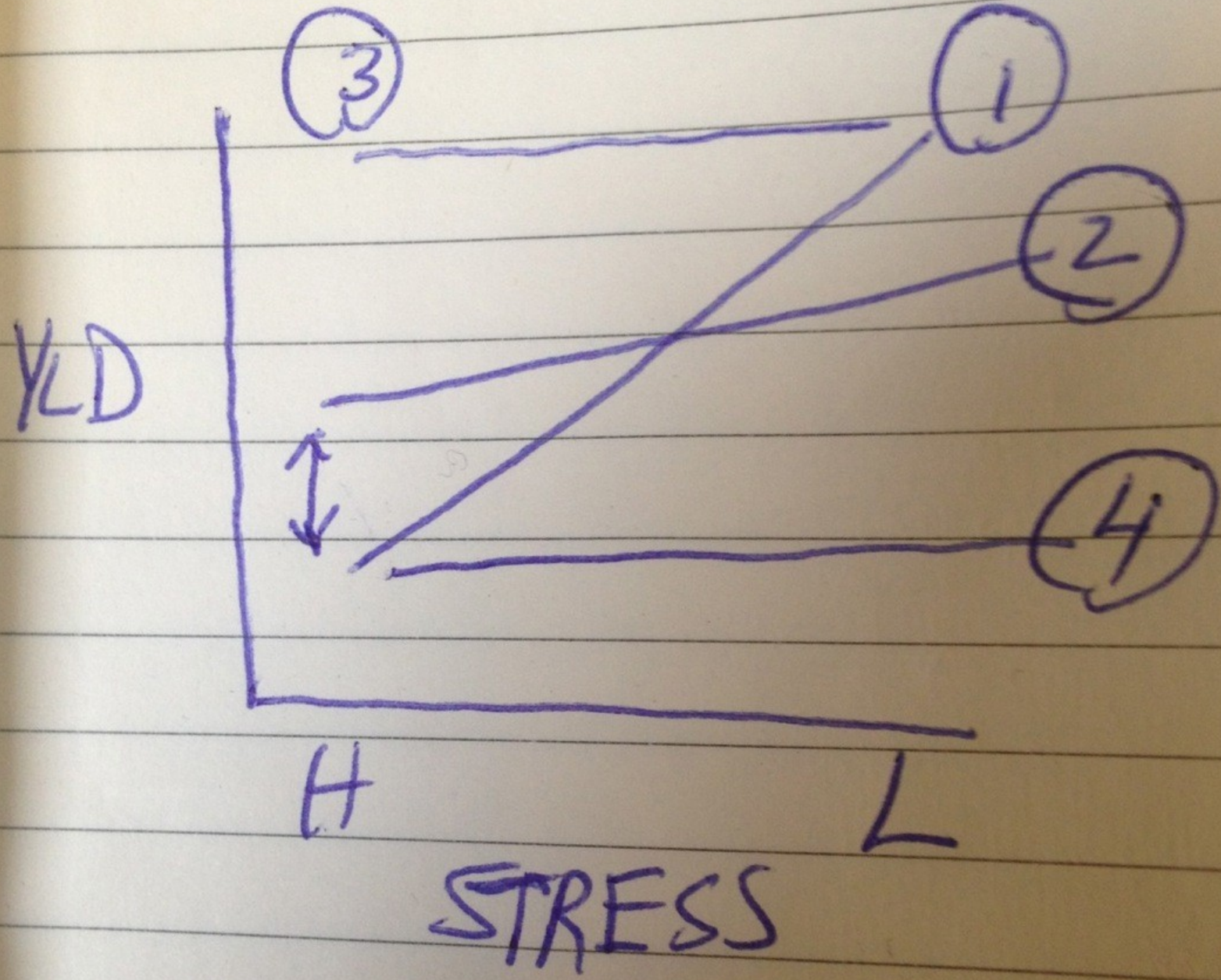
Next Generation Phenotyping and Breeding

Univ of MN
Plant Breeding
Symposium
15 March 2013

Patrick S. Schnable
Iowa State University
China Agriculture University
Data2Bio, LLC

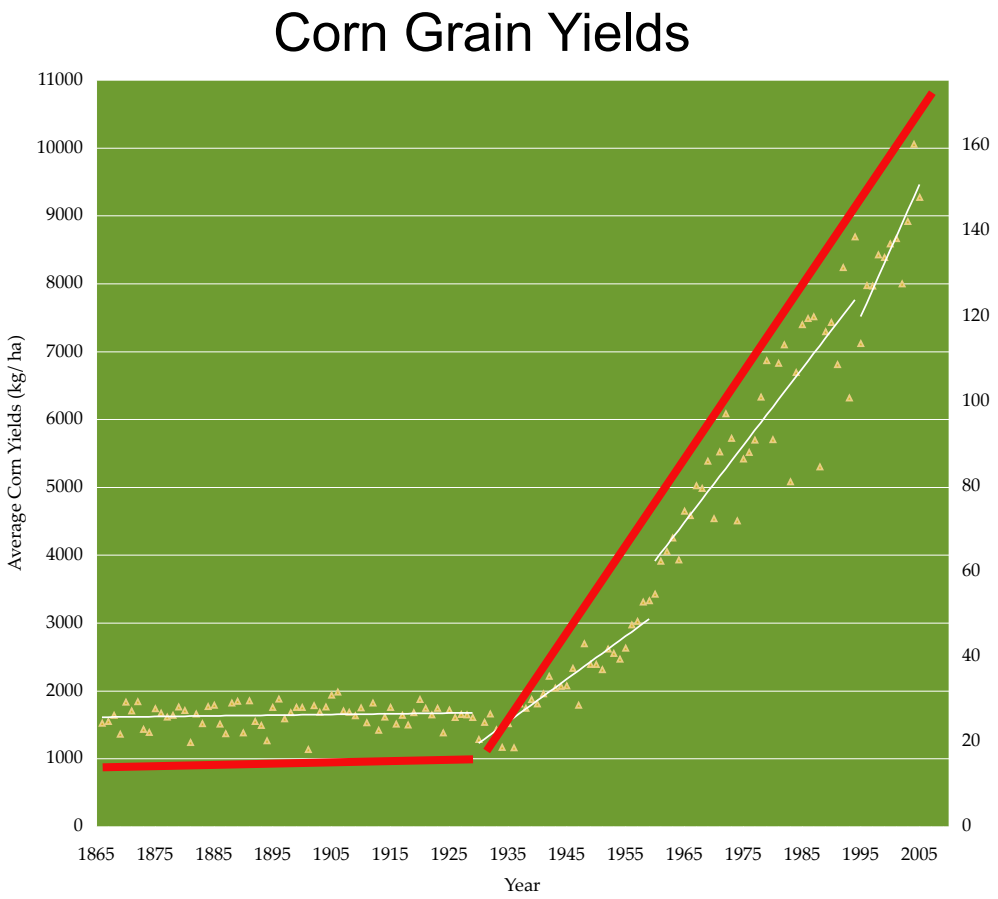

3
1
2
YLD
4
H
L
STRESS

Outline

- Next Generation Sequencing
- Next Generation Genotyping
- Next Generation Phenotyping

