Agricultural Applications for Genome Sequencing

AEIC April 2012

Presented by Joe Clarke, NGS Platform Lead, Syngenta RTP NC

Syngenta at Home and Abroad

Who we are and what we do

Syngenta is one of the world's leading companies with more than 26,000 employees in over 90 countries dedicated to our purpose: Bringing plant potential to life.

Our Crop Protection and Seeds products help growers increase crop yields and productivity. We contribute to meeting the growing global demand for food, feed and fuel and are committed to protecting the environment, promoting health and improving the quality of life.

Global R&D capabilities

JEALOTT'S HILL

Chemical Discovery
Weed Control
Formulation
Bioscience
Environmental Science

BEIJING
Syngenta
Biotechnology

GOA

Chemistry

GREENSBORO

Formulation
Environmental Science

RTP
Syngenta
Biotechnology

STEIN

Fungicides, Insecticides & Professional Products

Other sites:

- Marker-assisted and seed breeding capabilities
- Global field station network

syngenta

Syngenta Biotechnology

Proven delivery of biotech innovation with industry firsts

Delivering innovation is the focus of Syngenta scientists, who use a combination of bioscience and cutting-edge technology to develop innovative solutions that help farmers, food companies, and consumers meet tomorrow's challenges.

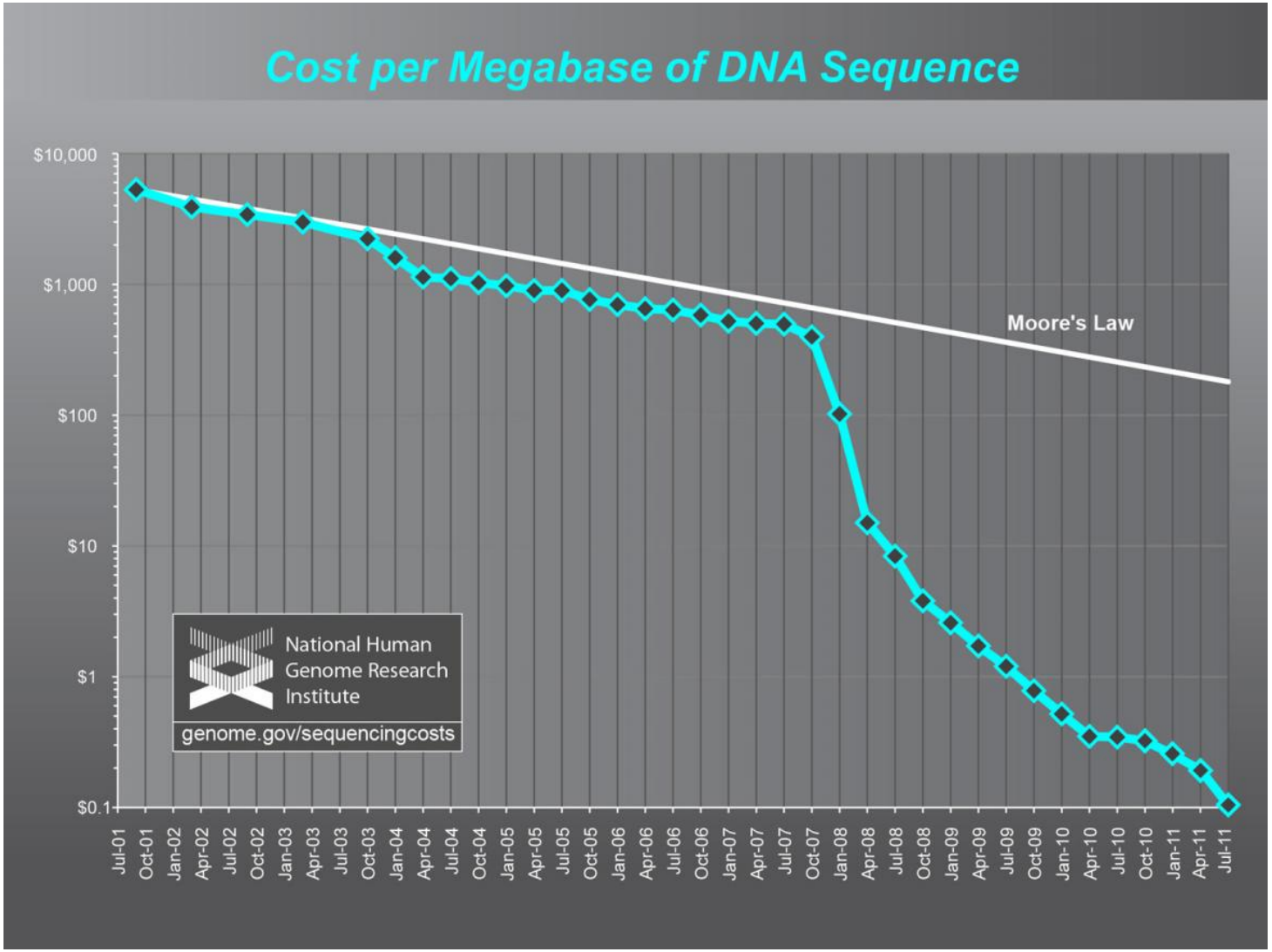
- Formed in 1984
- Located in Research Triangle Park, NC
- Strong ties with the community and local academic institutions
- State-of-the-art facilities
  - 120,000 sq ft of lab space
  - 50,000 sq ft of greenhouse space
  - 27,000 sq ft of office space
  - 100,000 sq ft of new office space
- Approximately 400 employees
  - Expansion underway

Next-Gen Sequencing and Genome Assemblies

A series of short stories describing platforms and capabilities

- A series of short stories describing downstream applications for genome assemblies
  - Brief overview of genome assembly
  - Highlighting projects requiring different assembly strategies
    - Marker discovery and trait association to gene cloning
    - Genotype by sequencing for RIL mapping
    - Population structure and GWAS analysis

The price point for sequencing continues to drop



Assemblers and Aligners

Table 1
Feature comparison between de novo assemblers for whole-genome shotgun data from next-generation sequencing platforms. OLC refers to the overlap/layout/consensus architecture. DBG refers to the de Bruijn graph architecture. The table is based on the literature cited in the text. It may not reflect the current state of each software package.

Algorithm Feature	Greedy Assemblers	OLC Assemblers	DBG Assemblers
Modeled features of reads			
Base substitutions			Euler, AllPaths, SOAP
Homopolymer miscount		CABOG	
Concentrated error in 3′ end			Euler
Flow space		Newbler	
Color space		Shorty	Velvet
Removal of erroneous reads			
Based on K-mer frequencies			Euler, Velvet, AllPaths
Based on K-mer freq and QV			AllPaths
For multiple values of K			AllPaths
By alignment to other reads		CABOG	
By alignment and QV	SHARCGS		
Correction of erroneous base calls			
Based on K-mer frequencies			Euler, SOAP
Based on Kmer freq and QV			AllPaths
Based on alignments		CABOG	
Approaches to graph construction			
Implicit	SSAKE, SHARCGS, VCAKE		
Reads as graph nodes		CABOG, Newbler, Edena	
K-mers as graph nodes			Euler, Velvet, ABySS, SOAP
Simple paths as graph nodes			AllPaths
Multiple values of K			Euler
Multiple overlap stringencies	SHARCGS		
Approaches to graph reduction			
Filter overlaps		CABOG	
Greedy contig extension	SSAKE, SHARCGS, VCAKE		
Collapse simple paths		CABOG, Newbler	Euler, Velvet, SOAP
Erosion of spurs		CABOG, Edena	Euler, Velvet, AllPaths, SOAP
Transitive overlap reduction		Edena	
Bubble smoothing		Edena	Euler, Velvet, SOAP
Bubble detection			AllPaths
Reads separate tangled paths			Euler, SOAP
Break at low coverage			Velvet, SOAP
Break at high coverage		CABOG	Euler
High coverage indicates repeat		CABOG	Velvet
Special use of long reads		Shorty	Velvet
Graph partitions			
Partition by K-mers			ABySS
Partition by scaffolds			AllPaths
Uses for mate pairs			
Constrain path searches			Euler, Velvet, AllPaths
Guide path selection			Euler, Allpaths
Detect misassembled contigs		CABOG, Shorty	
Merge contigs or fill gaps		CABOG, Shorty	Velvet, ABySS, SOAP
Transitive link reduction		CABOG	SOAP
Detect, avoid repeat contigs		CABOG	Velvet, SOAP
Create scaffolds		CABOG, Shorty	Euler, Velvet, AllPaths, SOAP

