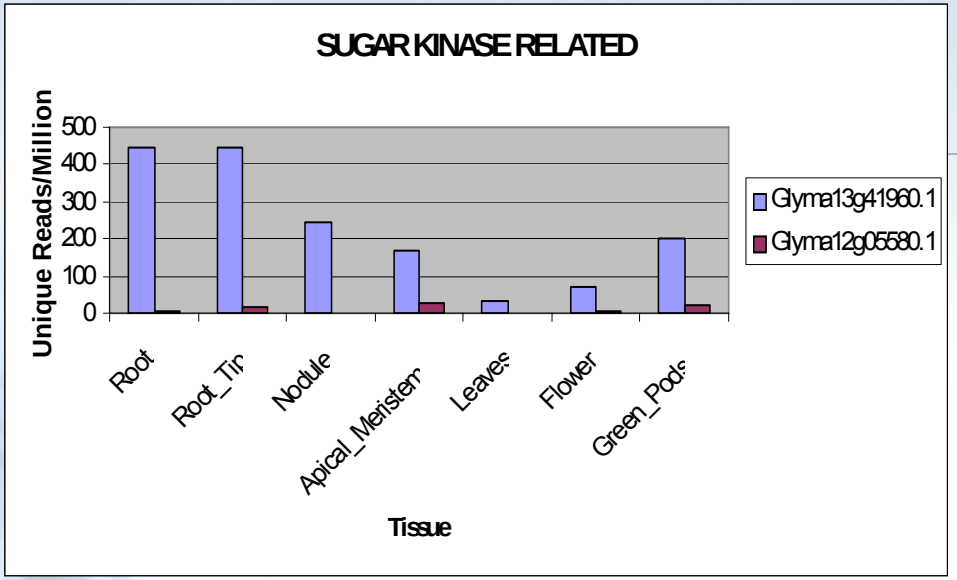
No Significant Inter-Homoeolog Differences