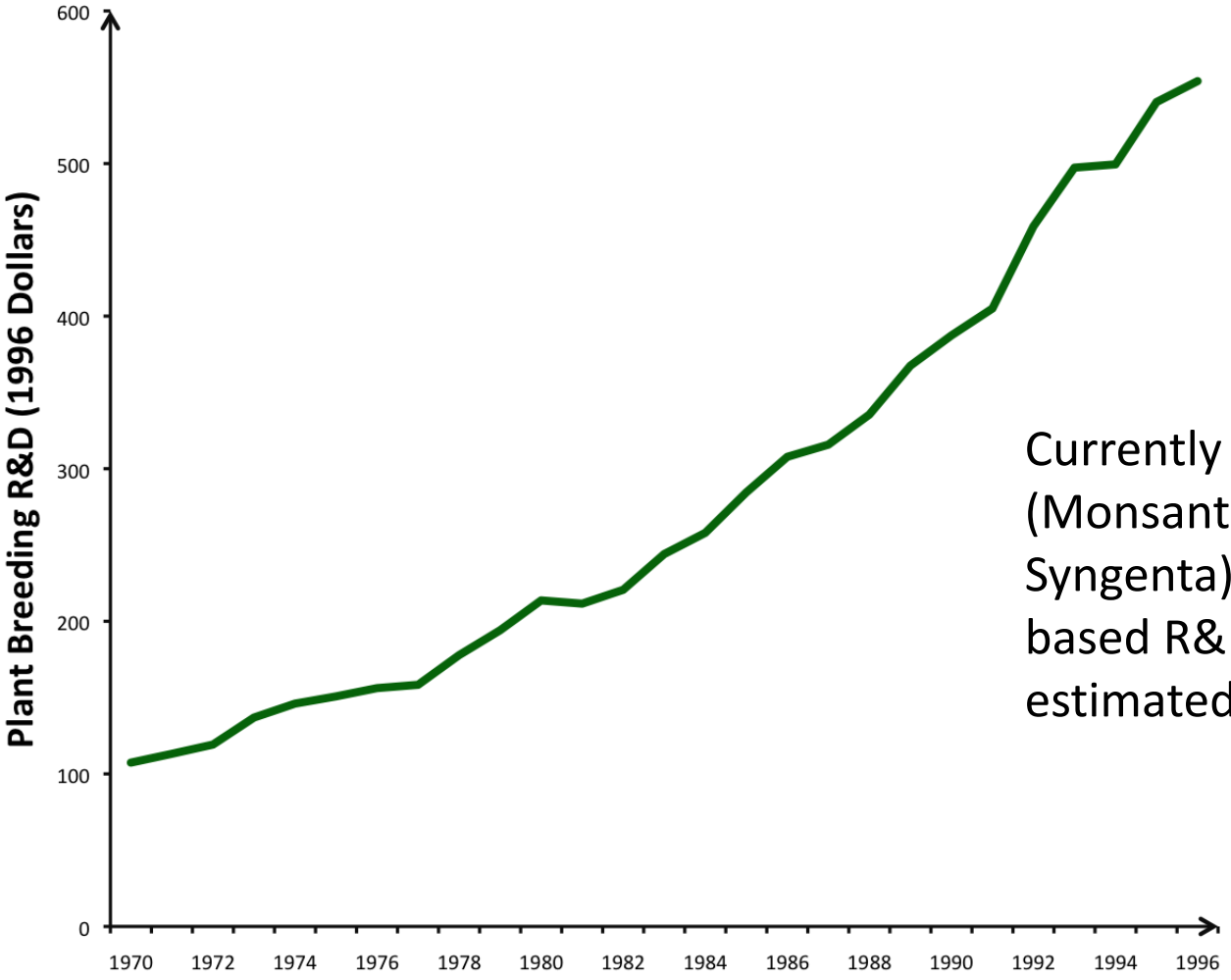
Plant breeders, agronomists and farmers have been successful

Sustained Yield Increases



Troyer (2006) Crop Sci. 46:528-543

Private-Sector R&D Investment in Plant Breeding is Increasing (inflation-adjusted constant dollars)



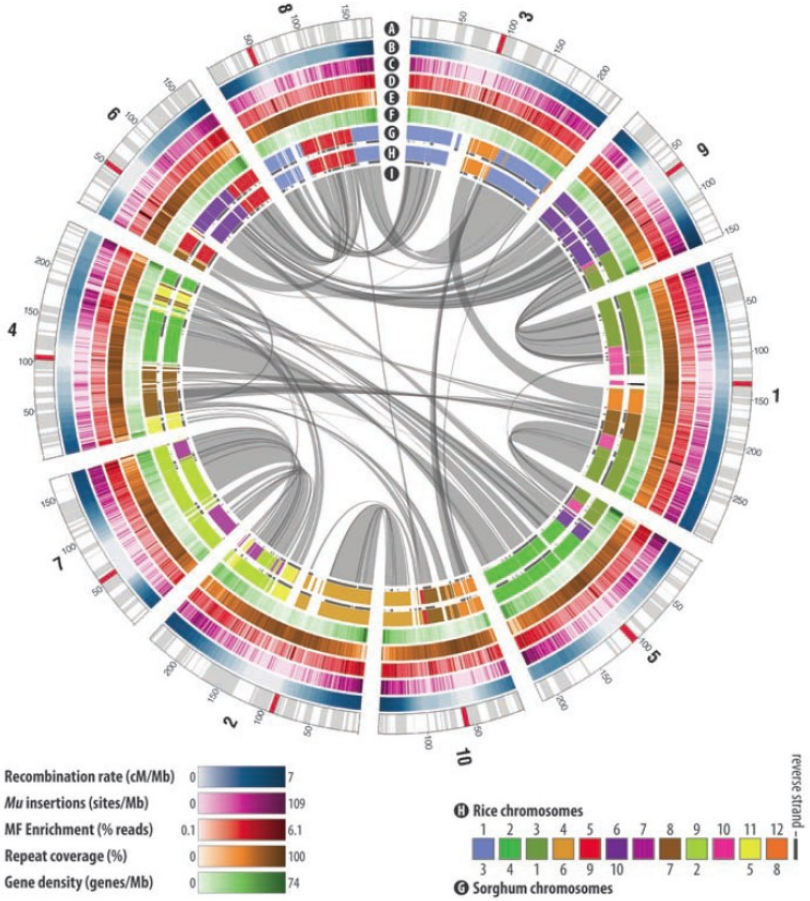
Currently (2011) the big three (Monsanto, DuPont/Pioneer, & Syngenta) have a combined seed-based R&D budget of an estimated ~$2B

USDA Report: The Seed Industry in U.S. Agriculture

Challenges Require Novel Approaches for Crop Improvement

- Agricultural inputs:
  - Increasing costs (e.g., nitrogen)
  - Reduced availability (e.g., water)
  - Undesirable ecological impacts
- Environmental stress
  - Droughts/floods
  - Temperature extremes
  - New pests & diseases

How to Translate Genomic Data into Crop/Livestock Improvement?