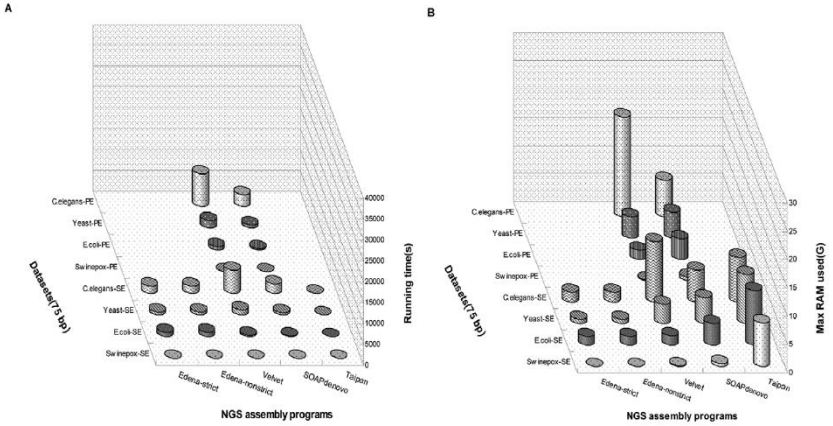


Figure 3. Computational running time and maximum memory occupancy of 75-mer short reads assembly procedures. (A) the computational times of each assembler for different datasets. (B) the maximum RAM used during the assembly process. No data is shown when the RAM is insufficient or the assembly tool is not suitable for the dataset.
doi:10.1371/journal.pone.0017915.g003

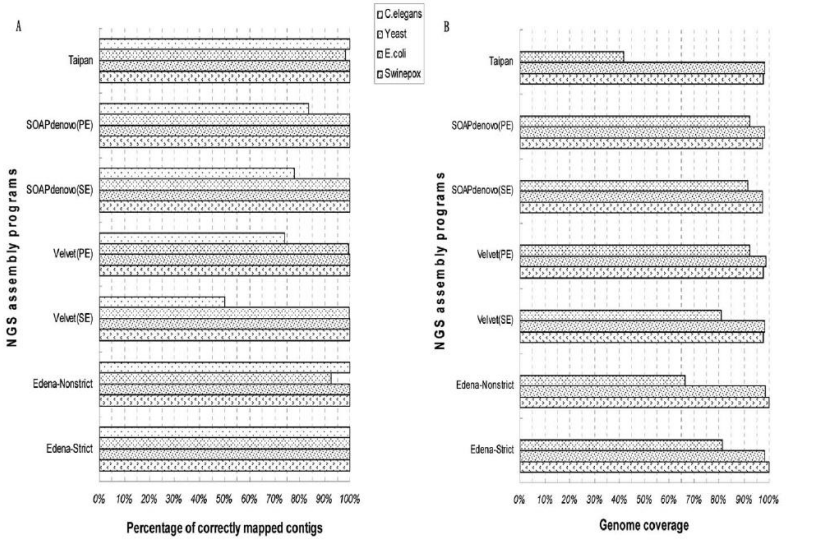
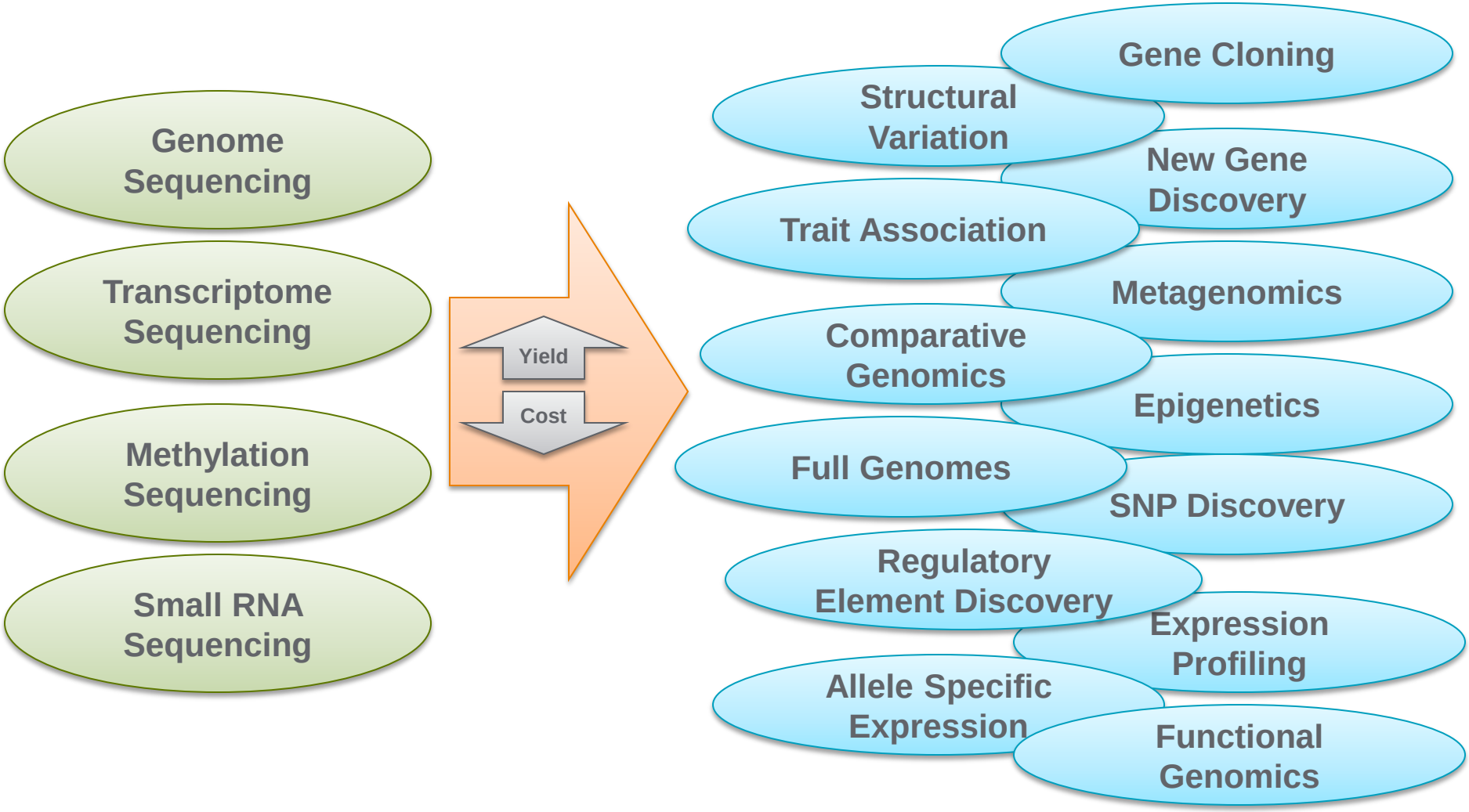