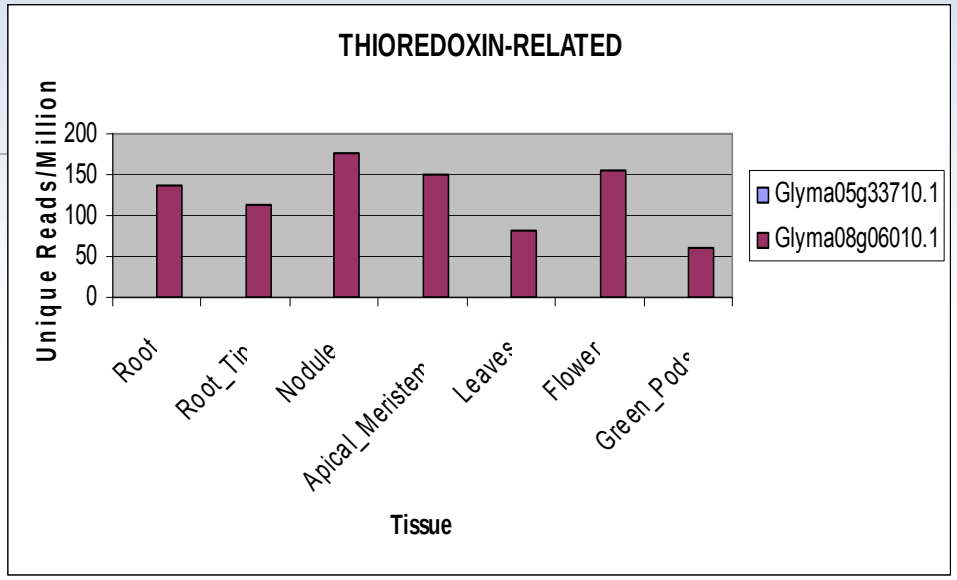


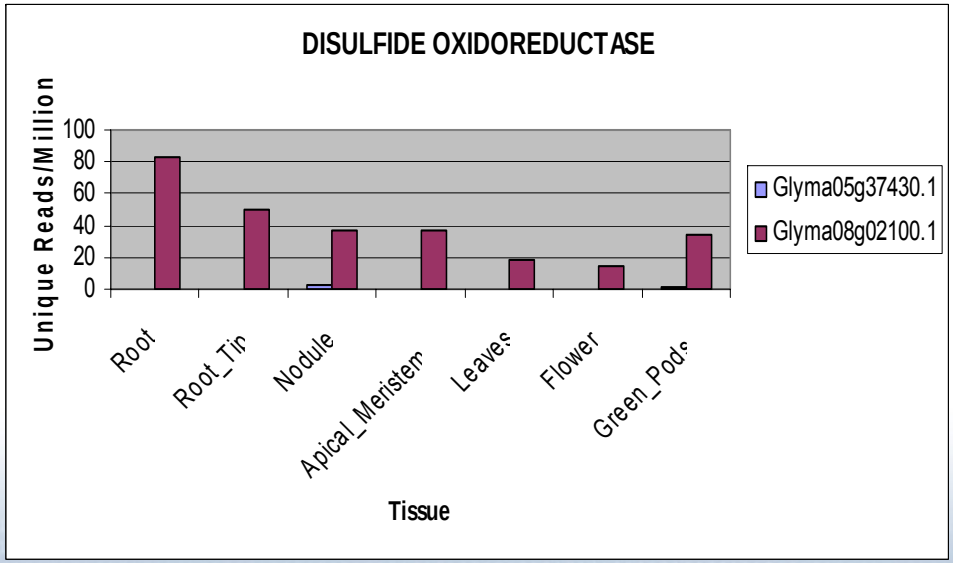




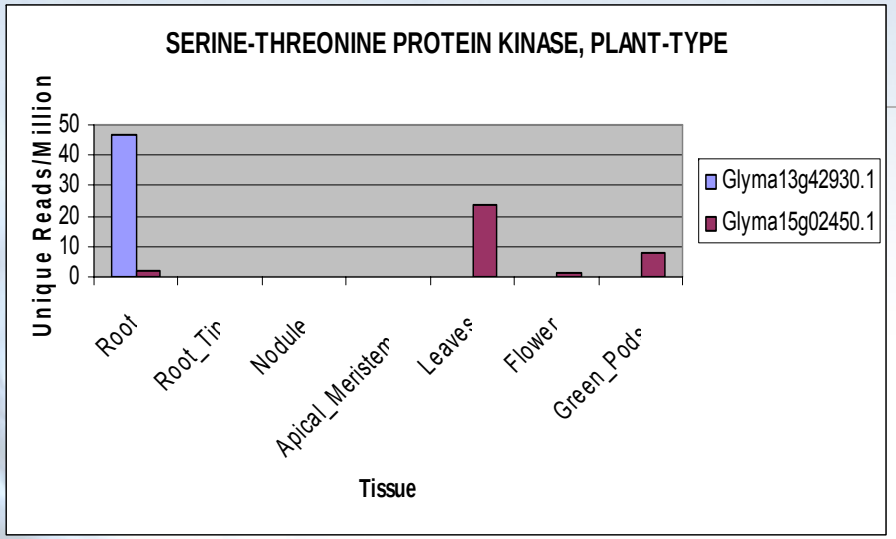
One Homoeolog Preferentially Expressed

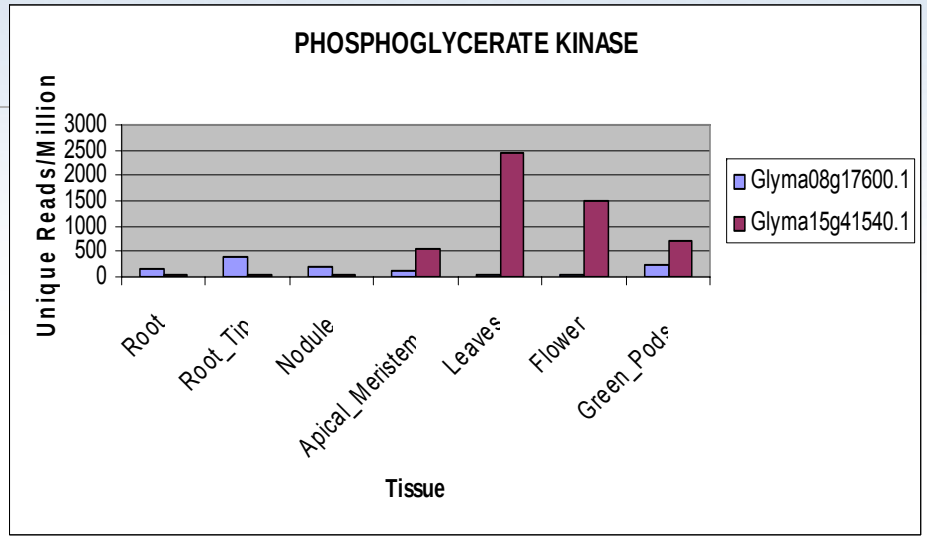


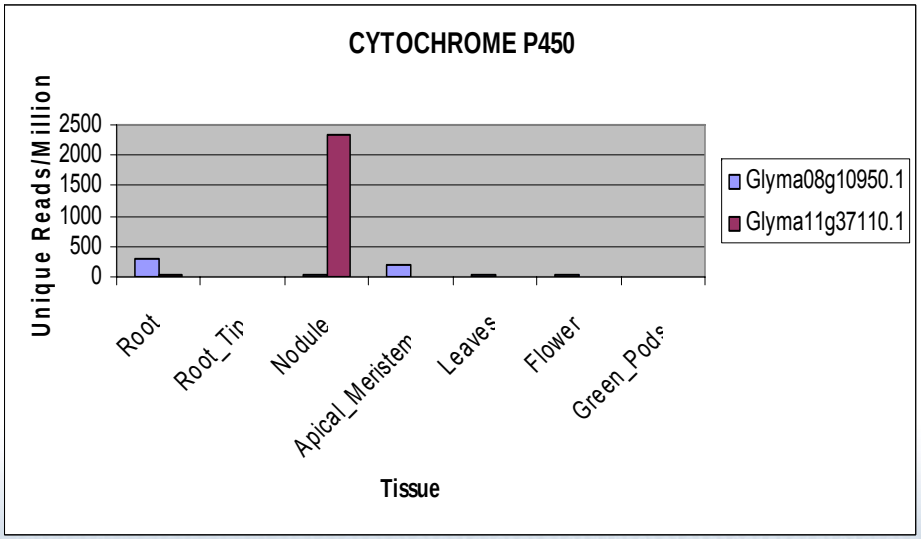




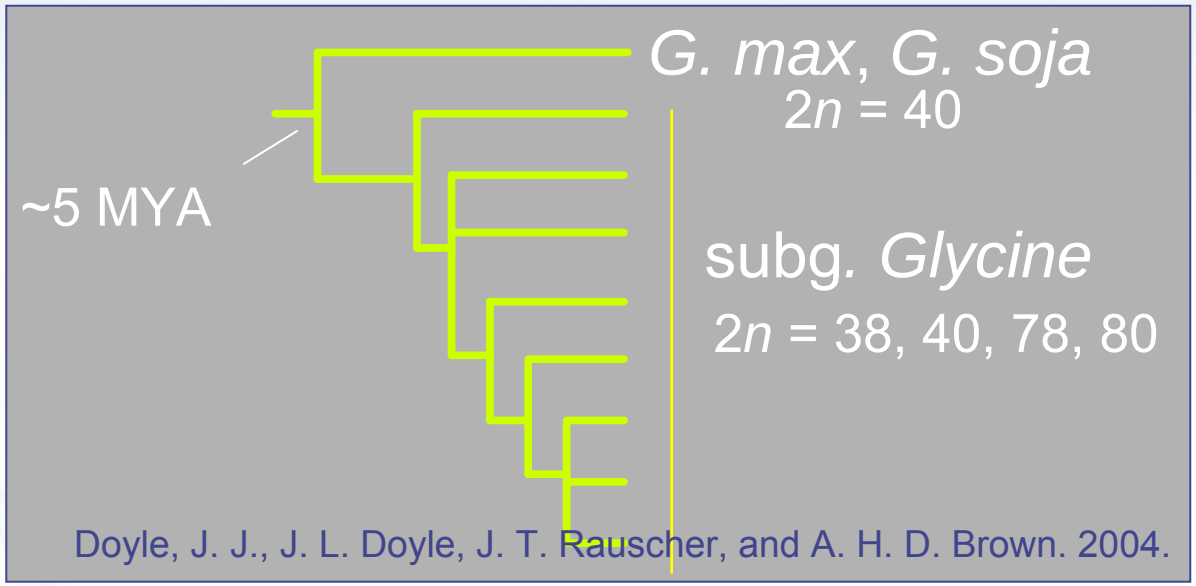
Subfunctionalized Homoeologs?







Annual and perennial *Glycine* diverged around 5 million years ago



Recent (ca. 50,000 yr.) allopolyploidy in subgenus *Glycine*

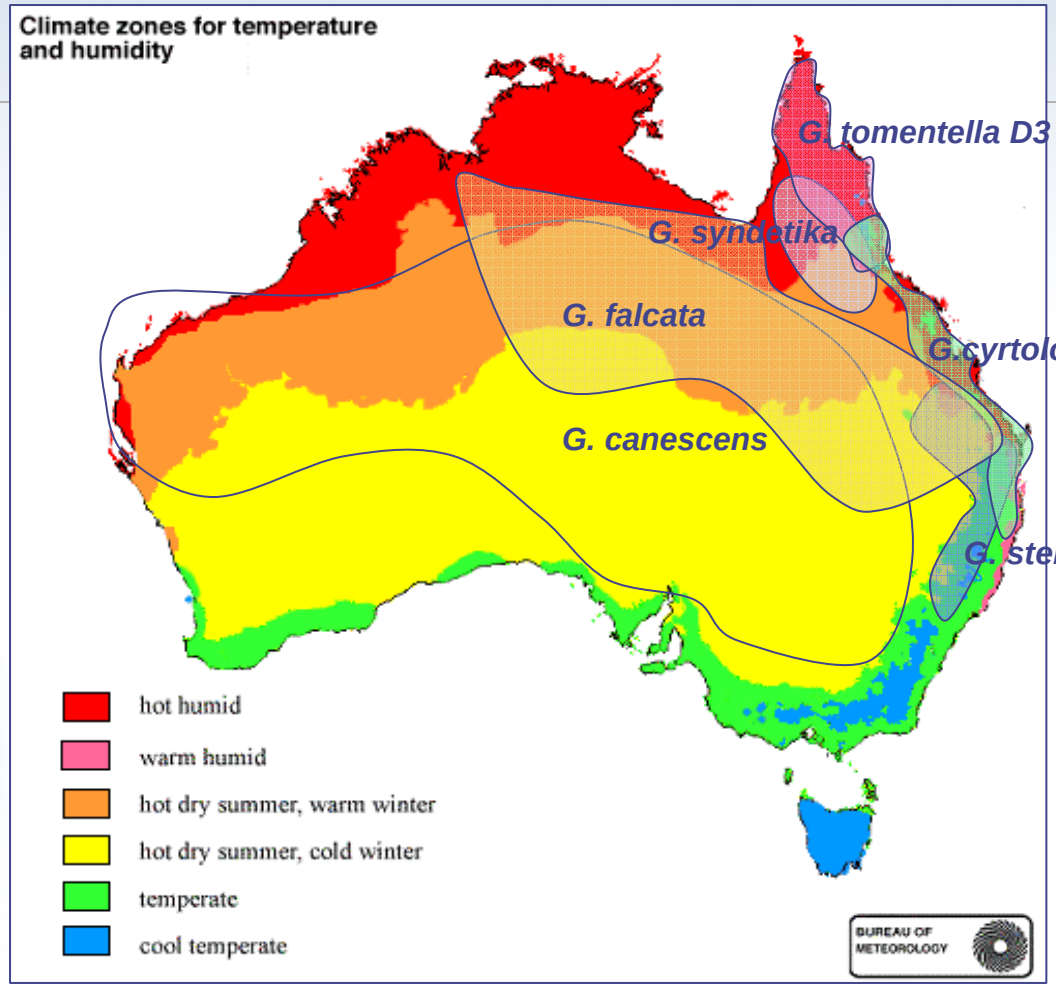
(From 2*n* = 38, 40 to 2*n* = 78, 80)

Cornell University

Jeff Doyle
Daniel C. Ilut
Thomas Owens

NCGR

Perennial *Glycine* species occur throughout Australia



What physiological changes might have promoted polyploid diversification?