Schnable, Ware et al. (2009) Science

The Maize Genome Sequencing Project, Rick Wilson, PI

Maize contains ~50,000 genes

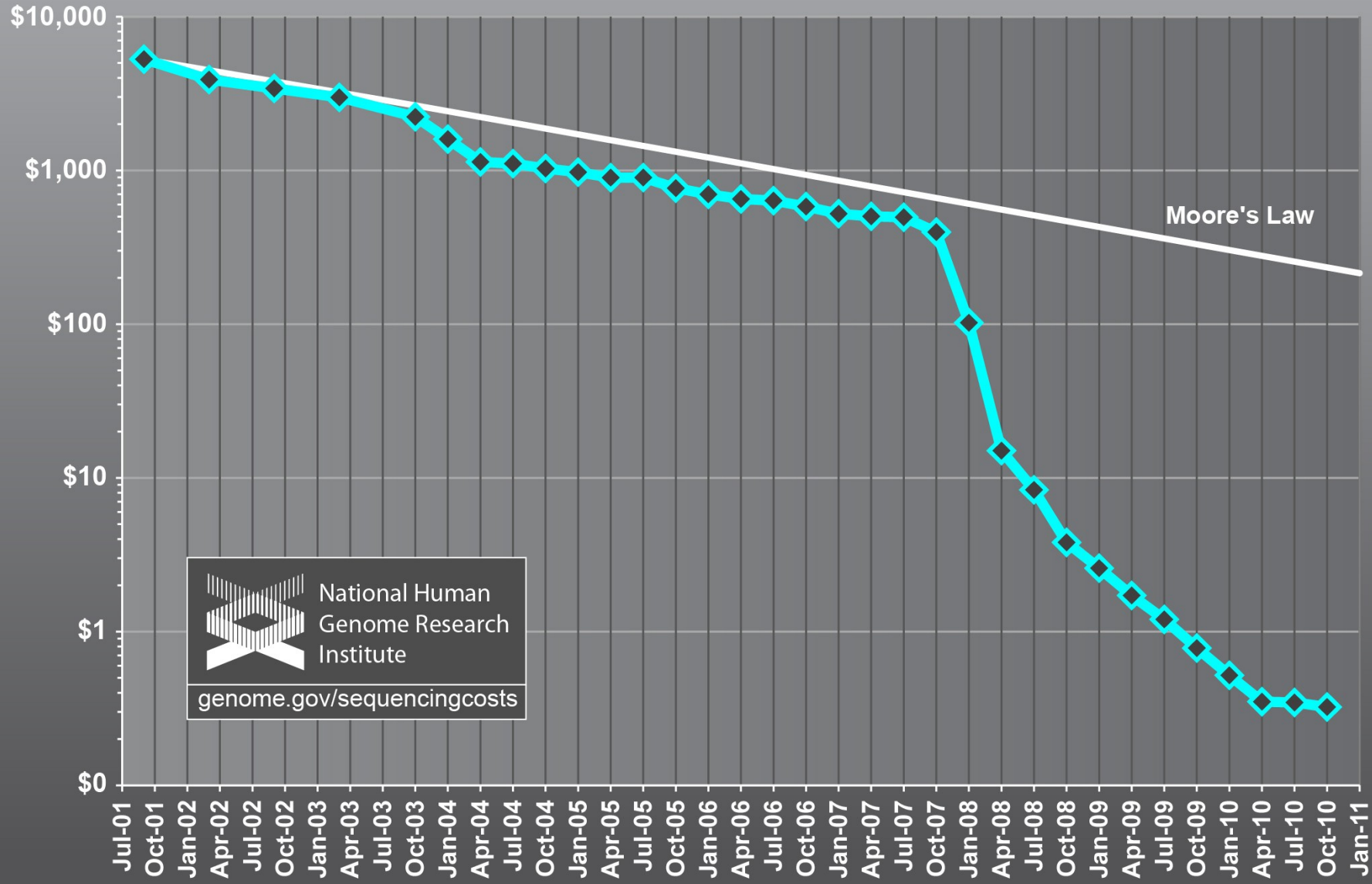
Genomic Approaches to Increase Genetic Gain

- Quantitative Trait Loci (QTL) Mapping & Genome-wide association studies (GWAS) link SNPs (markers) and genes to *qualitative* (single gene) and *quantitative (multi-genic)* traits
- Genomic Selection

Outline

- Next Generation Sequencing
- Next Generation Genotyping
- Next Generation Phenotyping

Cost per Megabase of DNA Sequence



Next Generation Sequencing (NGS) Technologies are Game Changing

- In 1980s PCR-based technologies revolutionized how we detect and quantify nucleic acids
- Currently, NGS technologies are changing how we:
  - Examine gene expression (RNA-Seq)
  - Identify polymorphisms and conduct genotyping (GBS)
  - Map and clone genes (BSR-Seq, GWAS)
  - etc.

Requirements for QTL & GWAS

- Genetic Diversity
- Genotyping Data
- Phenotypic (Trait) Data

Outline

- Next Generation Sequencing
- Next Generation Genotyping
- Next Generation Phenotyping

SNP Discovery

(Includes base substitutions, Insertions, and deletions)

SNP DISCOVERY WORKFLOW

1. **Raw Data 原始数据**
   Acquire data from internal data sets and public databases
2. **Reads Trimming 序列修剪**
   Performed using in-house short reads trimming or using lucy (--error 0.01 0.01 parameter)
3. **Repeat Filtering 循环过滤**
   Run in-house repeatmasking script and discard reads w/ >40% repeat content for 454 reads only
4. **AGPv2 Alignment 比对**
   Align to AGPv2 and Maize mitochondria (AY506529.1) and chloroplast (X86563.2)
5. **Remove Read Stacking 移除冗杂**
   Remove potential PCR artifacts from stacked reads having the same coordinates
6. **SNP Calling SNP检测**
   Call SNPs using in-house SNP calling pipeline

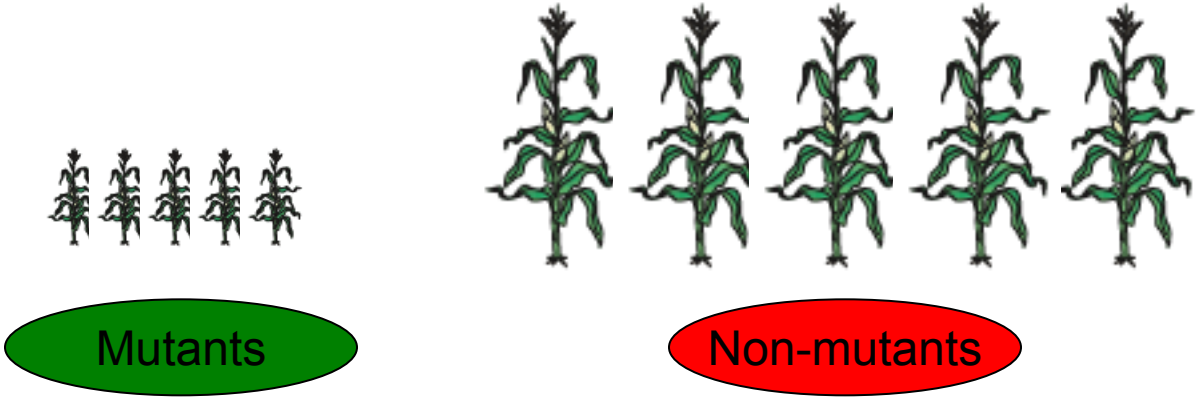
Bulked Segregant Analysis (BSA)

Objective:

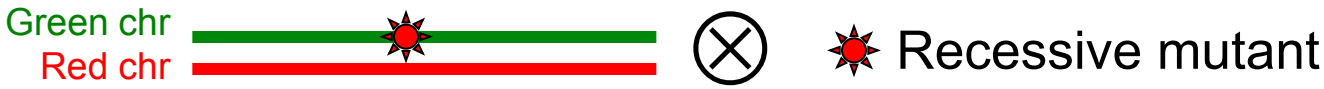
Rapidly genetically map mutant genes or QTLs
MICHELMORE, RW *et al.*, 1991, PNAS 88: 9828-9832.

