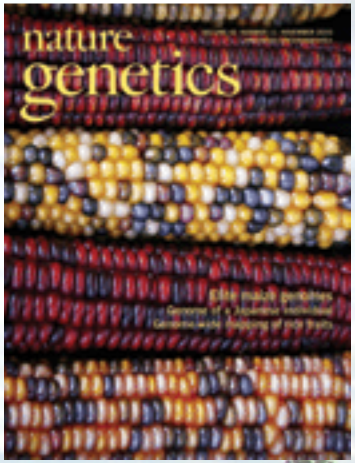
Extensive Intra-Specific Structural Variation among Maize Genes



Springer et al., PloS Genetics, 2009
Lai et al., Nat. Genetics, 2010

PAG XIX
San Diego, CA
16 January 2011

Patrick S. Schnable
Iowa State University

Kai Ying
(应开)

Yan Fu
(傅延)

Wei Wu
(吴薇)

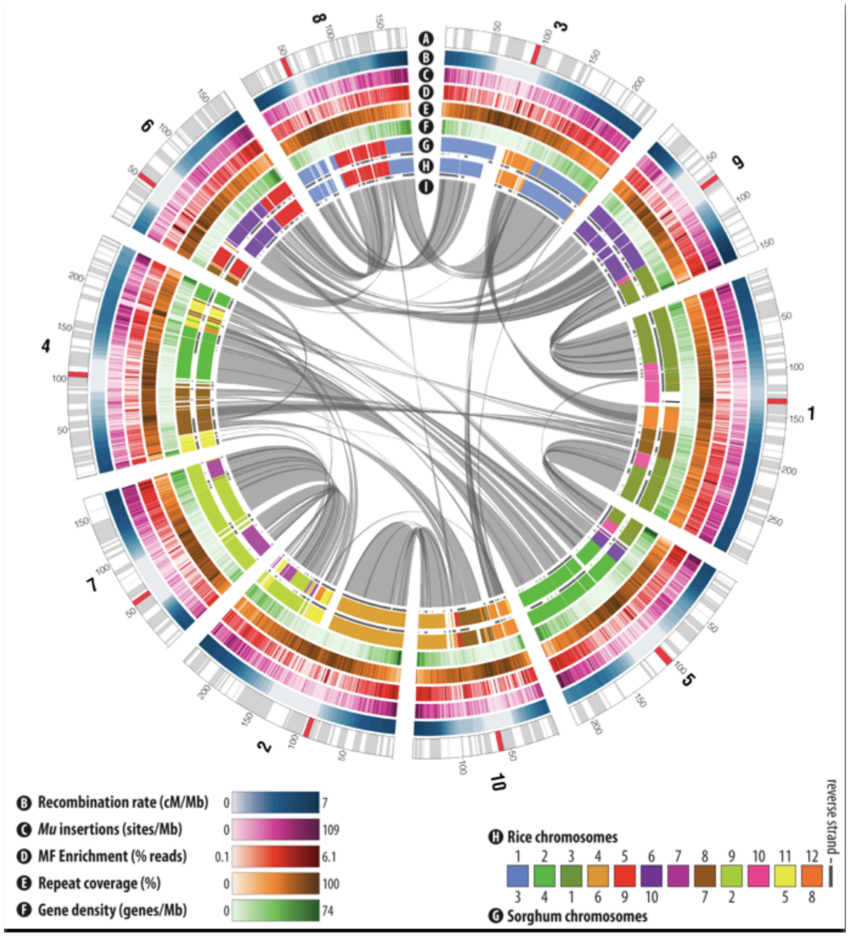
Nathan
Springer

Jeff
Jeddeloh

Brad
Barbazuk

Jinsheng
Lai

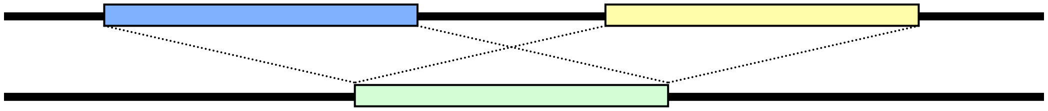
The $30M B73 Maize Genome Sequencing Project


A
B
C
D
E
F
G
H
I
1
2
3
4
5
6
7
8
9
10
B Recombination rate (cM/Mb) 0 7
C Mu insertions (sites/Mb) 0 109
D MF Enrichment (% reads) 0.1 6.1
E Repeat coverage (%) 0 100
F Gene density (genes/Mb) 0 74
H Rice chromosomes
1 2 3 4 5 6 7 8 9 10 11 12
3 4 1 6 9 10 7 2 5 8
G Sorghum chromosomes
reverse strand

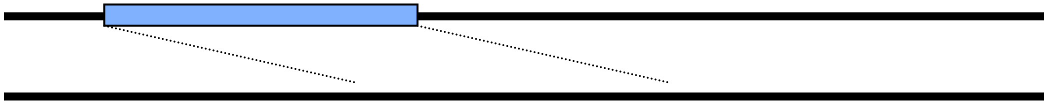
- WU Genome Sequencing Center (R. Wilson, PI); Arizona Genome Institute; Cold Spring Harbor Laboratory; Iowa State University
- Schnable et al., Science, 2009

Structural Variation (CNV & PAV)