Experimental Design

- Species

  Diploids → Tetraploid

  D3 x D4 → T2

- Daytime Light intensity:
  - Limiting light: 125 µmol $m^{-2}s^{-1}$
  - Excess light: 800 µmol $m^{-2}s^{-1}$
- Whole transcriptome shotgun sequencing (mRNA Seq)

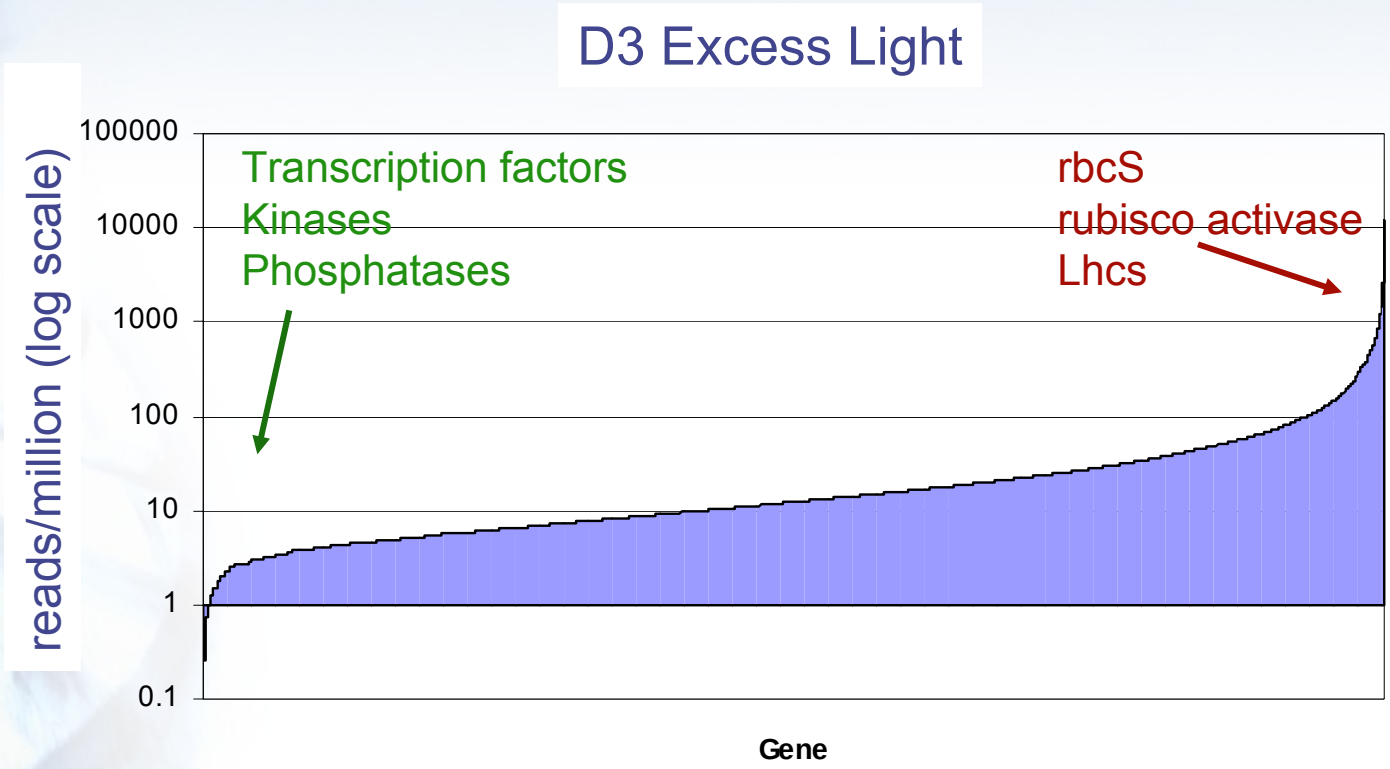
Solexa results

Summary Statistics:

Sample	Channels	Reads	Reads Aligned	Percent Aligned	Reads Uniquely Aligned	Percent Uniquely Aligned	Genes sequenced
D3E	**1**	5065610	3957656	78.13%	2000660	39.49%	43520
D3L	**1**	5148542	3928079	76.29%	1642540	31.90%	41311
D4E	**2**	732016	402852	55.03%	258535	35.32%	31898
D4L	**3**	11544369	8517865	73.78%	4611424	39.95%	47357
T2E	**2**	8383794	6057267	72.25%	3947004	47.08%	48294
T2L	**3**	13707519	10345139	75.47%	6831894	49.84%	49291

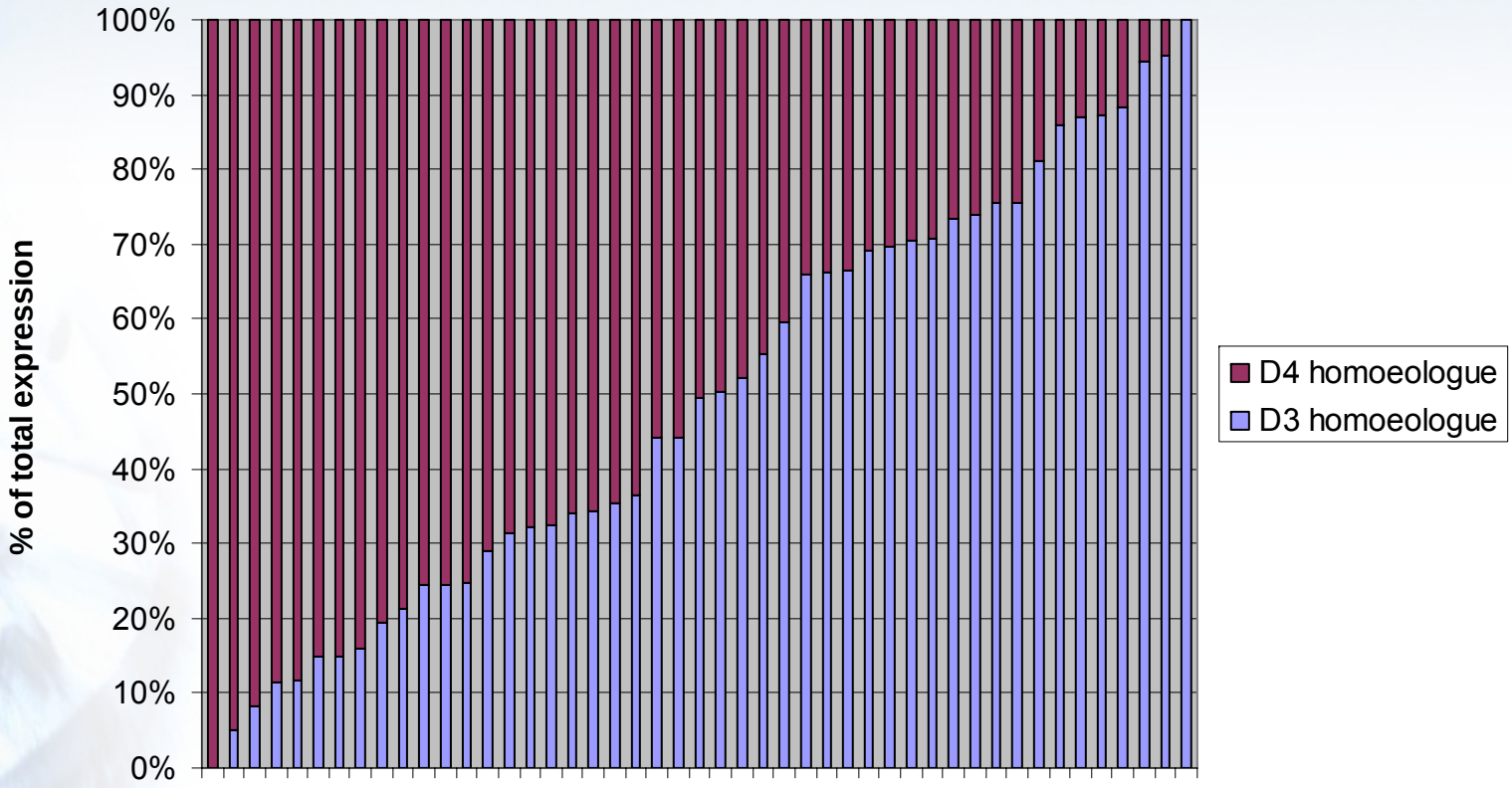
- >50.8 million reads = >1.8 Gb
- 55-78% of reads aligned
- 32-50% uniquely aligned
- Soybean genome draft 7X: 58,556 genes