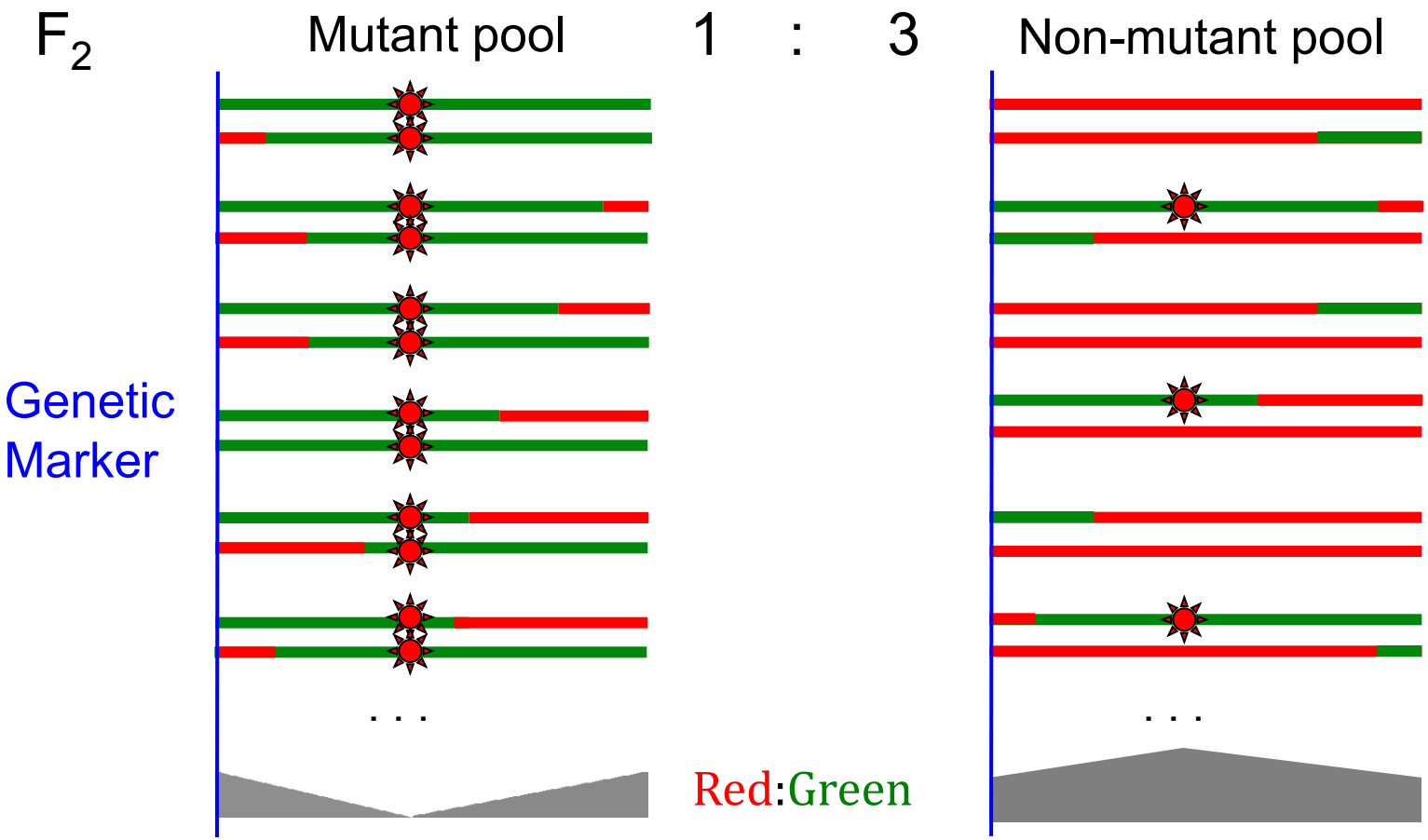
Rationale:

Linkage between genetic markers and the target gene is detected via comparative allelic quantification of genetic markers in pools of mutant and wild-type siblings.



BSA

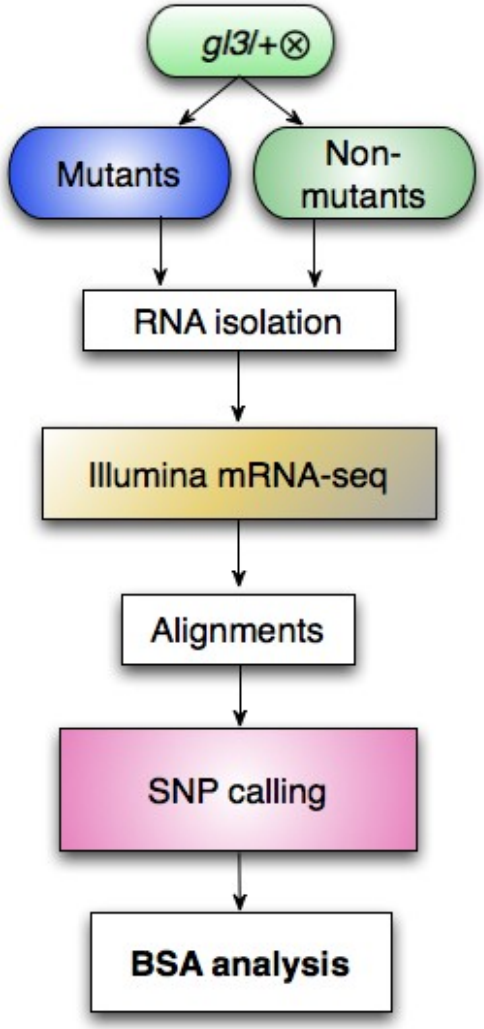




The ratio of Red:Green is proportional to the genetic distance between the mutant gene and the genetic marker.