CNV



PAV



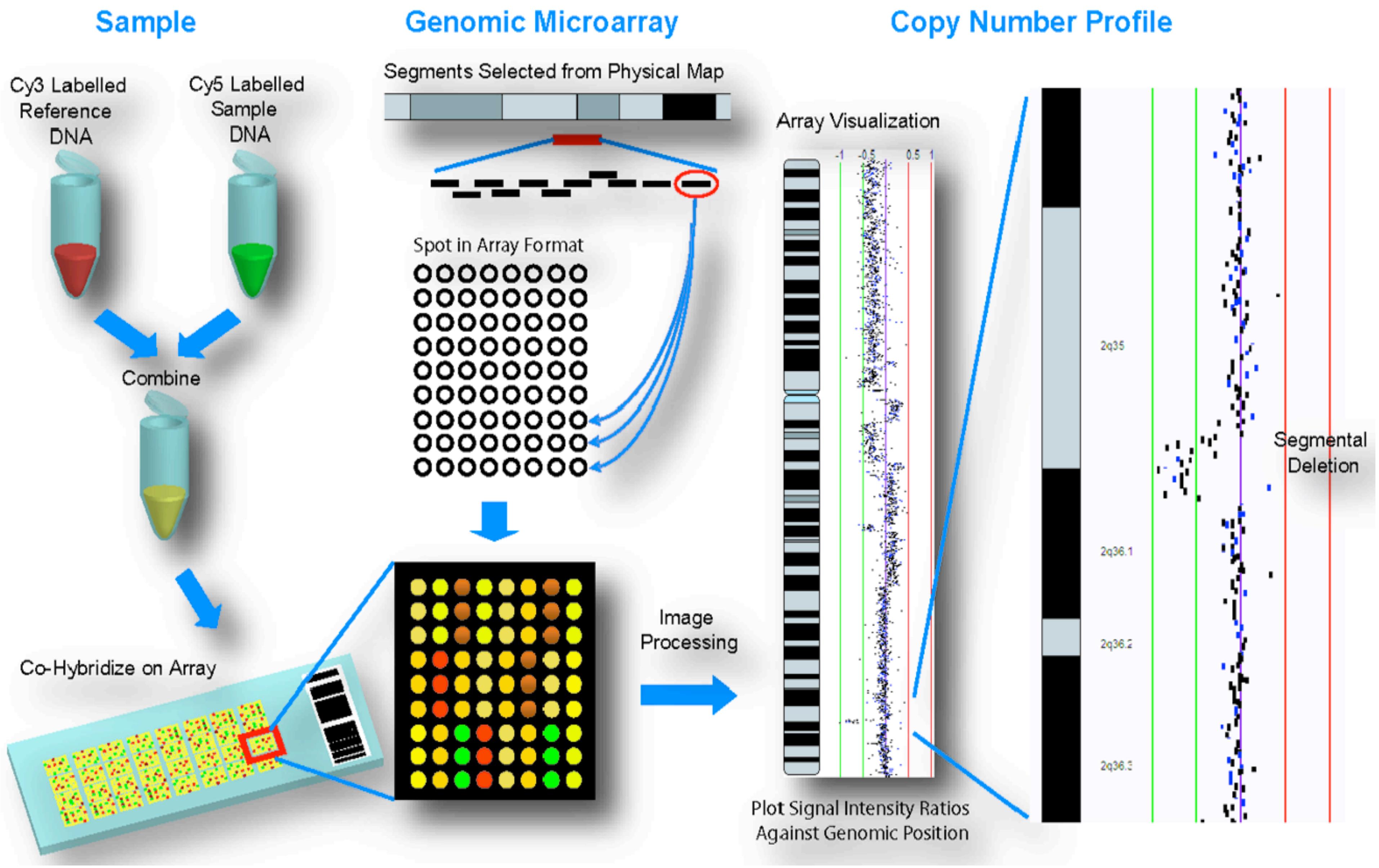
- •In humans SV can be associated with disease (“traits”)

- •In maize:
- •What is overall level of (genic) SV? (high)
- •Does SV contribute to phenotypic diversity? (yes)

Array-based Comparative Genome Hybridizations (CGH)

- Nimblegen's HD2 Array (~2.1M probes)
- Probes designed using a "frequency masked" 200 bp tile-path through the *draft* B73 genome sequence
- Genotypes: B73, Mo17 (different heterotic groups)