Figure 5. Accuracy and integrity for 75-mer datasets assembly. For short reads assembly, accurate and high genome coverage contigs are expected. Here, the quality of consequential contigs is shown with (A) the accuracy of assembled contigs and (B) the genome coverage of the assembled contigs. No data is shown when the RAM is insufficient or the assembly tool is not suitable for the dataset.
doi:10.1371/journal.pone.0017915.g005

Miller et al., Genomics 95 (2010) Zhang et al., Plos One 6 (2011)

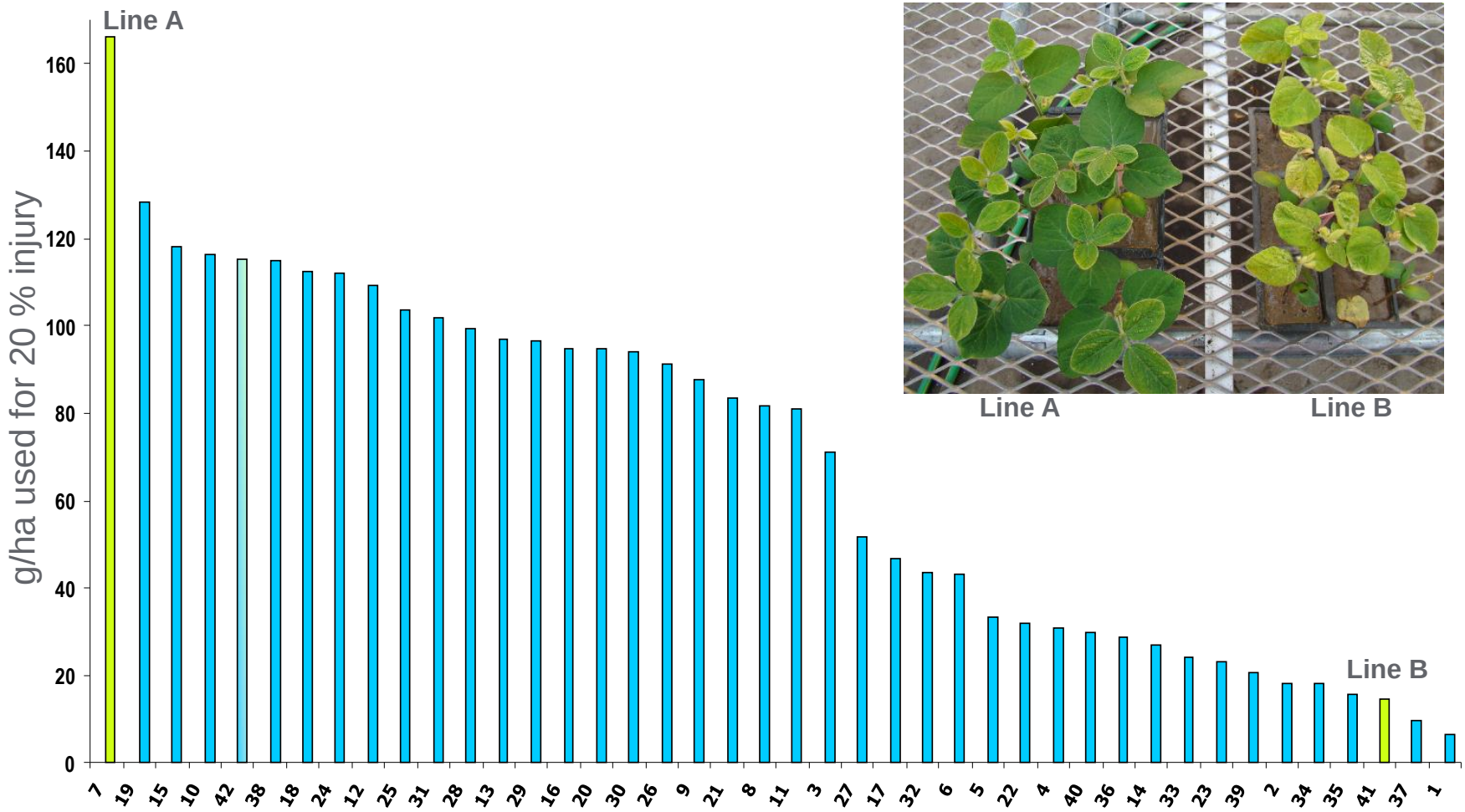
It is not about the sequence or the assembly



Shannon McDonald
Ernie Chilcott
Molly Dunn
Chris Basten
Becky Breitinger
Ju-Kyung Yu
Joe Curley
Harish Ghandi
Joe Clarke
Bernard Vernooij
Sheng Sheng Zhang
Lynn Senior
John Hipskind
Metabolon

Marker Discovery and Trait Association to Functional Validation

Syngenta conventional soybeans have range of tolerance to a soil applied herbicide



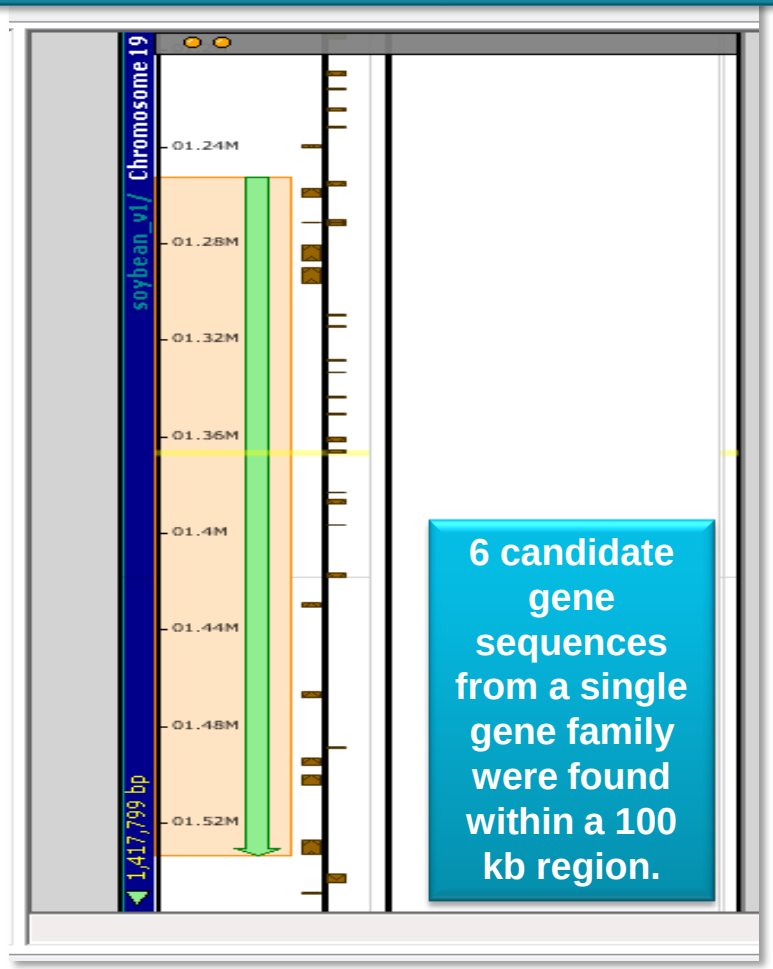
Bulked segregant analysis

- Cross highly susceptible and highly tolerant lines and make a segregating mapping population.
- Use highly tolerant and highly susceptible progenies to make pools which should be genotypically different at the trait linked regions only.
- Genotype the pools and identify markers to fine map major QTL.
- Search for candidate genes.
- Sequence candidates in segregating populations to look for polymorphisms that segregate perfectly with trait.