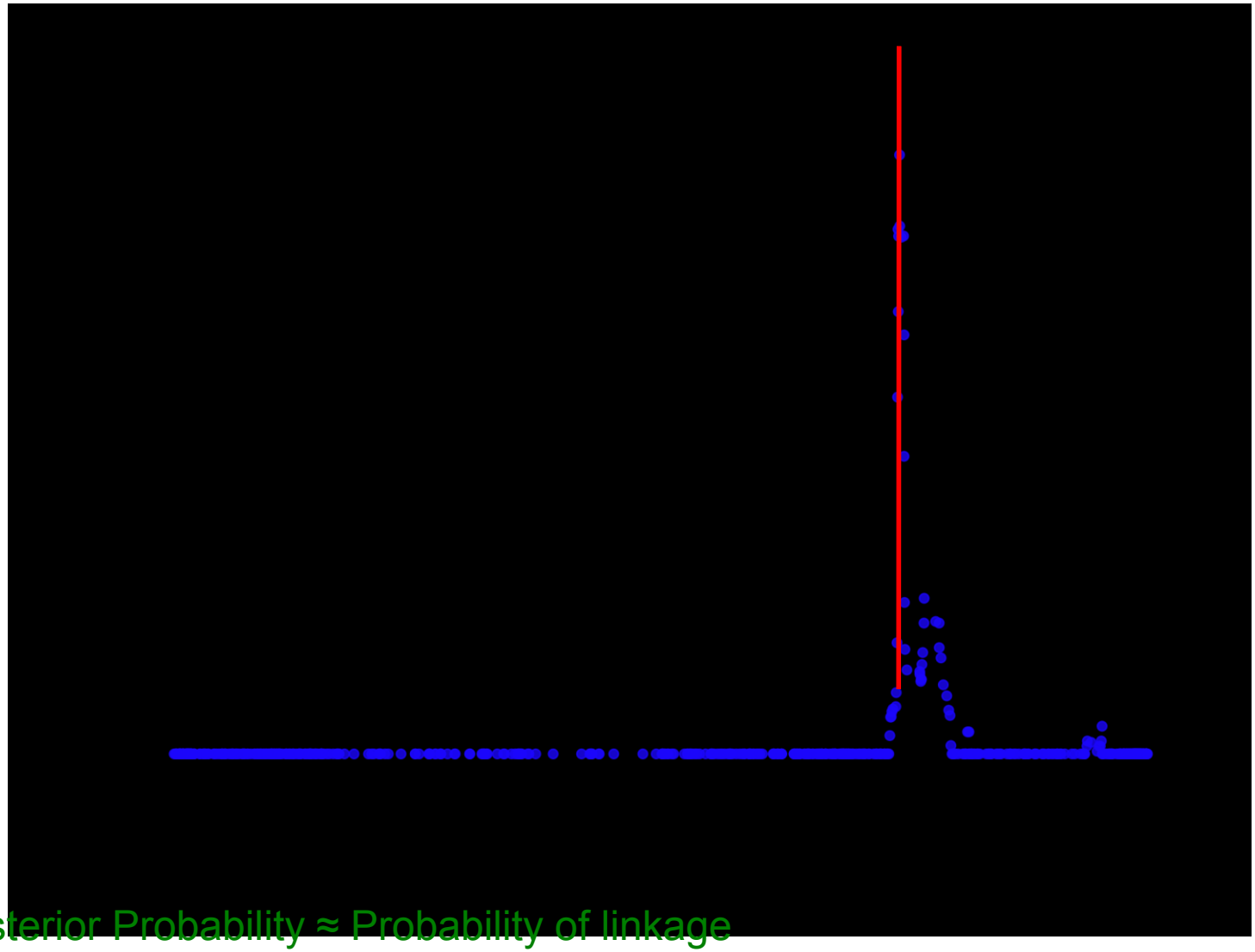
BSR-Seq: RNA-seq based BSA

- RNA-seq：
- Mutants vs. Non-mutants
- ~30 individuals
- Obtain SNP markers based on RNA-seq reads
- Map mutant gene location based on allele counts



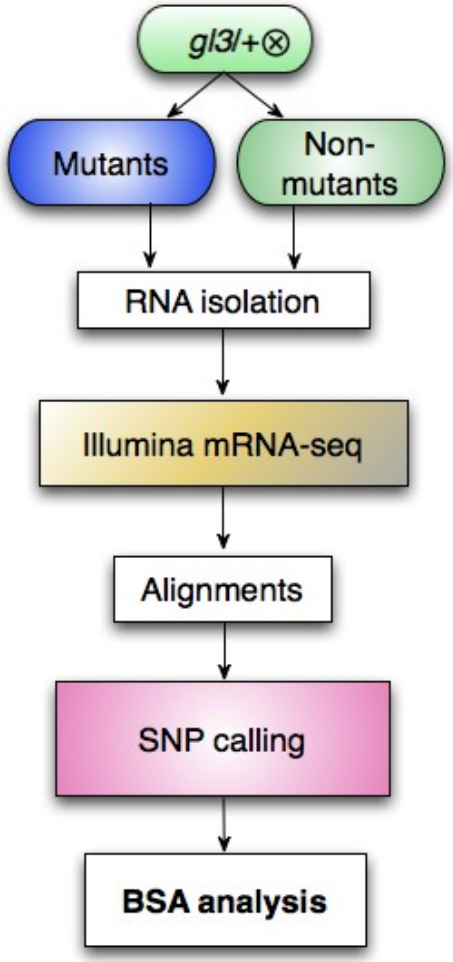
Liu et al., 2012 PLoS ONE

Location identified via BSR-Seq matches known physical location of cloned *gl3* gene



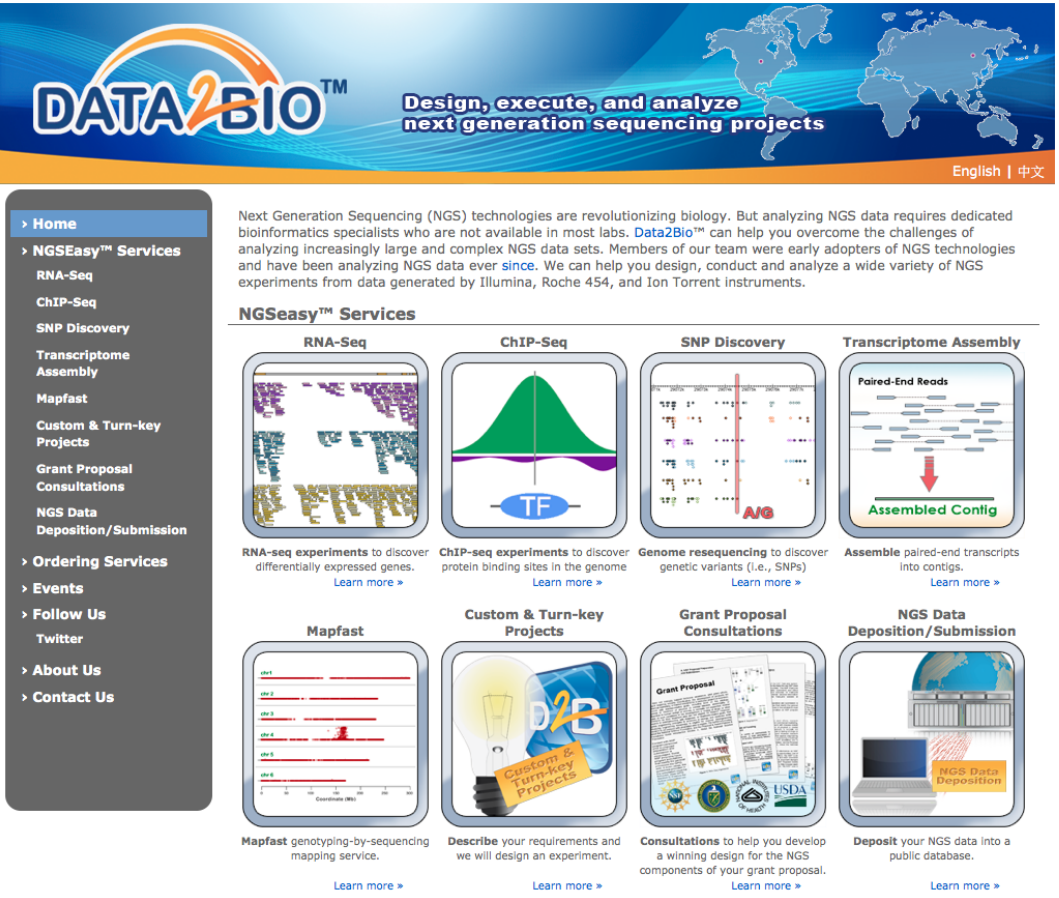
PP = Posterior Probability ≈ Probability of linkage