Introduction to CGH



Sample

Cy3 Labelled Reference DNA

Cy5 Labelled Sample DNA

Combine

Co-Hybridize on Array

Genomic Microarray

Segments Selected from Physical Map

Spot in Array Format

Image Processing

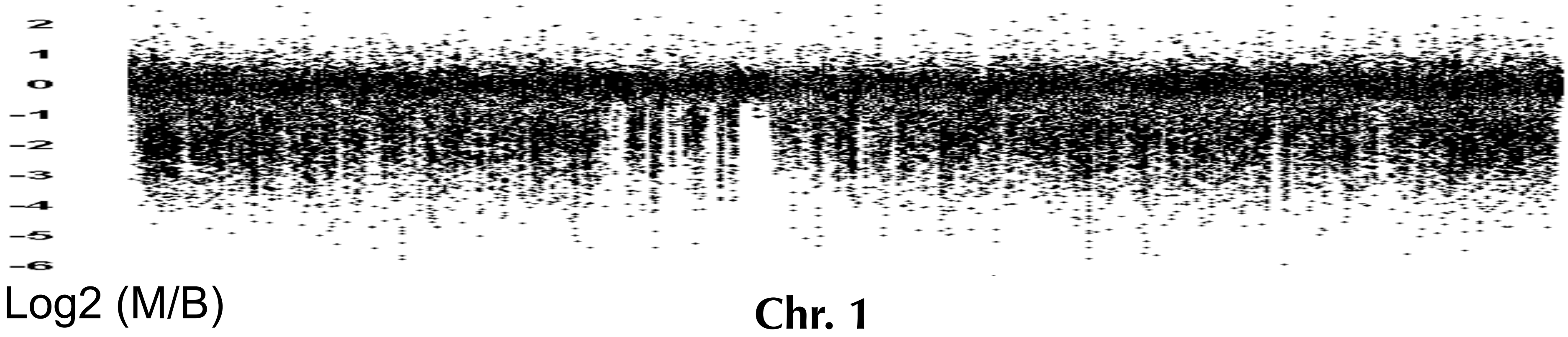
Copy Number Profile

Array Visualization

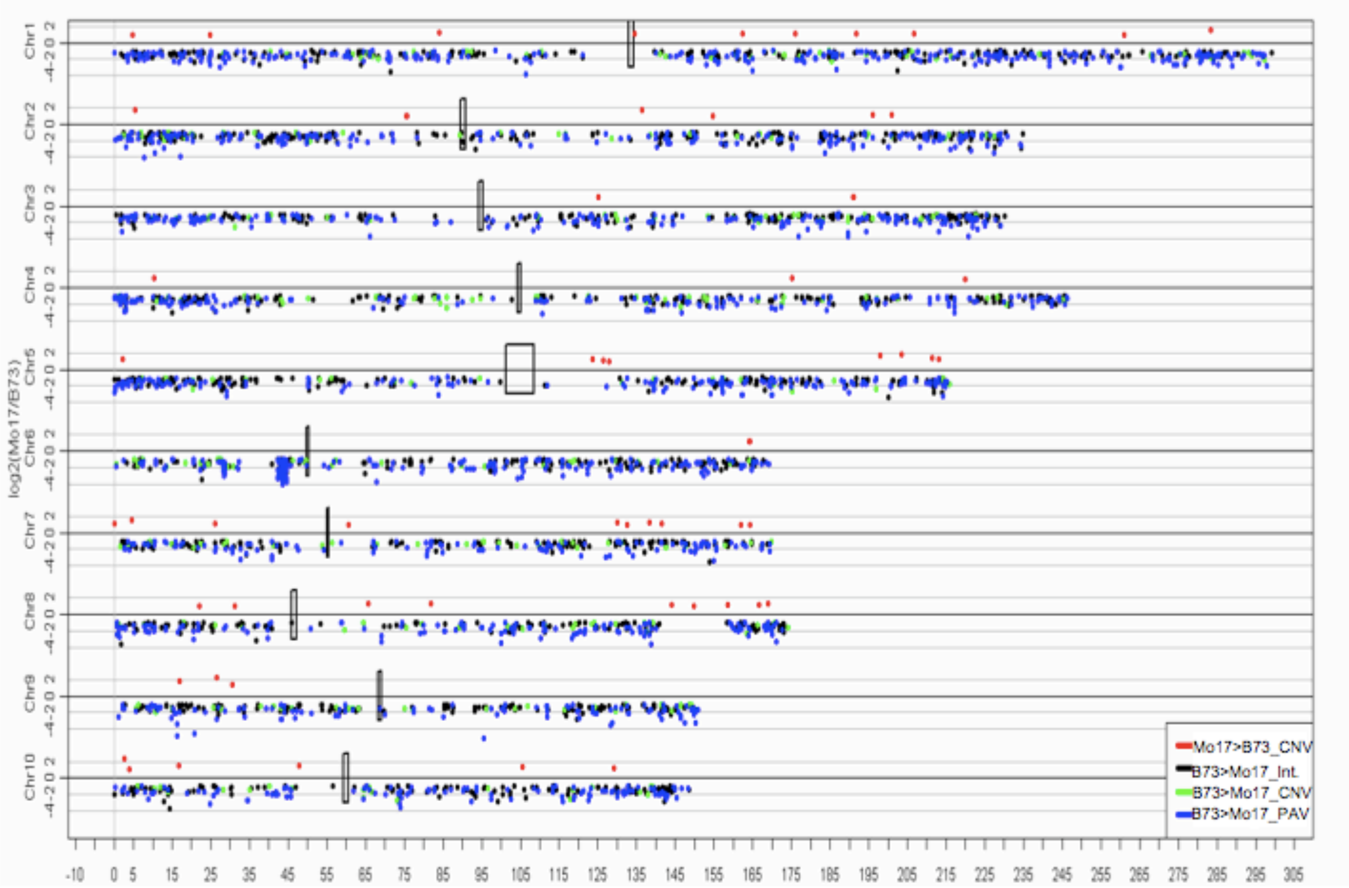
Plot Signal Intensity Ratios Against Genomic Position