As new technologies appeared, we adapted - Solexa sequence - based BSA

- Pools were created from the tolerant and susceptible lines as before.
- RNA extracted
- cDNA synthesized
- The transcriptome of the pools was sequenced
- SNPs specific to each pool were identified
- The same major QTL from the previous mapping study was detected
- New markers were identified that could be used to fine map the QTL .

17 new markers were identified and helped us fine map the region to 1 Mb



6 candidate gene sequences from a single gene family were found within a 100 kb region.

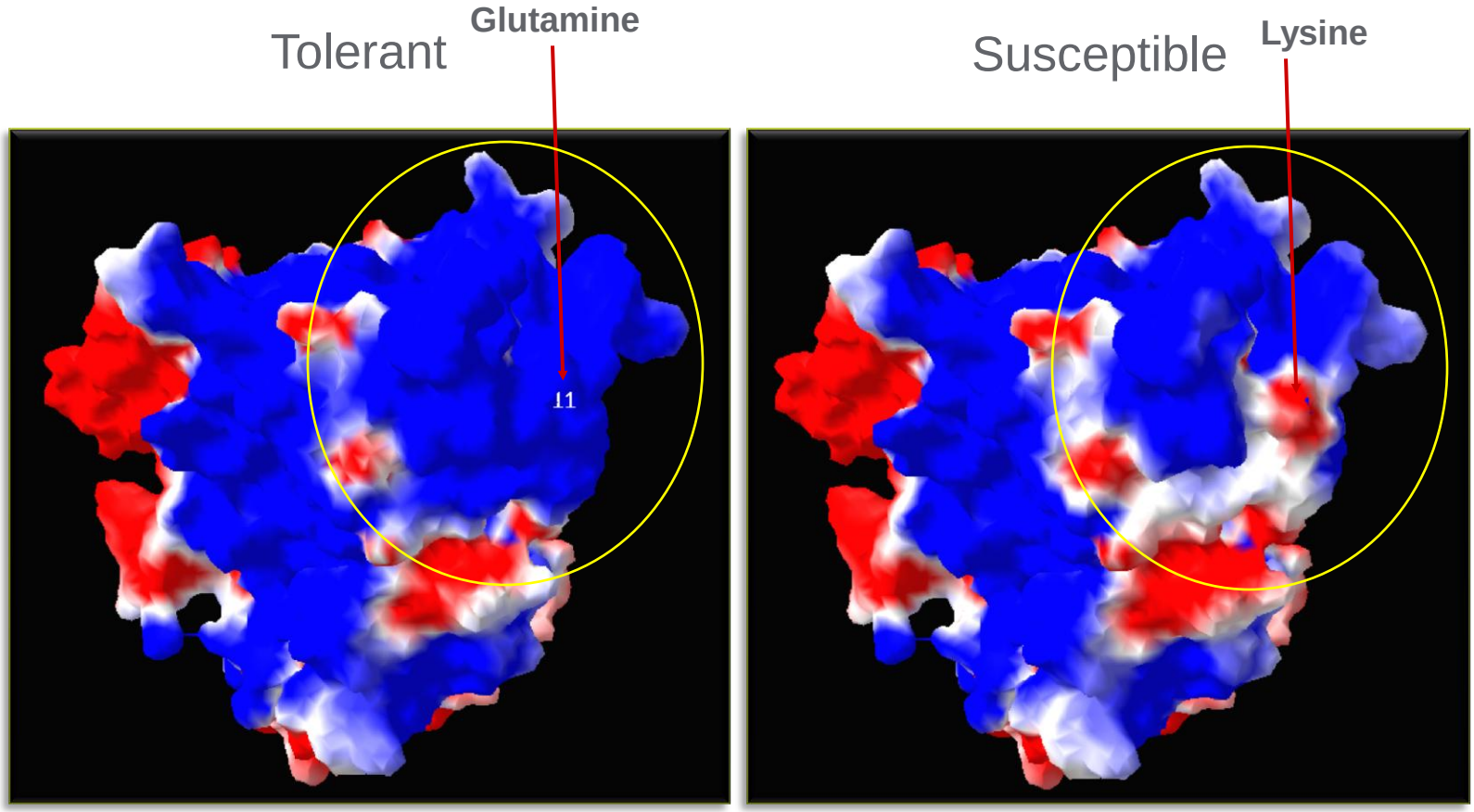
SNPs that appear to be associated with the phenotype

		1,385,702	1,385,735	1,418,379	1,418,379	1,467,691	1,467,820
Line	**Phenotype**	**Candidate gene 2**	**Candidate gene 2**	**Candidate gene 3**	**Candidate gene 3**	**Candidate Gene 5**	**Candidate gene 5**
Soy 1S	Suseptible	A	G	deletion	C	C	G
Soy 2S	Suseptible	A	G	deletion	?	C	G
Soy 3S	Suseptible	A	G	deletion	?	C	G
Soy 4S	Suseptible	H	G	deletion	C	C	G
Soy 5S	Suseptible	A	G	deletion	C	C	G
Williams82	Tolerant	G	A	no del	T	T	A
Williams	Tolerant	G	A	no del	?	T	?
Soy 1T	Tolerant	G	A	no del	T	T	A
Soy 2T	Tolerant	G	A	no del	T	T	A
Soy 3T	Tolerant	G	A	no del	T	?	?
Soy 4T	Tolerant	G	A	no del	T	T	A
Soy 5T	Tolerant	G	A	no del	T	?	?

Gene	**Polymorphisms**	**Changes in tolerant parent**
Candidate gene 5	2 SNPs	1st SNP Silent. Second causes a Lysine to Glutamine transition
Candidate gene 3	1SNP, 1 BIG Deletion	SNP in the last intron. Deletion resulted in a significant truncation.
Candidate gene 2	2 SNPs	none - SNPs in 5' untranslated region

syngenta

Effect on electrostatic surface charge



(Blue: positively charged region, Red: Negatively charged region)

In planta validation

- Out of 6 genes from the QTL region, one was determined to be a pseudogene and was not validated further.
- Transient expression of all five genes + controls was done in early 2011 in tobacco leaves
- Constructs were made using a constitutive promoter
- After infection, herbicide treated tobacco leaf samples were sent for metabolite profiling to check for herbicide breakdown products
- Plan was to choose one or two most effective gene family members for further study