Detection of transcription over a wide range of expression levels



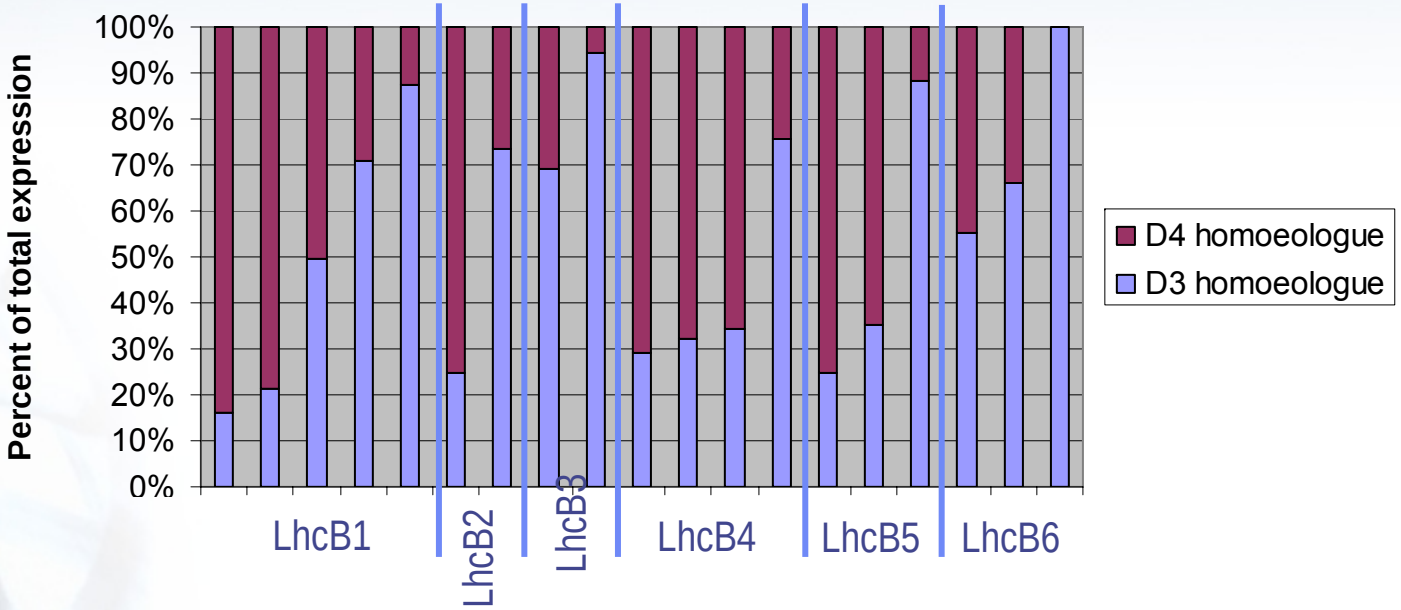
43,520 genes ranked by expression

Homoeologue expression biases are common, and occur in both directions



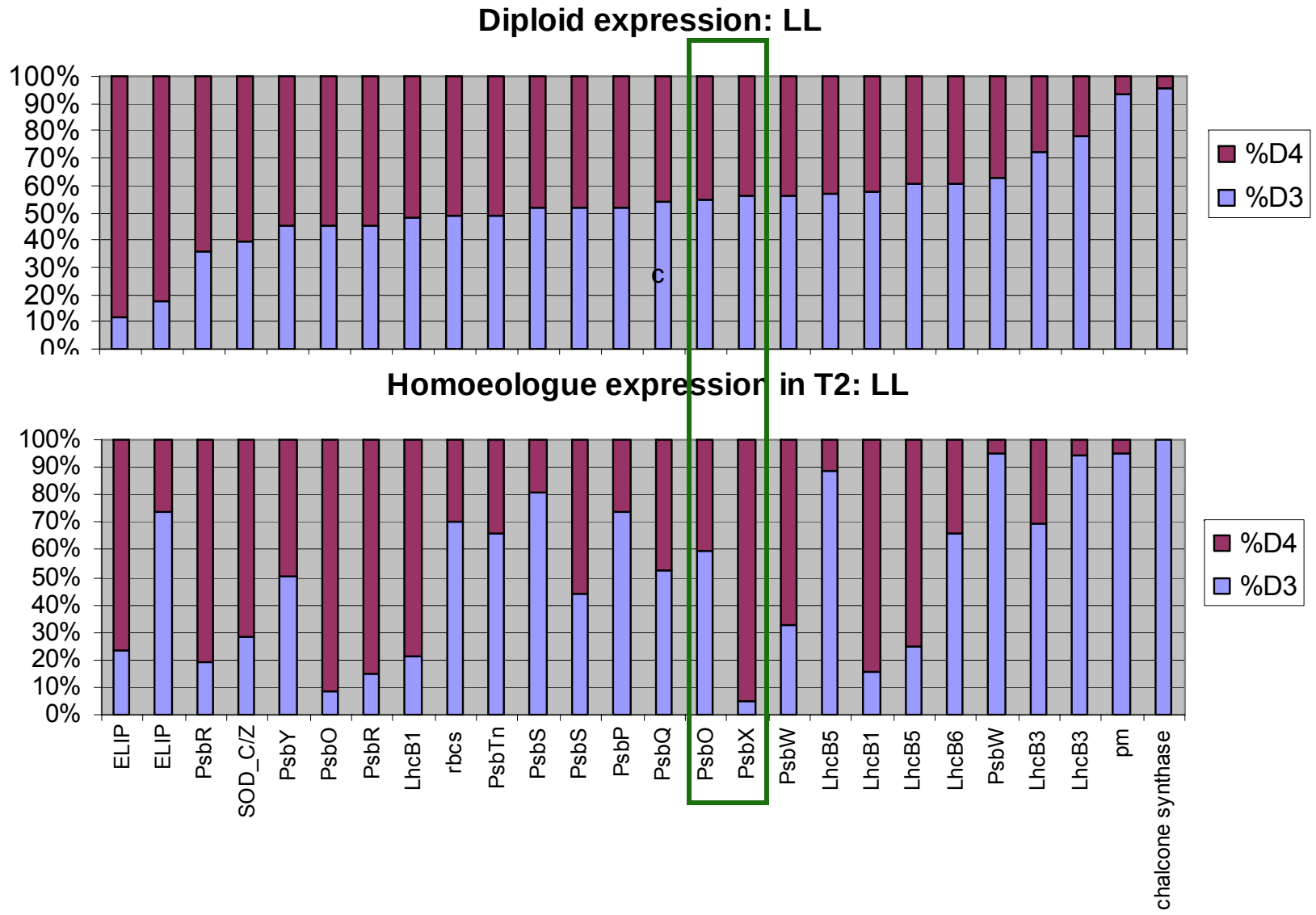
Contributions of homoeologues to total expression in limiting light for 47 genes encoding subunits of photosystem II.