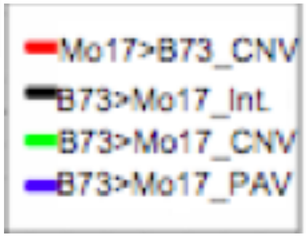
Segmental Deletion

Detection of CNV via CGH signal intensity



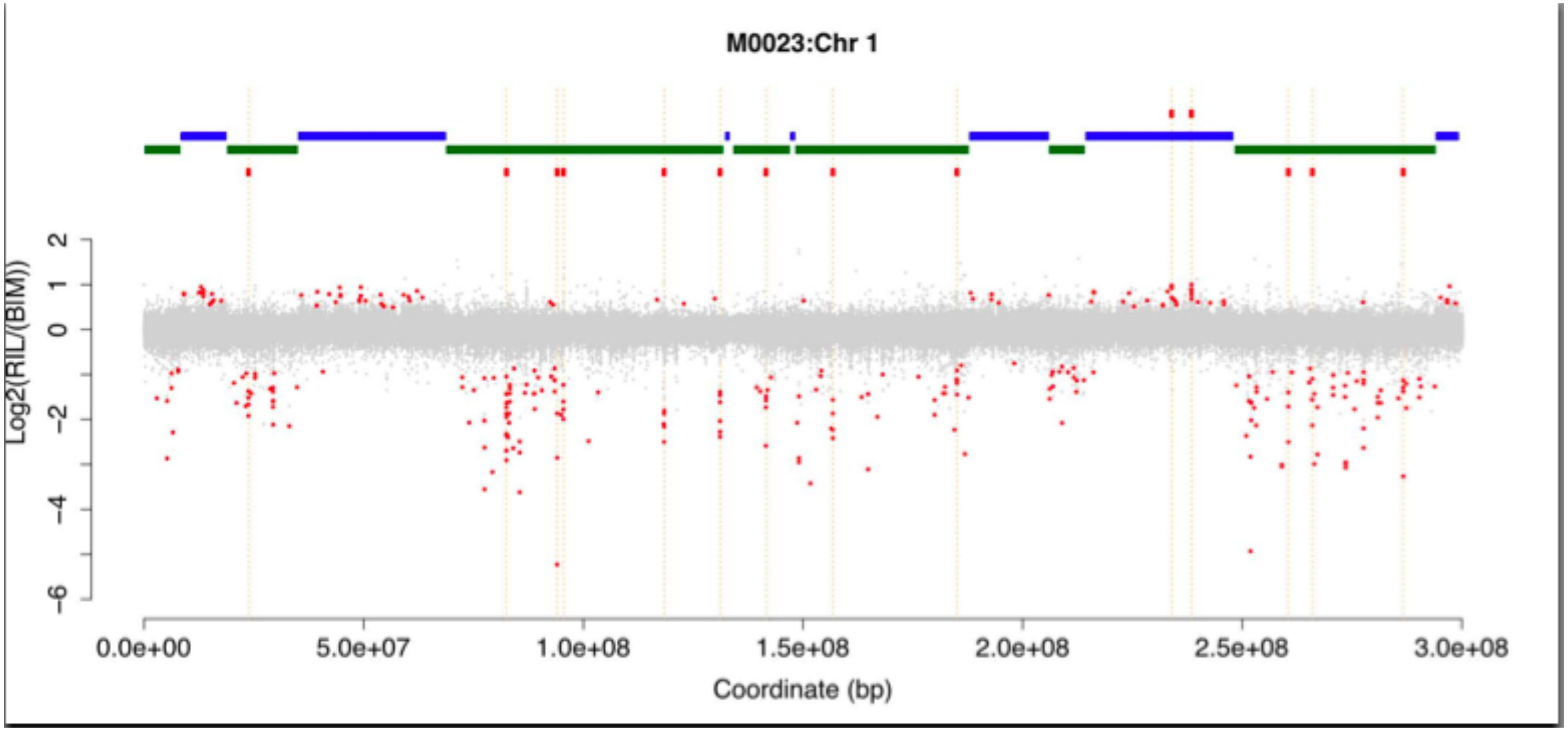
Several hundred *intact,* expressed, phylogenetically conserved genes exhibit CNVs and PAVs





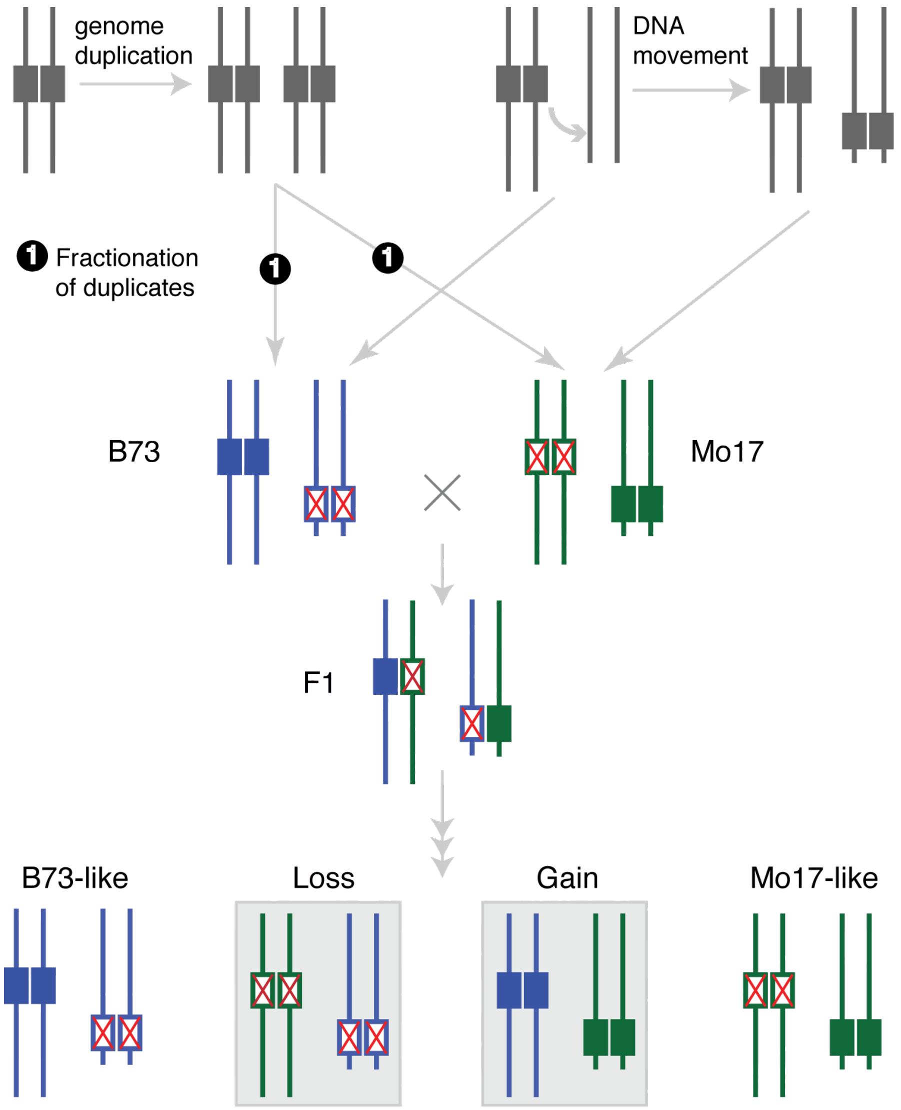
(more on this topic during BGI and Roche Workshops)