Bob Dietrich
Linda Ambroso
Rex Dwyer

The watermelon RIL mapping story

Genotype by Sequencing: PI0007082

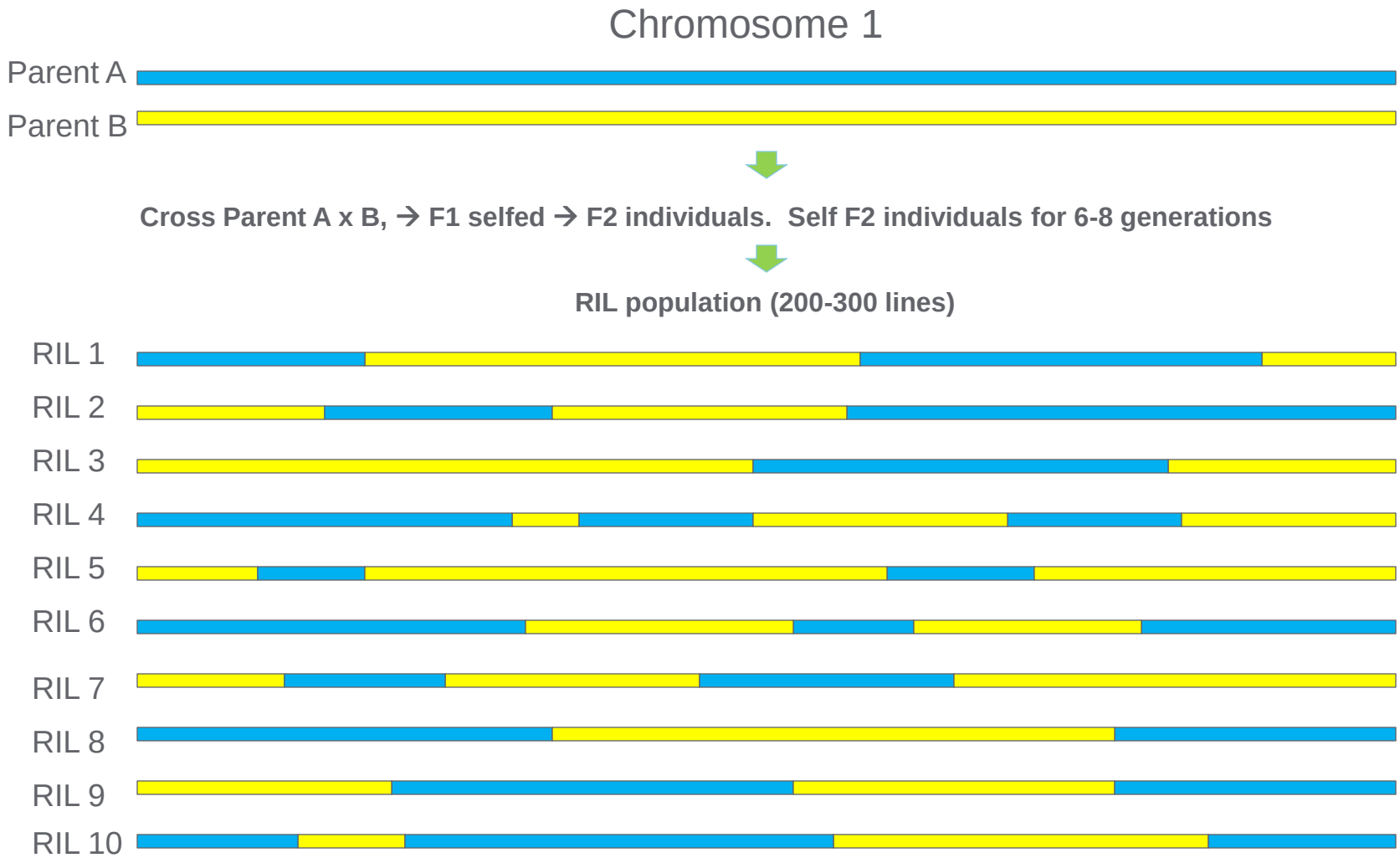
Genotyping by Next-Generation Sequencing (GBS)

- High-throughput genotyping by whole-genome resequencing
  - Huang et al., Genome Res. 2009 19: 1068-1076

- Resequenced 150 rice recombinant inbred lines (RIL)
  - Average coverage of the genome in each line: 0.02X
  - Analyzed the data using a “sliding window” approach

- Results
  - Claimed genotyping accuracy of 99.94%
  - Identified recombination breakpoints with 40kb resolution

Recombinant inbred lines



Pilot project to map a watermelon RIL population through GBS

- Watermelon recombinant inbred population
  - Parental lines sequenced for SNP discovery
    - 200,000 high confidence SNPs found
    - Used as reference for genotyping by sequencing

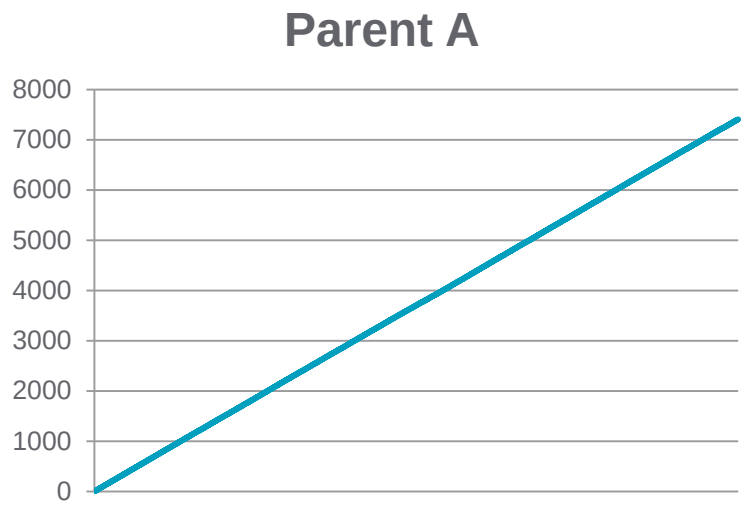
- 84 recombinant inbred lines sequenced on GAIIX

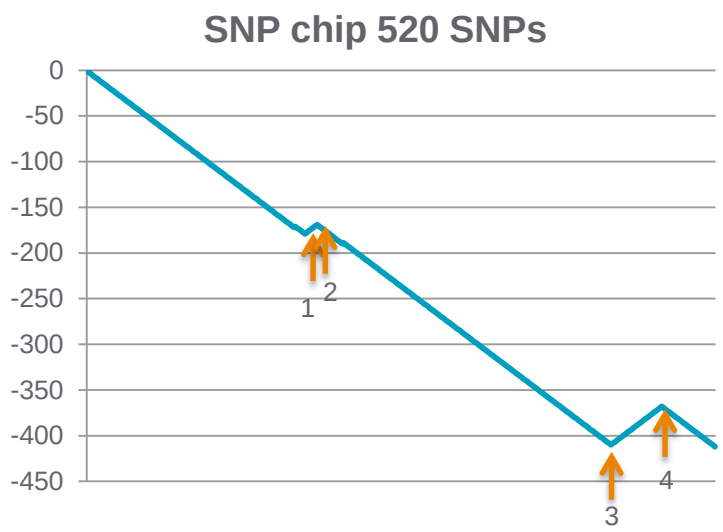
Crop	Genome size	SNP rate between Parents	Coverage
Rice	340 Mb	3.2 SNPs/kb	0.02X
Watermelon	424 Mb	0.5 SNPs/kb	0.36X

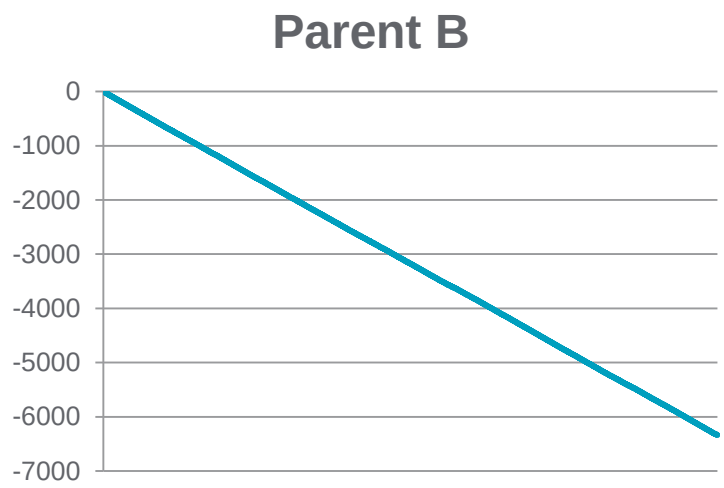
- Affymetrix watermelon SNP chip genotyping conducted to compare platforms and cross validate results