Advantages of BSR-Seq



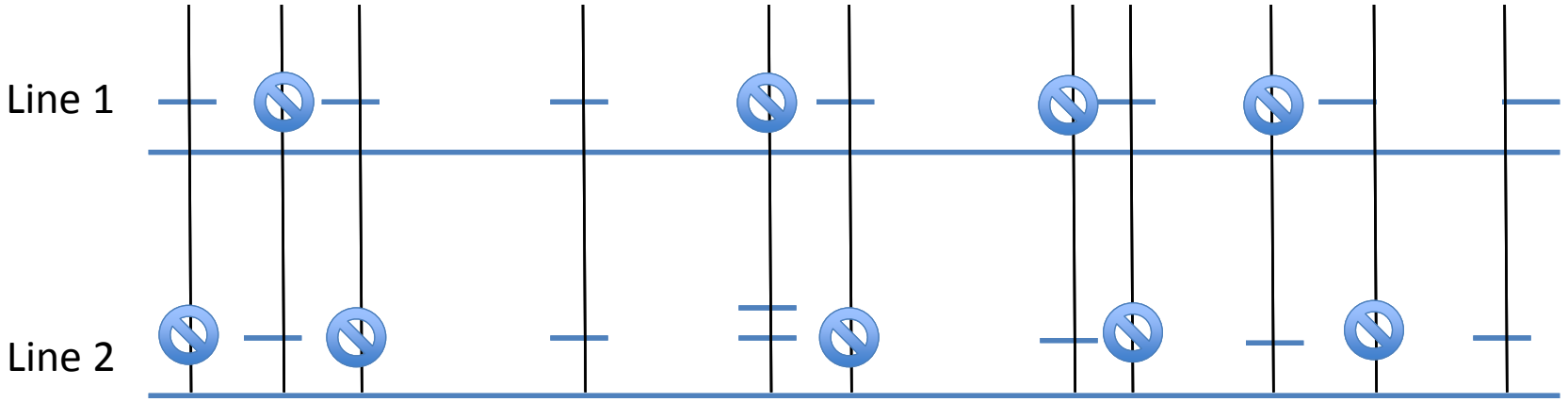
- Maps mutant gene to a small chromosomal interval
- Identifies SNPs in this chromosomal interval, which can be used for fine-scale mapping
- Defines genes in this chromosomal interval that are differentially expressed, and which are therefore candidate genes
- Yields genome-wide transcription profiles of mutants vs. non-mutants -> clues to mutant function
- Works for any species with a sequenced (*or not yet sequenced*) genome

Data2Bio, LLC



- Founded in 2010, Data2Bio helps scientists design, conduct and analyze next generation sequencing projects
- Core strengths are bioinformatics and genomics
- Offices in the US and China; Academic and private-sector customers in North American, Europe and Asia
- Proprietary bioinformatic pipelines and genomic technologies associated with DNA barcoding and genotyping-by-sequencing (GBS)

Conventional GBS



~1M sites, i.e., more than is required for most applications

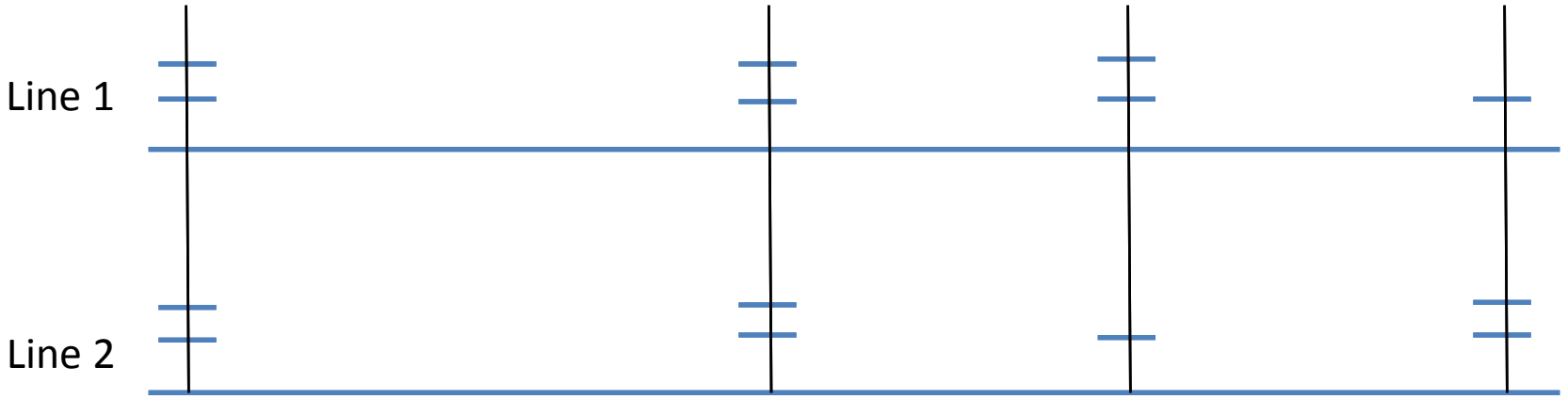
SNP calls based on few reads (error rate)

Can not call heterozygotes

Limited overlap of sites among lines -> missing data across lines

Requires *imputation* (heavy-duty bioinformatics)

tGBS: Value of Read Selection



tGBS more stringently controls fraction of genome that is sequenced and genotyped ("read selection"). Hence, given the same number of reads/line, reads are clustered at fewer sites:

- *Less missing data* across lines
- 1,000-60,000 SNPs ("tunable") -> adequate for most applications
- *High confidence SNP calls*
- *Can confidently genotype heterozygous loci*

Outline

- Next Generation Sequencing
- Next Generation Genotyping
- Next Generation Phenotyping

Contributions to Phenotype (traits)

- Phenotype (P) = Genotype (G) + Environment (E) + GxE
- E and GxE complicate predictions of hybrid performance (P) from genotype (G)
- Two strategies for dealing with “E” and “GxE”

1st Strategy: Controlled Environments

- Conduct phenotyping under controlled environments (“E”), e.g., growth chambers or greenhouses
- This can remove E (and GxE) from the equation, allowing accurate prediction of the performance of a given hybrid *in a given growth chamber or greenhouse*

2nd Strategy: Field Conditions

- Conduct phenotyping under *relevant field conditions*, but this requires growing genotypes in *many characterized environments* (and appropriate experimental design and statistical analyses)
- But if we want to *understand and model* the influences E and GxE on hybrid performance (P) we need to conduct experiments under conditions that match production conditions as closely as possible