Novel CGH Patterns in RILs

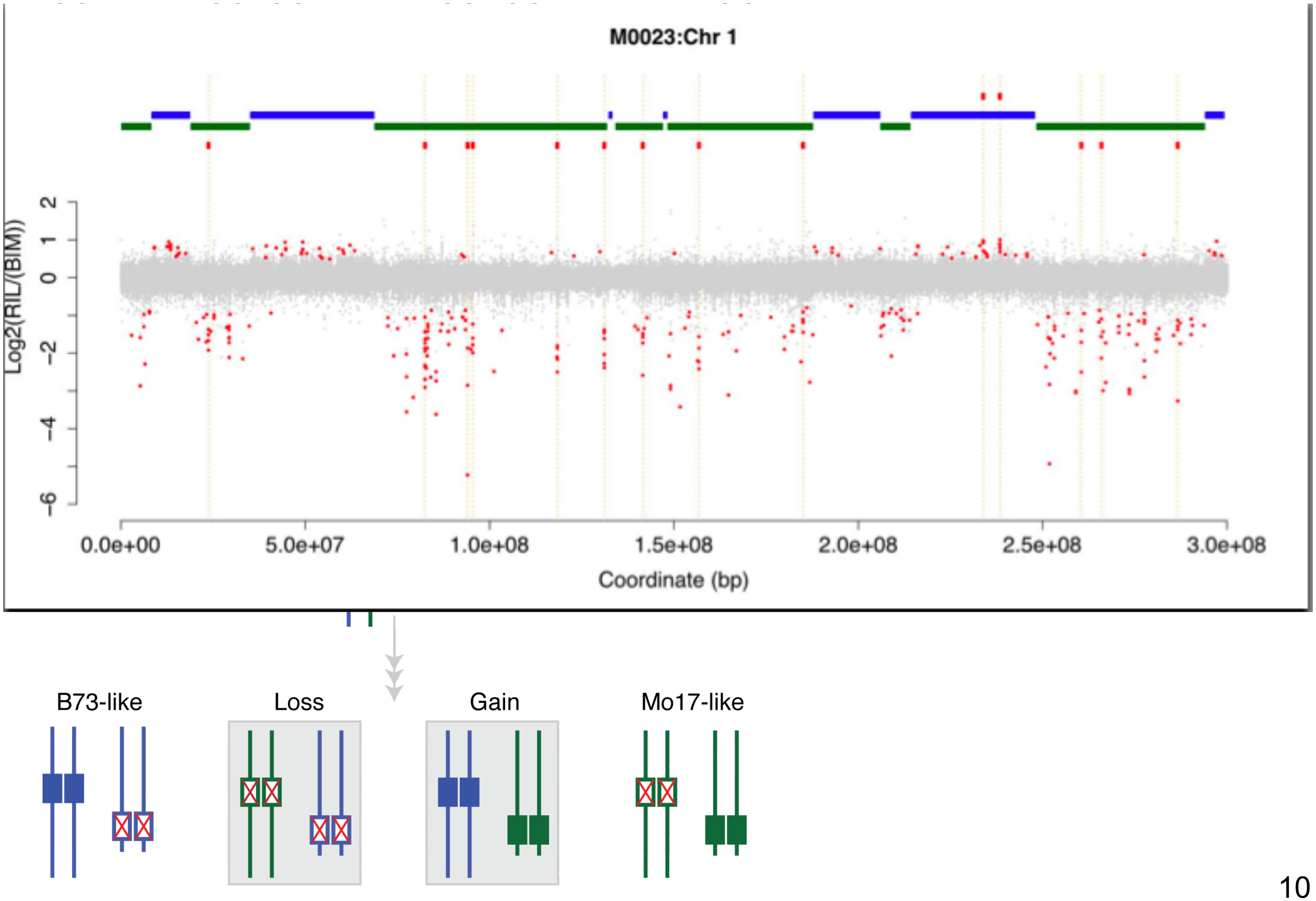


CGH signals for genes present in both B73 and Mo17, but at non-allelic positions (unlinked)