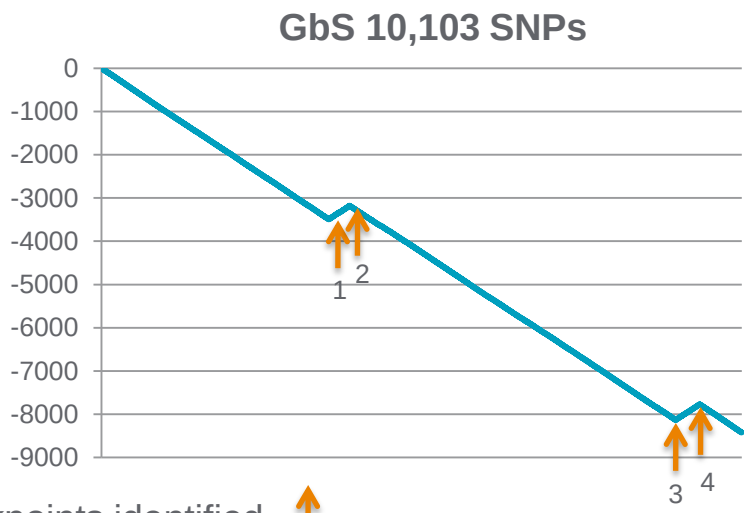
syngenta

Identifying recombination: chromosome 7 example





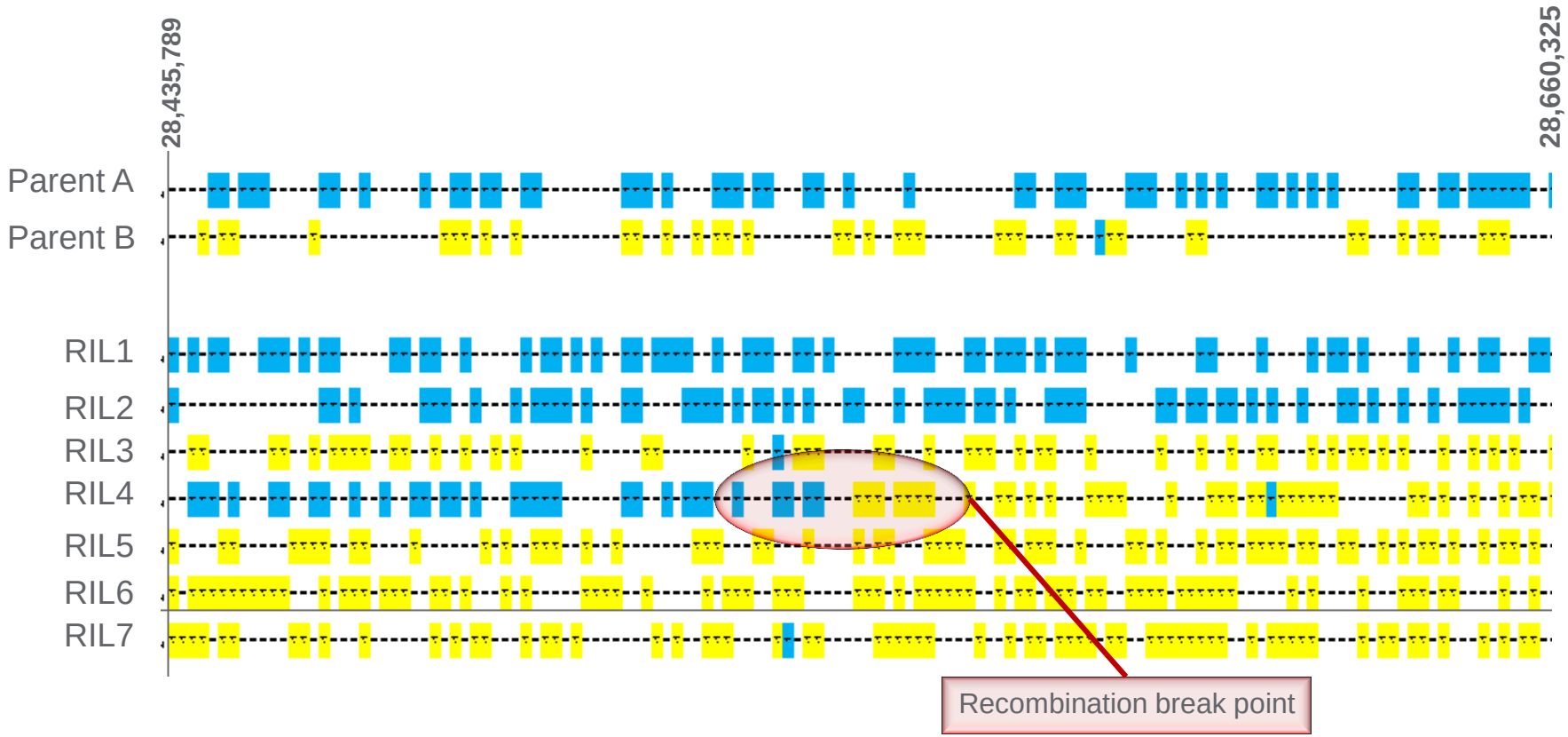




Recombination breakpoints identified

syngenta

Visualization of recombination break points



Genotyping results from a segment of chromosome 7 after variant analysis across seven RIL lines

Contrasting GBS and Chip genotyping for chromosome 7

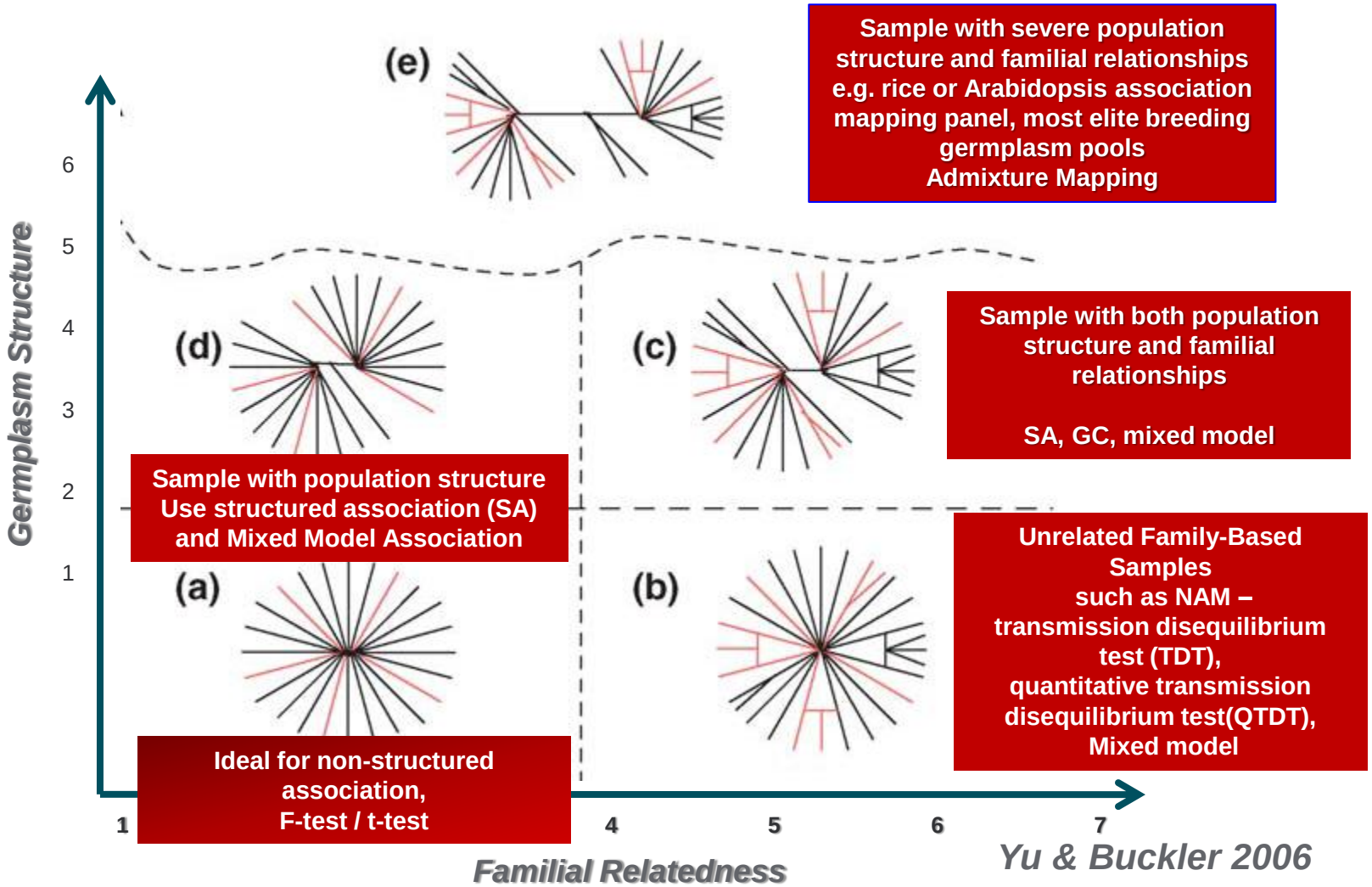
	SNP Chip	Genotype by Seq
Range, breakpoint interval size	28 kb-13 mb	62 bp -224 kb
Ave. breakpoint interval size	1,286,000 bp	55,000 bp
Ave. number of SNPs scored/line	518	8,000
SNP density	1 SNP/60 kb	1 SNP/4 kb