Reciprocal homoeologue expression biases within gene families



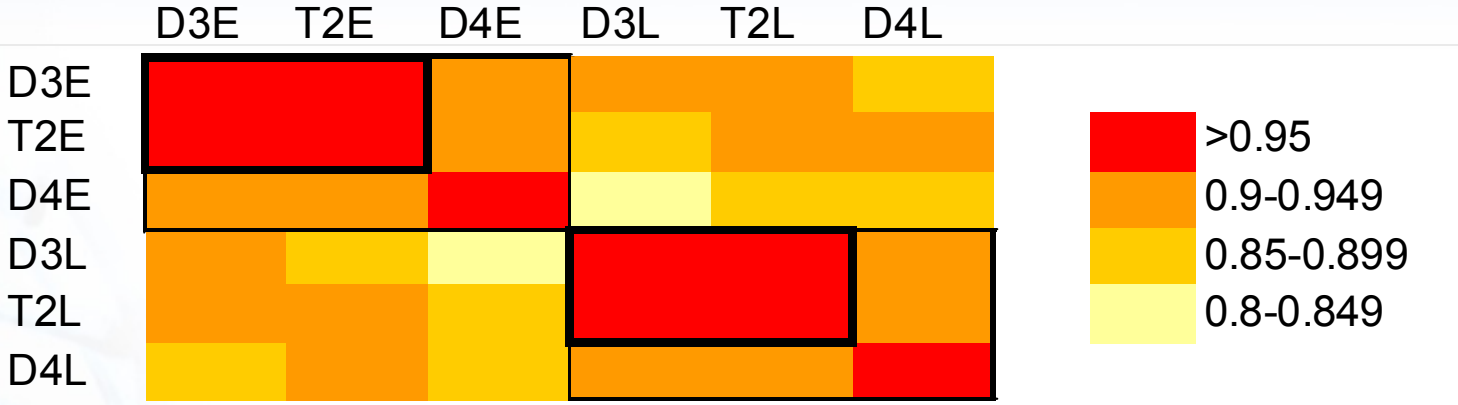
Are homoeologue expression biases simply carried over from diploids?

Homoeologue expression is not simply carried over from diploids



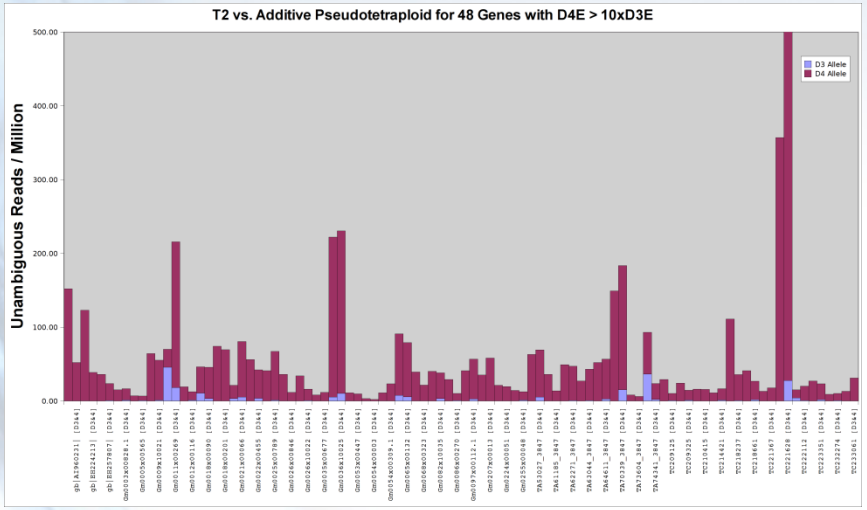
Does the overall pattern of gene expression in the tetraploid look more like one diploid progenitor than the other?

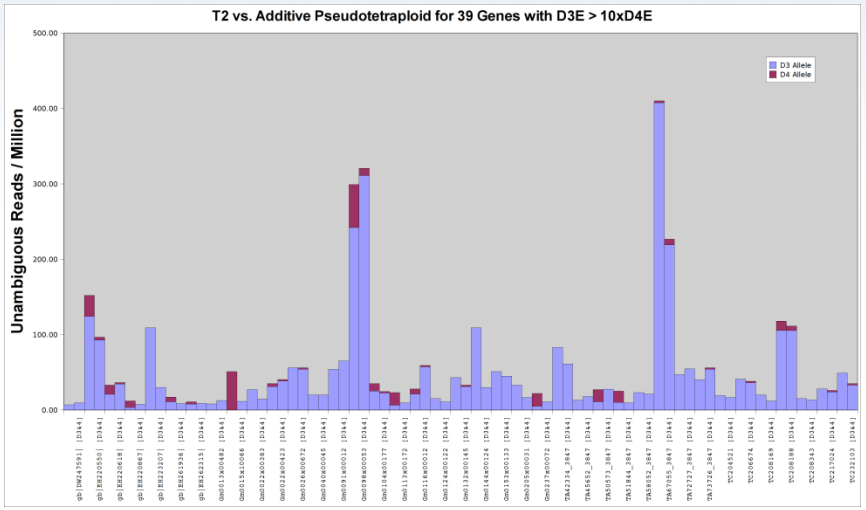
Yes, T2 is more like D3 under both high and low light conditions



Heat map of pairwise Pearson correlation coefficients, using unique RPM and all genes.

The tetraploid primarily express the homoeologue of the highly expressed diploid parent





(D4E >10x D3E)

(D3E >10x D4E)