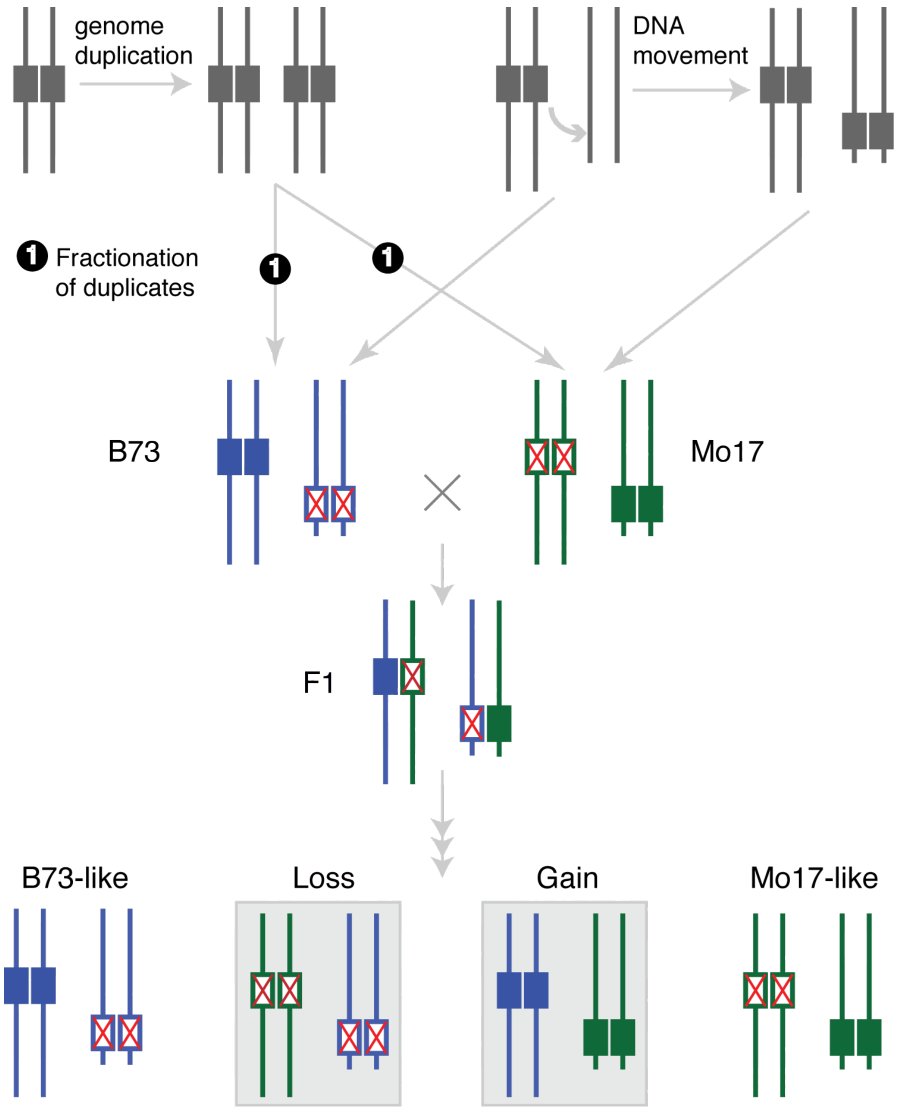
Segregation of Non-Allelic Gene Copies Generates PAVs/CNVs and Novel Phenotypes



Segregation of Non-Allelic Gene Copies Generates PAVs/CNVs and Novel Phenotypes



Segregation of Non-Allelic Gene Copies Generates PAVs/CNVs and Novel Phenotypes



Changes in gene complement among RILs.

Explanation for transgressive segregation?

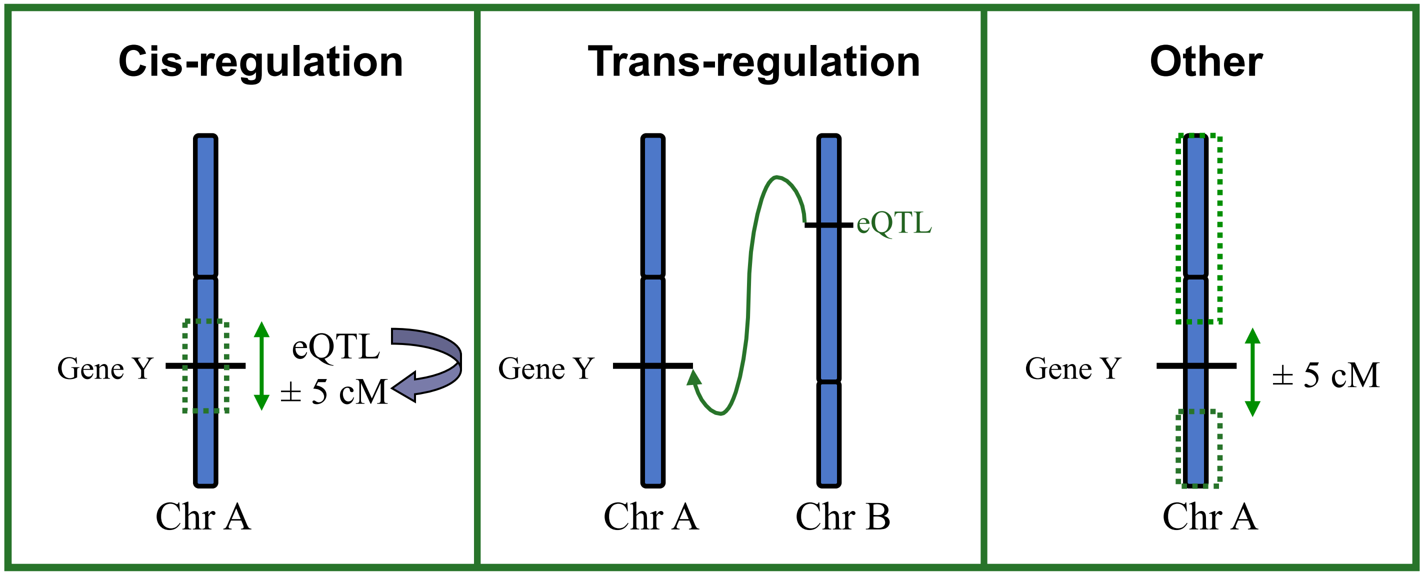
Strong statistical support for association between gene loss and yield component traits in IBM RILs

How prevalent is epistatis?