Illustrates the difference between RIL mapping with markers (closed system) and direct sequencing (open system)

Resolving Population Structure and Conducting GWAS

Elhan Ersoz
Joe Clarke
Nicolas Martin
Keith Allen
Christine Chaulk-Grace
Rex Dwyer
Sarah Forrester
Eric Ganko
Suresh Kadaru
Julie Leonard
Jinwei Liu
Tom Prest
Dale Skalla
Daolong Wang
Todd Warner
Chris Zinselmeier

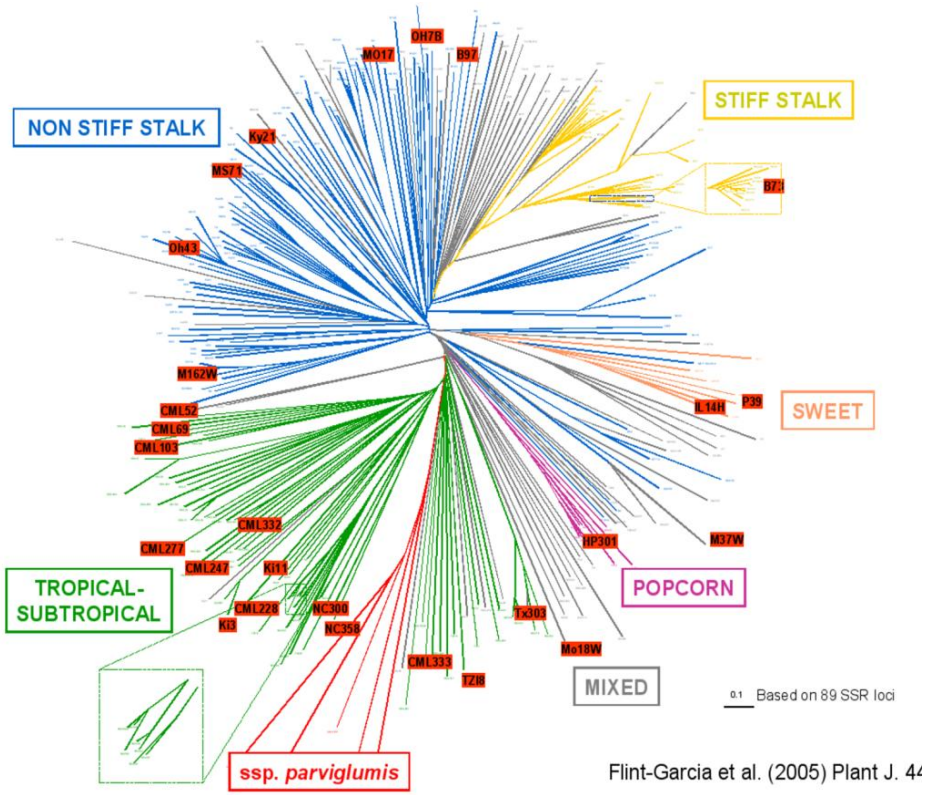
Different population structure requires different GWAS methods



Yu & Buckler 2006

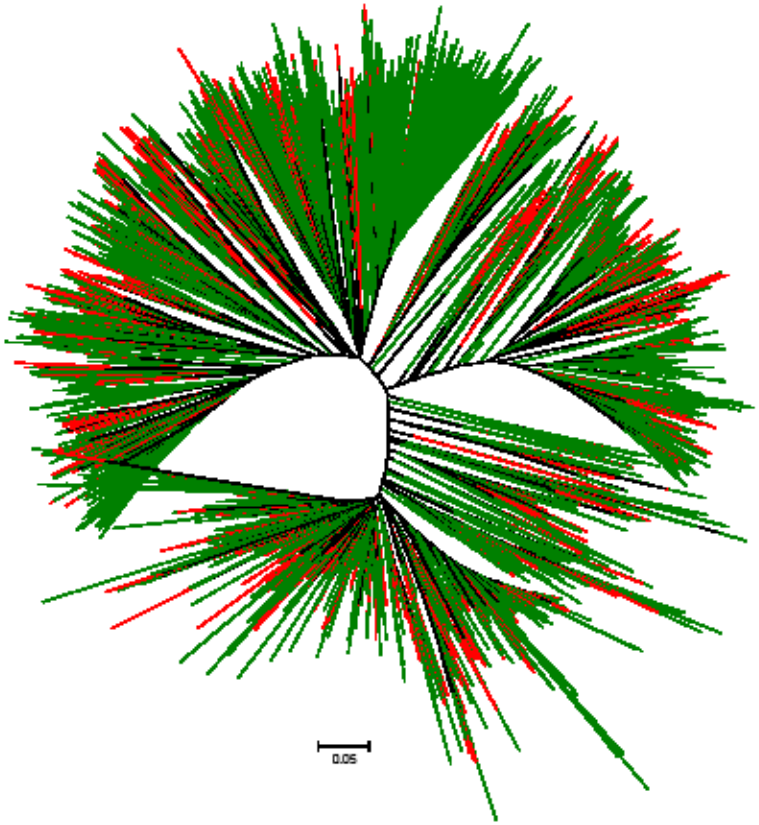
Public and Syngenta Germplasms Have Different Structures