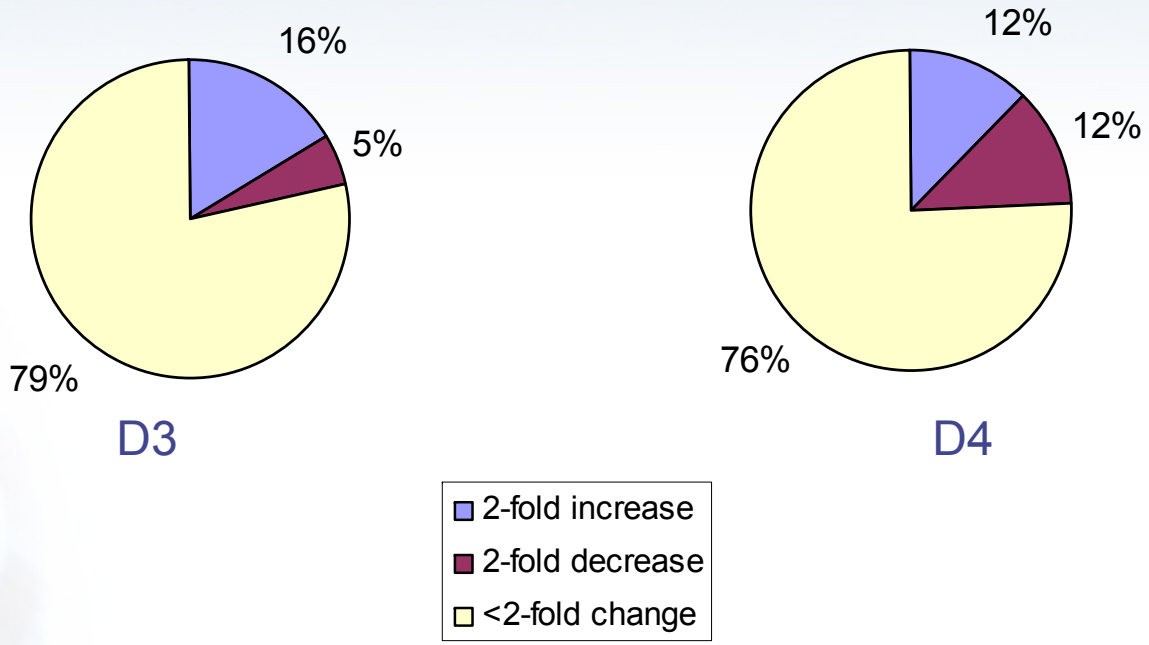
Reflects inherited bias from diploids

- D3-allele
- D4-allele

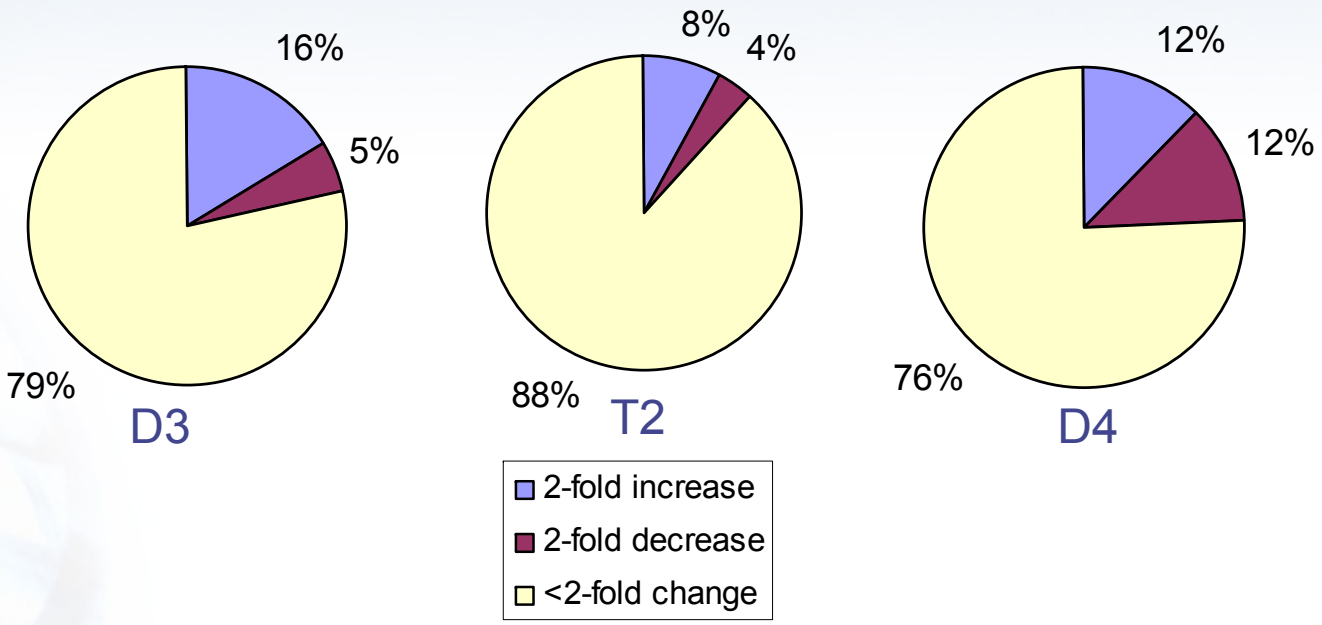
Are similar suites of genes up- or down-regulated in diploids and the polyploid under excess light conditions?

What fraction of transcriptomes show >2x response on shift from low to excess light?

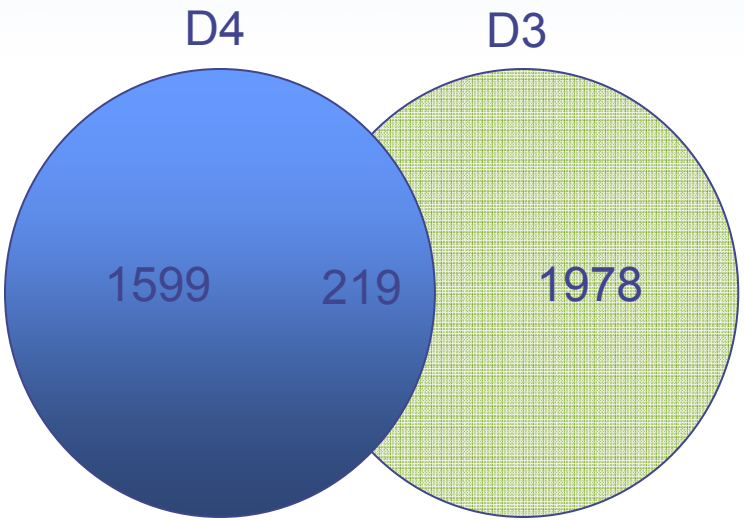
Over 20% of diploid transcriptomes are light responsive



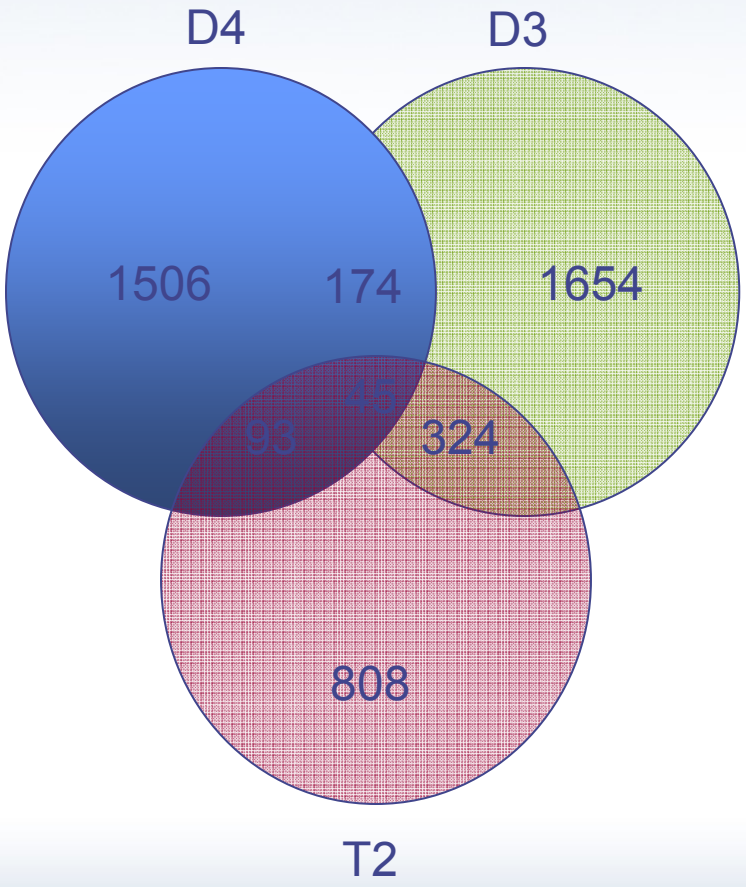
Only 12% of the polyploid transcriptome is light responsive



Very different suites of genes are up-regulated under excess light in the two diploids



Tetraploid has a unique response, distinct from either diploid



What is the biology of the genes showing different expression patterns?