- Global tests (lots of markers and lots of traits) to maximize chances to detecting epistasis
- **What data set?**
- How to analyze?

Summary of eQTL Mapping

Identified >4,000 eQTL associations (FDRs 1-10%)



14.5% 75% 10.5%

Ruth
Swanson-Wagner

But based on only 30 IBM RILs

Swanson-Wagner et al., Science 2009

2^{nd} Data Set

- Strand-specific oligo-microarray
- Detects sense and anti-sense transcripts
- Analysis of twice as many (56) IBM RILs identified many eQTL affecting accumulation of sense or anti-sense transcripts
- >12,000 “traits”; > 1,000 markers (no missing data)

Yi Jia (贾毅)

How prevalent is epistatis?

- Global tests (lots of markers and lots of traits) to maximize chances to detecting epistasis
- What data set?
- **How to analyze?**

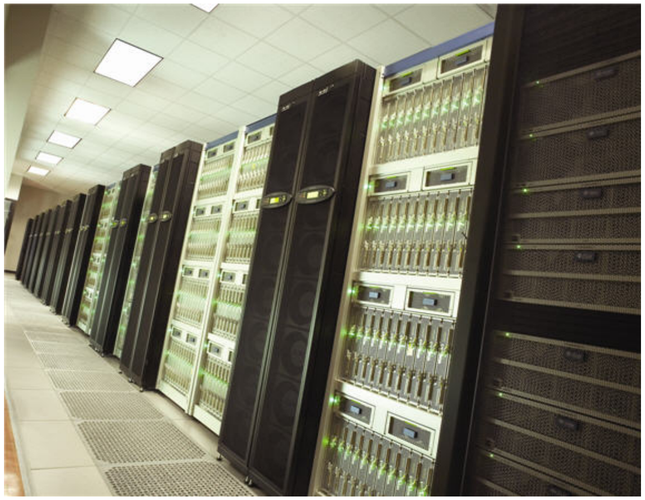
Challenge

- Quadratic increase in problem size in relation to marker number
- Linear increase with the number of lines (need for statistical power to detect smaller effects)
- Developed statistical methodology whose per-test calculation involves comparatively small number of arithmetic operations (F-test applied to a linear contrast of genotype means coupled with p-value determination via permutation and corrected for false discovery rate)
- Even so, initial run time estimate was **1,634 years** of computer time, based on quick implementations done in R and Python.
- Partnership with the NSF-funded iPlant, which made available high performance computing (HPC) expertise and machine resources at TACC.
- Then analyzed the real data with the algorithm

Hardware Used

- Results produced with two NSF TeraGrid resources at TACC
- Ranger:
  - 62,976 cores of AMD Barcelona
  - 2GB RAM per core
  - Up to 4,000 cores used for this code
- Longhorn
  - 2,048 Intel Nehalem cores
  - 6GB RAM per core
  - 512 GPUs (not used here)

Allocations of compute time on both these systems are available thru iPlant or TeraGrid for researchers in the US or with US-based collaborators



Ranger

Longhorn

Performance

- As compared to the initial estimate of the runtime for this problem (**1,634 years**) the final run time was **4.5 hours** on 128 processors of a cluster, a performance improvement of 3.2 million times!
  - Port to Fortran: 1,000x improvement.
  - Code optimization: 25x
  - On-node and inter-node parallelism: 125x
- The good news: 6 weeks of optimization saved a *millennium*.

The Bad News: No Evidence of Epistatis Detected

- Number of lines (RILs) too small?
- Data too noisy? (microarray)
- Statistical test too conservative?
- Effect sizes are small? No epistasis?

Steps forward

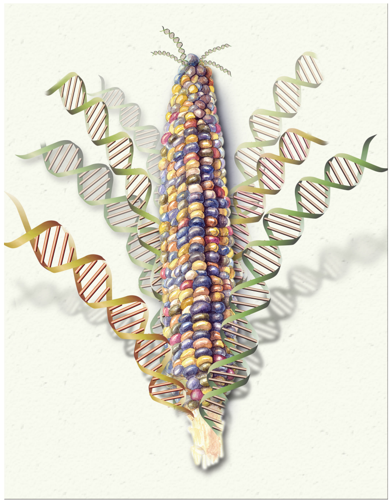
- **Simulations:** Statistical properties of our test not fully understood. What effect sizes will be detectable as the numbers of lines and markers are scaled? (etc) Will analyze simulated data that contains known epistatic interaction (Scott Chapman and Mark Dieters)
- **New data sets:** RNA-Seq data ("cleaner") being generated on larger number of RILs (more power) in collaboration with the NSF-funded SAM project (Mike Scanlon, Gary Muehlbauer, Jianming Yu, Marja Timmermans and Diane Janick-Buckner) will be analyzed with iPlant pipeline

Summary

- Maize exhibits *unprecedented* levels of SV (CNV and PAVs), affecting several hundred *genes*
- Evidence that SV contributes to the extraordinary phenotypic diversity in maize
- In collaboration with iPlant an efficient pipeline was developed to conduct genome-wide tests for epistasis
- Simulations and new datasets are coming…

Collaborators

- Srinivas Aluru (ISU)
- Yan Fu (ISU-> Monsanto)
- Jinsheng Lai (China Agriculture Univ)
- Dan Nettleton (Statistics, ISU)