Buckler Association Panel

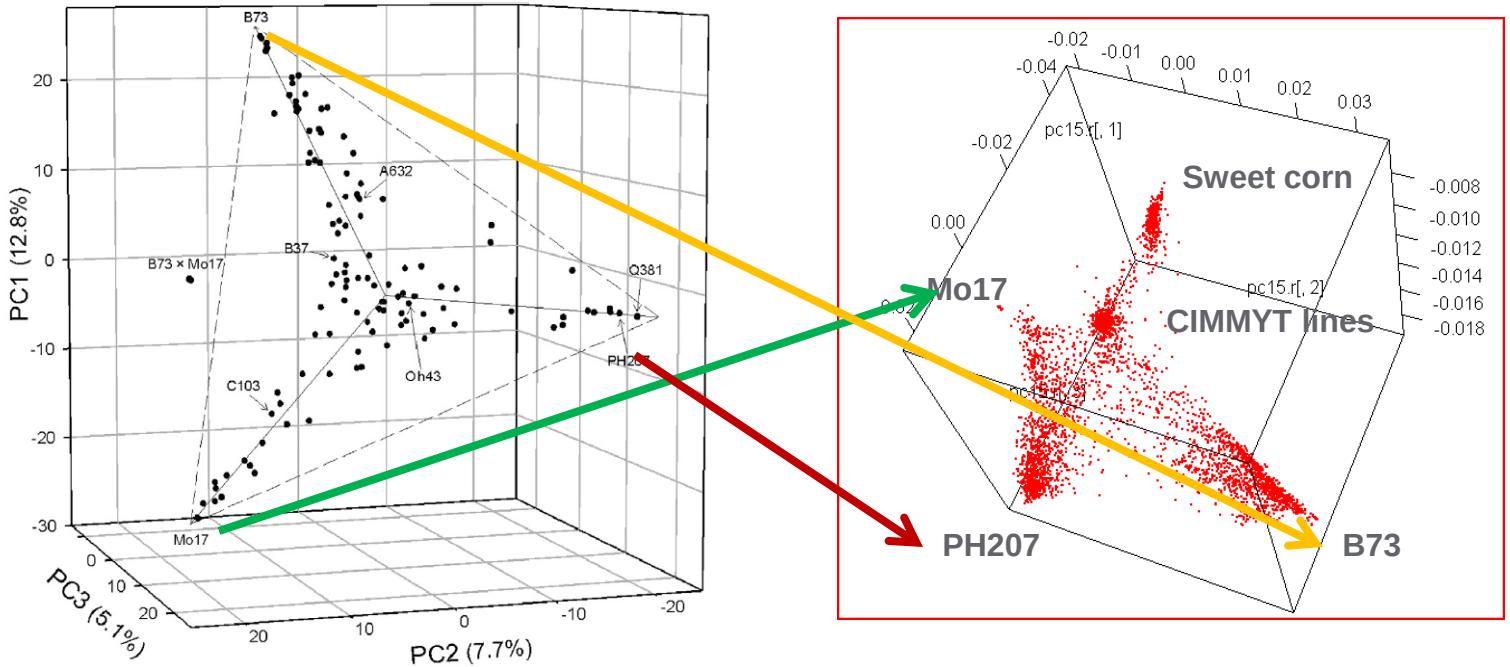


Flint-Garcia et al. (2005) Plant J. 44

Syngenta Association Panel



Syngenta Germplasm Population Structure similar to ex-PVPs'

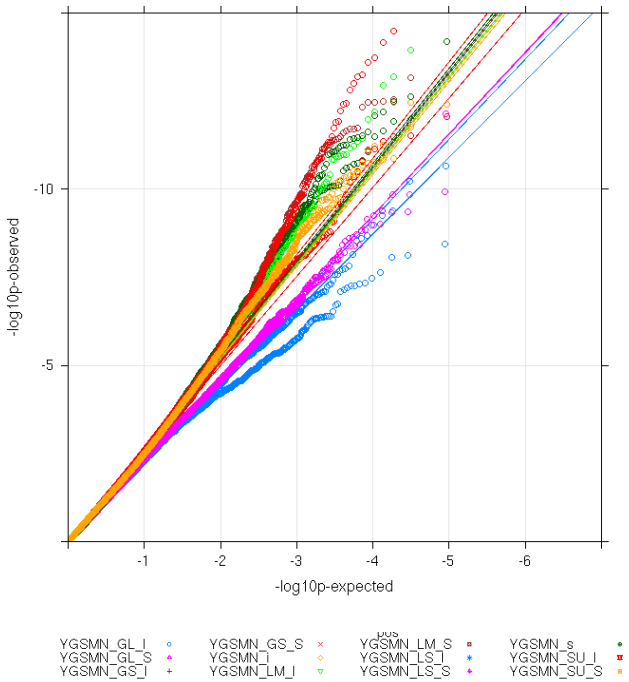


Nelson et al. 2008

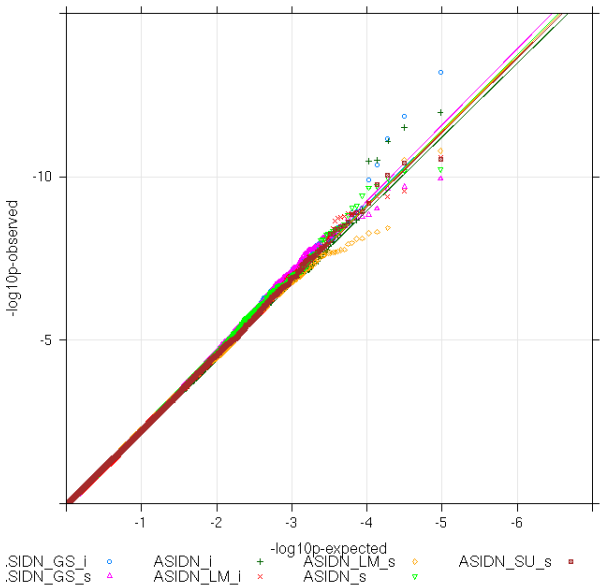
First 3 Principal Components

- 55K chip data
  - 3500 Syngenta lines
  - 500 CIMMYT elite inbreds

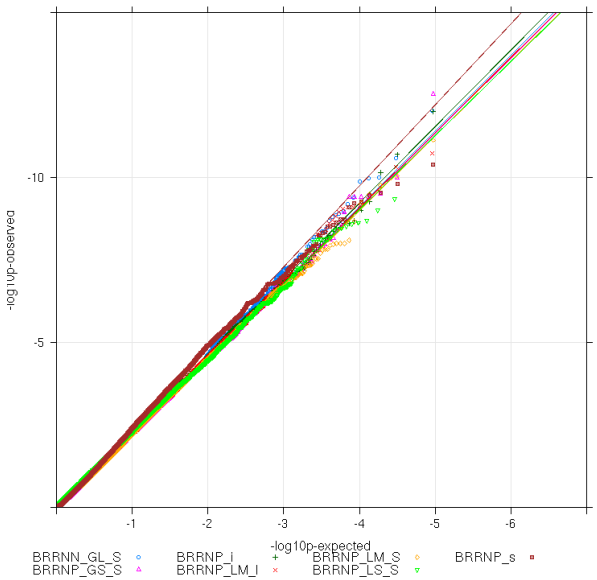
Structure control in GWAS with Unified Mixed Model Association with P3D and Compression : $Y = Q + K_1 + K_2 + E$



Yield



Barrenness



ASI

Significant new drought-tolerance loci identified

- False positives due to populations structure influence on the traits were reduced by 25-50%.

- We exploited
  - 55K SNP genotyping chip (http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0028334)
  - 1.1M proprietary Syngenta SNPs from RNAseq-based GBS
  - 1.4M HapMap SNPs imputed onto our population using lines in common with HapMap v1.
  - Total 2.1M SNPs, 30% of these mappable to ~¾ of available gene models.
  - Average 22 SNPs per gene model and approximately 1 SNP per 1000 bps.

- Our successful GWAS on Syngenta germplasm for yield component traits identified new leads for future Agrisure Artesian™ Product Development.

Bringing plant potential to life

Thank you for your attention!