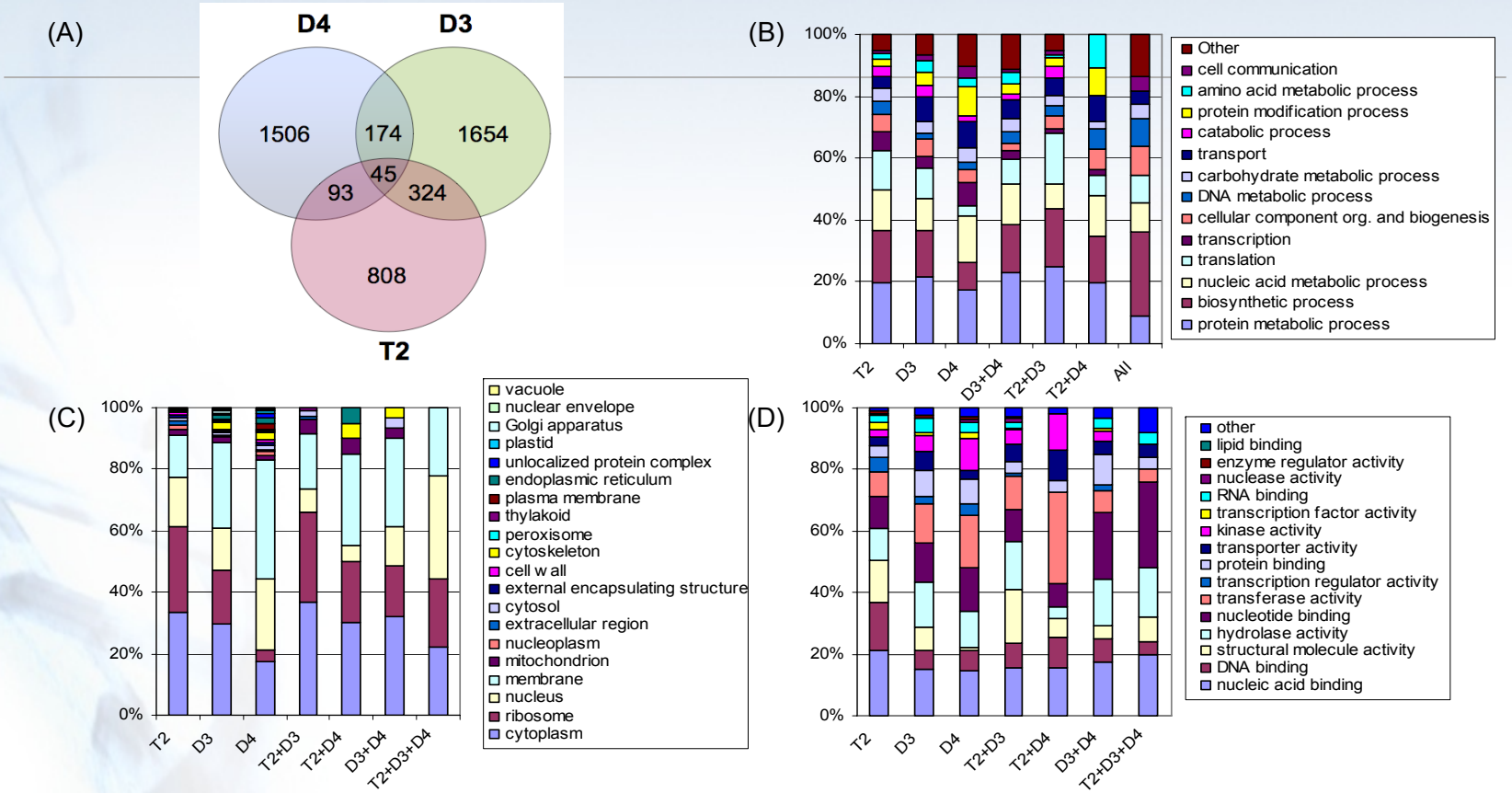


Fig. 2. (A) Genes up-regulated under excess ight conditons in D3, D4, and T2, showing overlapping sets of genes. (B,C, and D) GO ŅBiological ProcessÓ(B), ŅCellular ComponentÓ(C), and ŅMolecular FunctionÓ(D) category classifications for each of the gene sets from the Venn diagram (e.g., the 808 genes up-regulated only in T2 are broken down in the histograms labeled ŅT2Ó)The histograms represent about 25% of the genes from the Venn diagram; around 50% are unclassified, and an additional approximately 25% are classified in the two most abundant classes for each category, which show similar levels in all seven comparisons, and have been omitted to facilitate visualization of the remaining classes.

Summary

Perennial *Glycine* includes an allopolyploid complex that has formed recently, multiple times, and often bidirectionally from extant diploid progenitors

Glycine allopolyploids can differ from their diploid progenitors in key aspects of photosynthesis

G. tomentella T2 departs from additivity in the amount and relative homoeologue expression of many genes

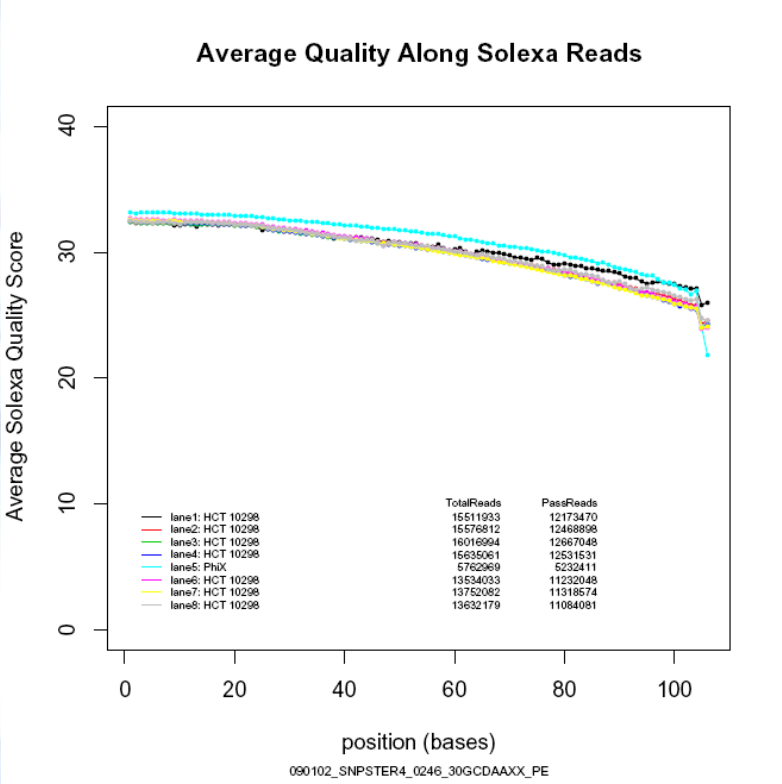
What does it all mean … about photoprotection or about allopolyploidy?

mRNA Sequencing Provides...