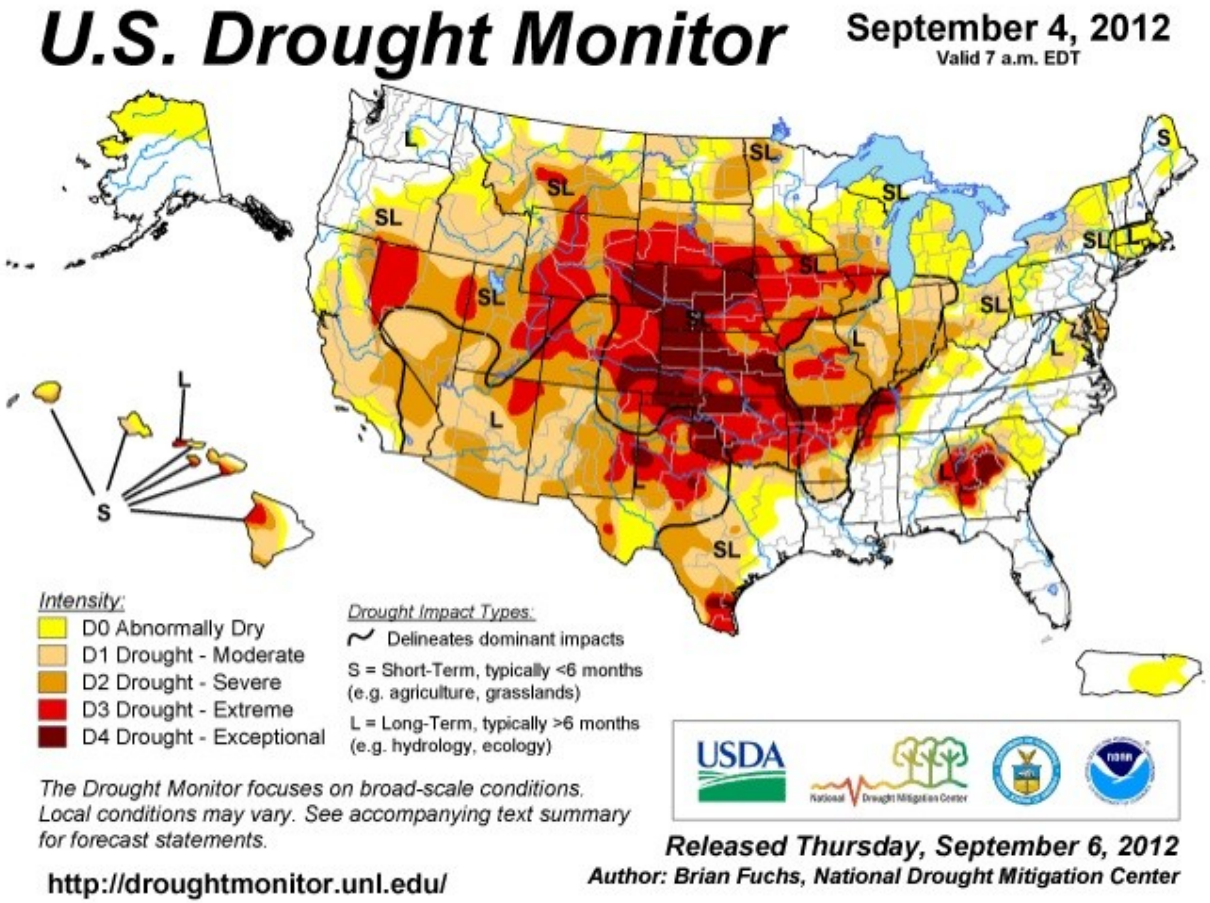
Predictive Models for Hybrid Performance

- Once we understand the relationships and interactions among genotype, environment and phenotype (P= G + E + GxE) we will be able to:
  - More efficiently develop hybrids that exhibit yield stability across environments and/or that are engineered for particularly challenging environments
  - Much more accurately predict the performance of a given hybrid in a given field in a given year

What is required to define the effects of G, E and GxE on phenotype?

- Measure traits from *many* genotypes, in *many* diverse (and commercially relevant) environments
- Quantity of data to be collected *requires new phenotyping technologies and strategies*

Number of locations matters (a lot):



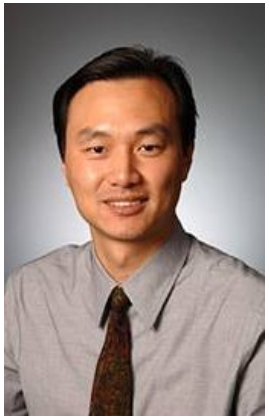
Vision #1 for High-Throughput, *Field-based* Phenotyping

- Conduct a coordinated phenotyping project involving MANY regional stations across the U.S. (e.g., at Land Grant Universities) to study how corn performs under diverse environmental conditions
- Possibility for public-private partnership
- Develop novel high-throughput field-based phenotyping technologies

IOWA STATE UNIVERSITY

Patrick Schnable

Maize Genomics
Agronomy



Lie Tang

Robotics Specialist
Ag & Biosystems
Engineering

Maria Salas Fernandez

Sorghum Breeder
Agronomy

USDA United States Department of Agriculture National Institute of Food and Agriculture

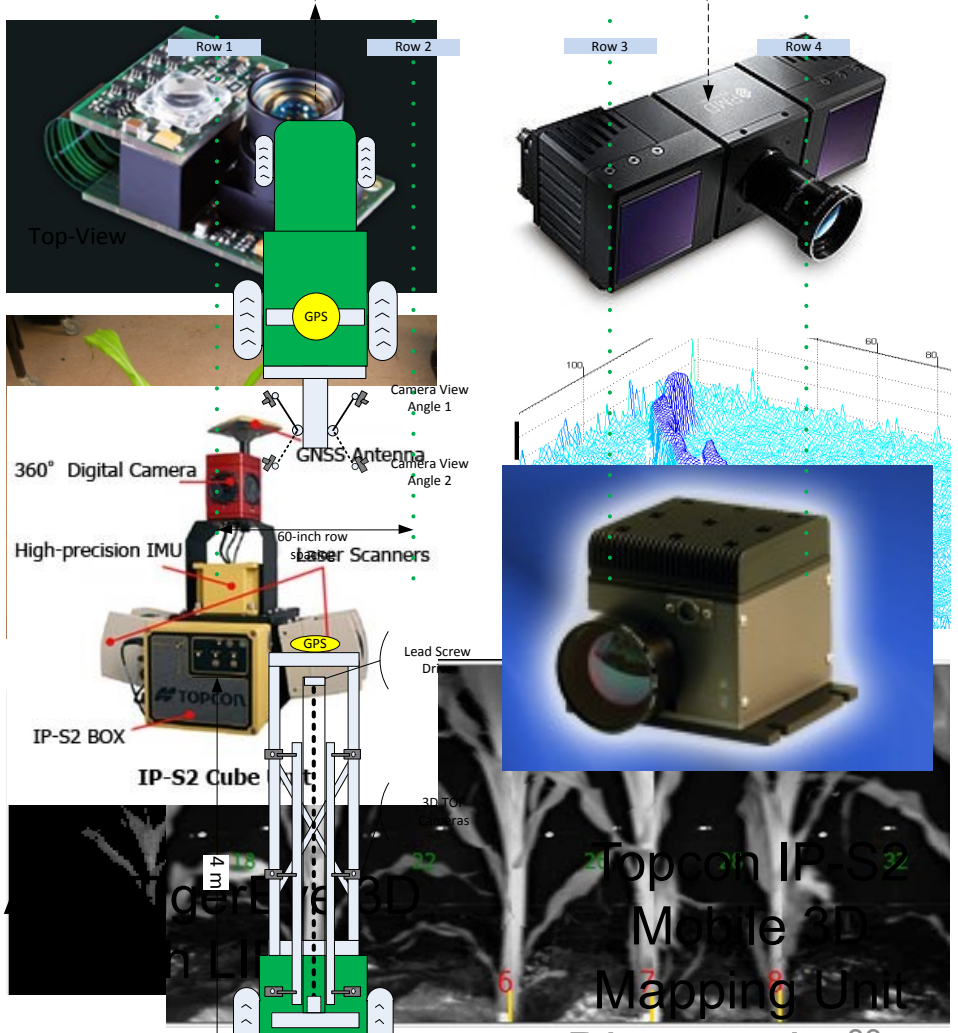
Grant #: 2012-67009-19713

Design of the automated phenotyping robot for biomass yield-related architecture traits

Mobile Carrier with RTK-GPS

John Deere Sub-Compact Utility Tractor Equipped with Topcon Universal Auto-Steer System

Multiple Light 3D Time-of-Flight Cameras



Identify sorghum biomass yield traits via genome-wide association studies (GWAS)

Diversity Panel

387 photoperiod insensitive lines

- Converted tropical sorghums
- Elite materials
- Historically important elite lines

Yu Panel

300 photoperiod sensitive lines

- A subset lines selected based on biomass potential.