UF UNIVERSITY of FLORIDA

Brad Barbazuk

UNIVERSITY OF MINNESOTA

Nathan Springer

The Maize Genome Sequencing Project, Rick Wilson, PI

NimbleGen HIGH - DEFINITION GENOMICS℠ Roche

华大基因 BGI

er S

Jeffrey Jeddeloh:

Todd Richmond; Leonard Iniguez
Heidi Rosenbaum; Jacob Kitzman

Jun Wang

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

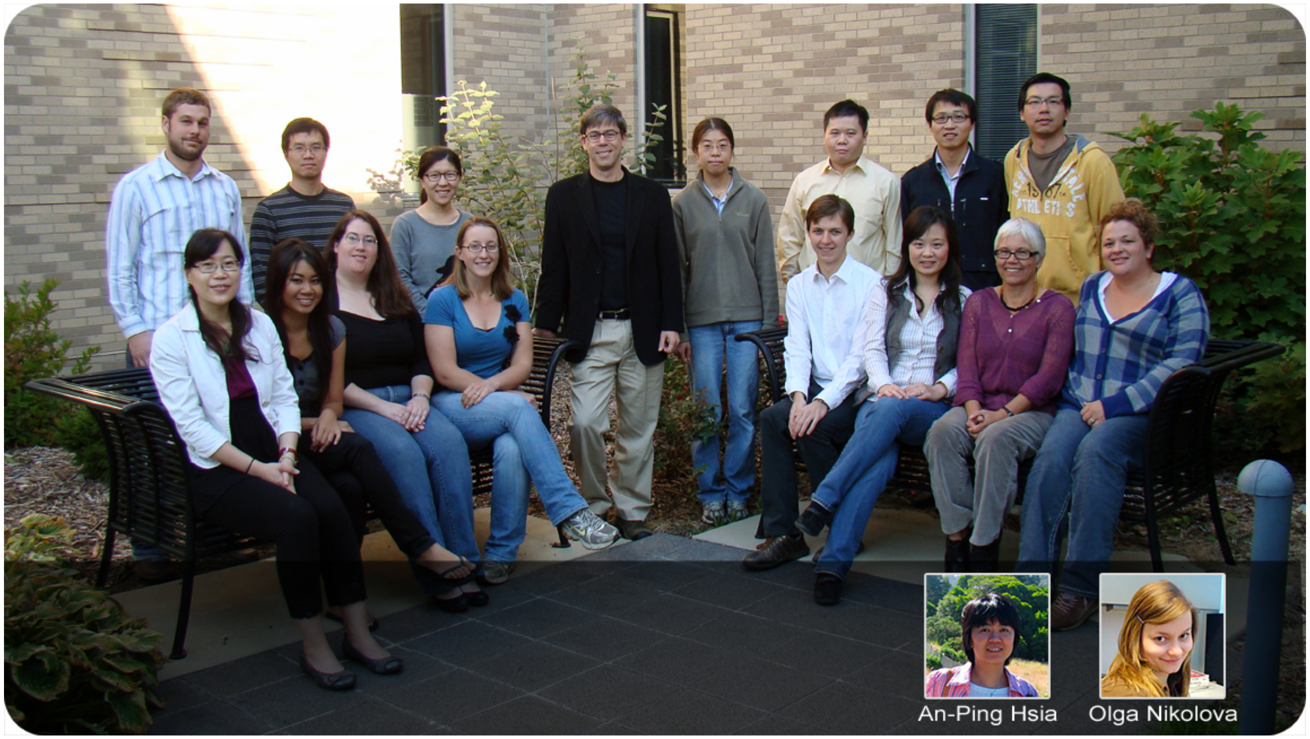
DOE Joint Genome Institute
Enabling Advances in Bioenergy & Environmental Research

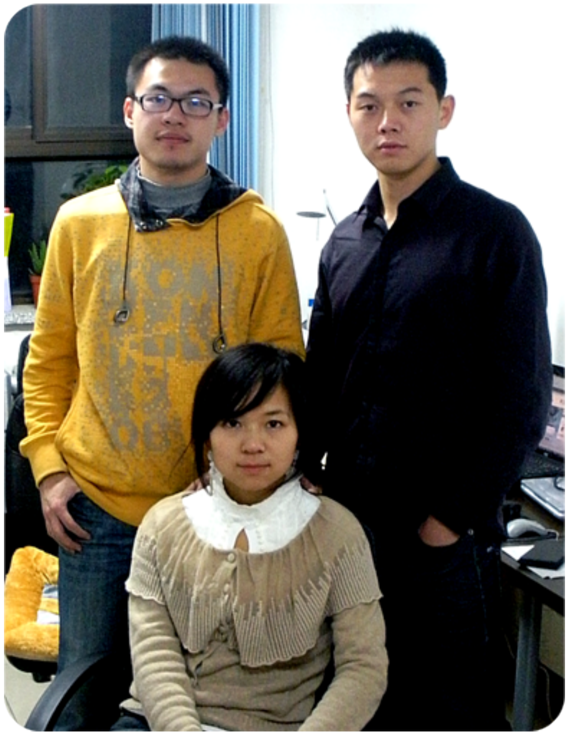
Dan Rokhsar

iPlant Collaborative™

Steve Welch, Lars Koesterke, Wacek Kusnierczyk, Mattt Vaugh, Dan Stanzione, Stephen Goff

SCHNABLE LAB
Plant Genomics





IOWA STATE UNIVERSITY

China Agricultural Univ

中國農業大學
CHINA AGRICULTURAL UNIVERSITY
1905

NSF

CSREES
EDUCATION
USDA
RESEARCH · EXTENSION