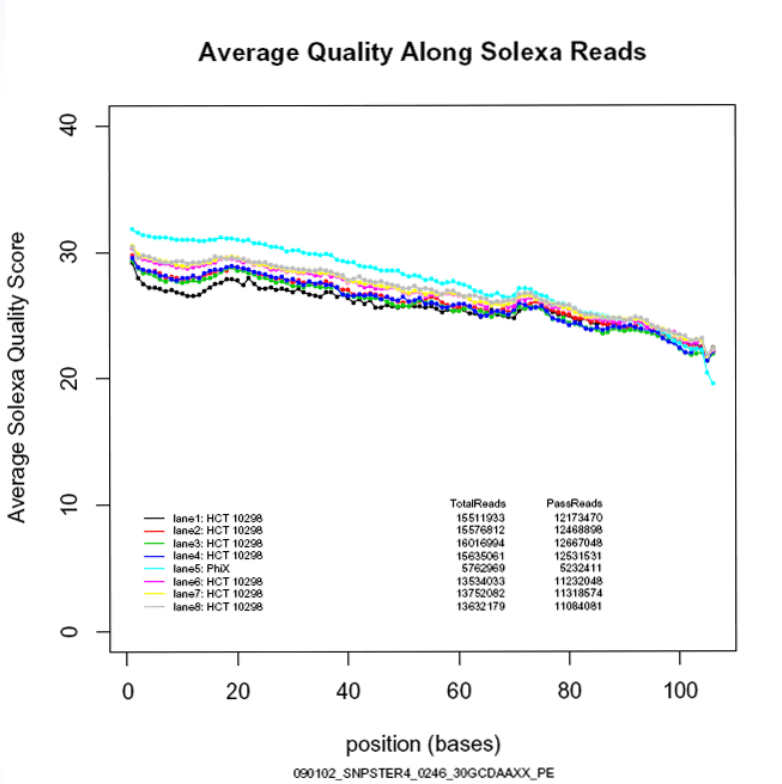
- Global gene expression profiling.
  - Expression level comparisons
  - Identification of structural variants (alternative splicing)
  - Gene copy (duplicate gene) - specific expression
- The ability to identify genetic variants.
  - Population level genetic diversity of complex genomes
  - Understand impact of genetic variants on gene expression

What about determining allelic- or homoeolog-specific expression in less studied species?

First Read



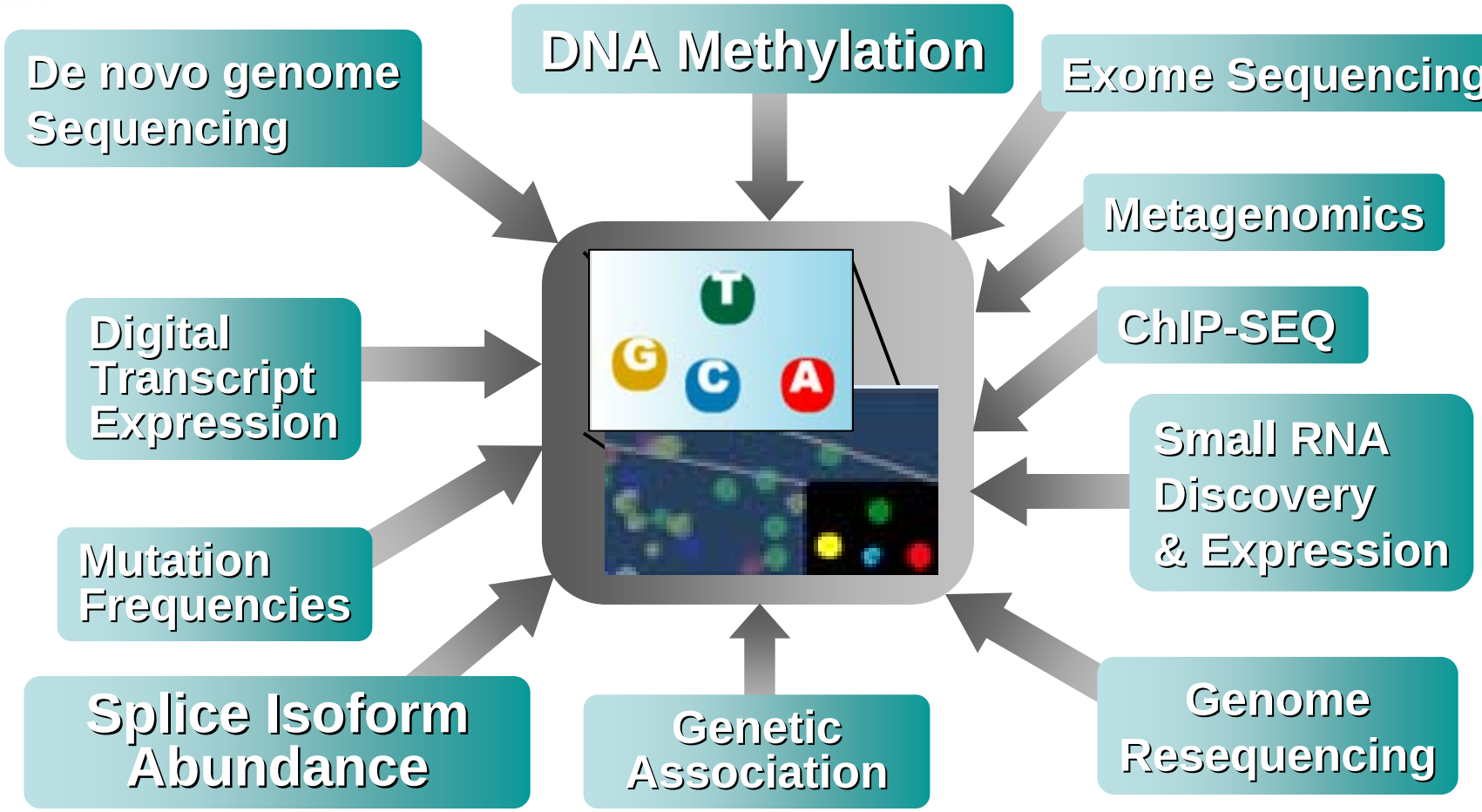
Second Read



>25 Gbp Illumina >90bp Paired-End Run

National Center for Genome Resources

Illumina Sequencing Applications



Alpheus DB Stats

- >50 instances (db + website)
- >50 organisms

Human
Bos taurus (bovine)
Sus scrofa (pig)
Rat
Chinese Hamster
Mouse
Gallus gallus (chicken)
Catfish
Tuna

Tuberculosis
Euglena
Staphylococcus aureus
Vibrio fischeri
Bacillus sp.
Pseudomonas fluorescens
Burkholderia
P. capsici
P. Phaseoli
Yeast
Mosquito
Moth
Stink Bug
Tobacco Bud Worm
Tarnished Plant Bug
Corn Earworm

Arabidopsis thaliana
Rice
Barley
Maize
Sorghum
Brachypodim
Sunflower
Soybean
Chickpea
Pigeonpea
Medicago truncatula
Watermelon
Pepper
Tomato
Citurs
Cotton
Sugarcane
Poplulus
Cocao

Acknowledgements

- **NCGR - Sequencing Center**
  - Greg May
  - James Huntley
  - Ryan Kim
  - Jennifer van Velkinburgh
  - Leonda Clendenen
- **NCGR – Informatics, Statistics and Software Engineering**
  - Joann Mudge
  - Andrew Farmer
  - Faye Schilkey
  - Ingrid Lindquist
  - Neil Miller
  - Stephen Kingsmore
  - Ernie Retzel
  - Lar Mader
  - Kamal Gajendran
  - Selene Virk
  - Forrest Black
  - Kathy Myers
  - Dan Weems
  - Melodie Rice
  - Raymond Langley
- **USDA-ARS, Iowa State University**
  - Randy Shoemaker
  - Steven Cannon
  - Michelle Graham
  - Bindu Joseph
  - Nathan Weeks
- **USDA-ARS, University of Minnesota**
  - Carroll Vance
  - Yung-Tsi Bolon
  - Nevin Young
  - Roxanne Denny
- **University of Minnesota**
  - Gary Muehlbauer
- **University of Illinois-UC**
  - Brian Diers
- **University of Nebraska**
  - James Specht
  - NCSRP Grant to Specht et al.
- **DOE-JGI**
  - Jeremy Schmutz
  - Therese Mitros
- **USDA-ARS**
  - Rich Wilson
- **Illumina**
  - Gary Schroth
  - Irina Khrebtukova
  - Shujun Luo
- **Carleton College**
  - Susan Singer
  - Sonja Maki
- **Cornell University**
  - Jeff Doyle
  - Daniel C. Ilut
  - Thomas Owens
- **University of Missouri**
  - Gary Stacey
  - Marc Libault
  - Laurent Brechemacher
- **Purdue University**
  - Scott Jackson
  - Jessica Schlueter