Vision #2 for HTP, Field-based Phenotyping is based on validated social networking/local review paradigms

- Yelp: cell phone app that provides restaurant reviews
- Tens of thousands of individual users *create value for all users* by contributing their individual data (i.e., restaurant reviews) to a shared database
- Tremendously successful (71M unique visitors/month); advertising-driven business model

Vision #2 for High-Throughput, Field-based Phenotyping

- Farmer Participatory Phenotyping (FP2)
- Many farmers already have GPS and yield monitors on their combines
- Collect consistent data at high spatial resolution from commercial production fields; upload to ISU servers

Farmer Participatory Phenotyping

- Combine yield data with soil maps, weather data etc.
- With thousands of potential participants it will be possible to collect unprecedented amounts of data
- Information technology resources will be vital to store and interpret the vast amounts of collected data
- Develop statistical models that accurately predict phenotypes (hybrid performance) under specific environmental conditions
- Determine which hybrids are suited for particular environments (soil types, moisture levels, fertilization rates, and other agronomic practices)

Resulting Predictive Models will:

- Be relevant to production fields
- Enhance our ability to efficiently produce superior hybrids
- Improve ability of farmers to select appropriate hybrids for local conditions
- Improve ability to predict yields *prior to planting* and *during* growing season, reducing volatility of commodity prices

Longer-term Vision for FP2

- Develop novel sensors that can detect more subtle phenotypes that could be placed on tillage equipment or center pivot irrigation booms to collect stress phenotypes (water/nutrient/disease/insect) early in season

Value of FP2 to Farmers

- Because farmer-participants in FP2 will be able to access phenotypic data for each commercially available corn variety from thousands of fields with known soil types and that are being managed with different (but defined) agronomic practices, the process by which farmers select varieties for planting and the agronomic practices they employ will be driven by a rich set of unbiased data, thereby maximizing farmer profits

Bottom Line

- Advances in Next Generation Sequencing technologies have allowed for the development of Next Generation Genotyping technologies (e.g., multiple flavors of GBS)
- In turn, Next Generation Genotyping technologies offer the potential to greatly increase the efficiency of plant breeding and the development of hybrids with enhanced stress resistant and/or yield stability
- But to realize the benefits of Next Generation Genotyping we now need to develop “Next Generation Phenotyping” strategies:
  - Controlled conditions (e.g., greenhouses)
  - Field based
    - Coordinated regional phenotyping stations (N=hundreds)
    - Farmer Participatory Phenotyping, FP2 (N=thousands)
  - New data collection technologies/sensors

Collaborators

IOWA STATE UNIVERSITY

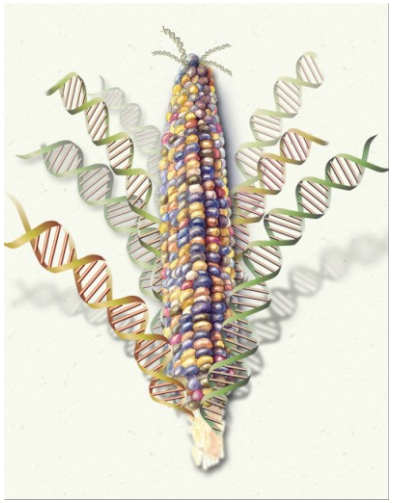
Srinivas Aluru
Hui-Hsien Chou
Carolyn Lawrence
Dan Nettleton

UF UNIVERSITY of FLORIDA

Brad Barbazuk

UNIVERSITY OF MINNESOTA

Nathan Springer



The Maize Genome Sequencing Project

NimbleGen HIGH - DEFINITION GENOMICS™ Roche

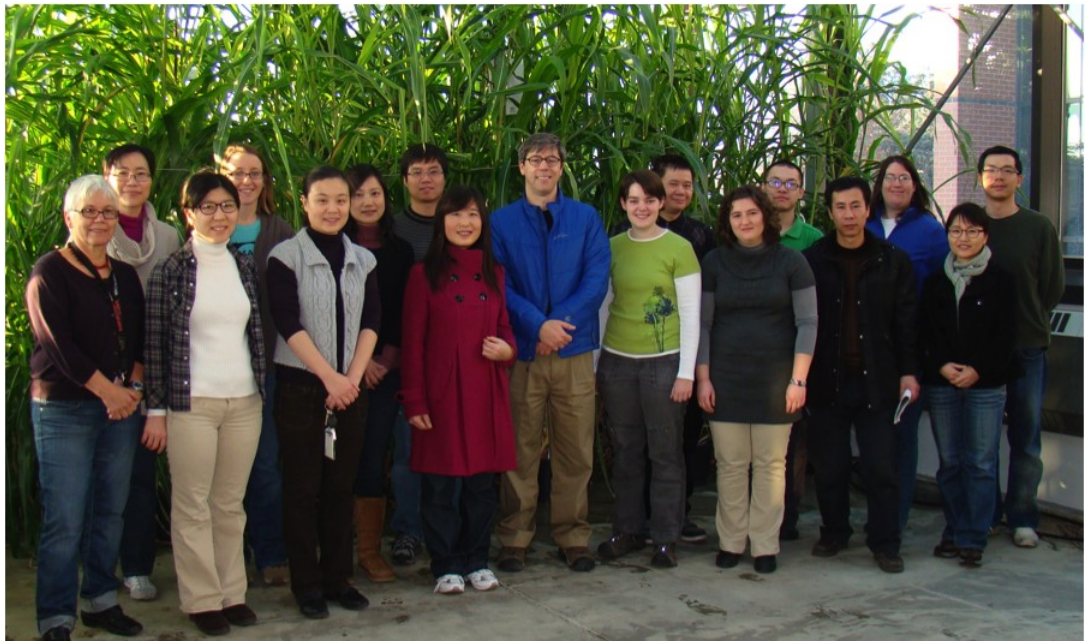
Jeffrey Jeddeloh

SAM MAIZE SHOOT APICAL MERISTEM

Mike Scanlon (PI, Cornell)
Jianming Yu, M. Timmermans,
G. Muehlbauer,
D. Jannick-Buckner

中國農業大學 CHINA AGRICULTURAL UNIVERSITY 1905

Jinsheng Lai



SCHNABLE LAB
Plant Genomics

IOWA STATE UNIVERSITY
Plant Sciences Institute

IOWA STATE UNIVERSITY

NSF

USDA

life technologies

PSS has substantial IP and equity interests in Data2Bio LLC

CHINA AGRICULTURAL UNIVERSITY