Applying Genotyping by Sequencing (GBS) to Corn Genetics and Breeding

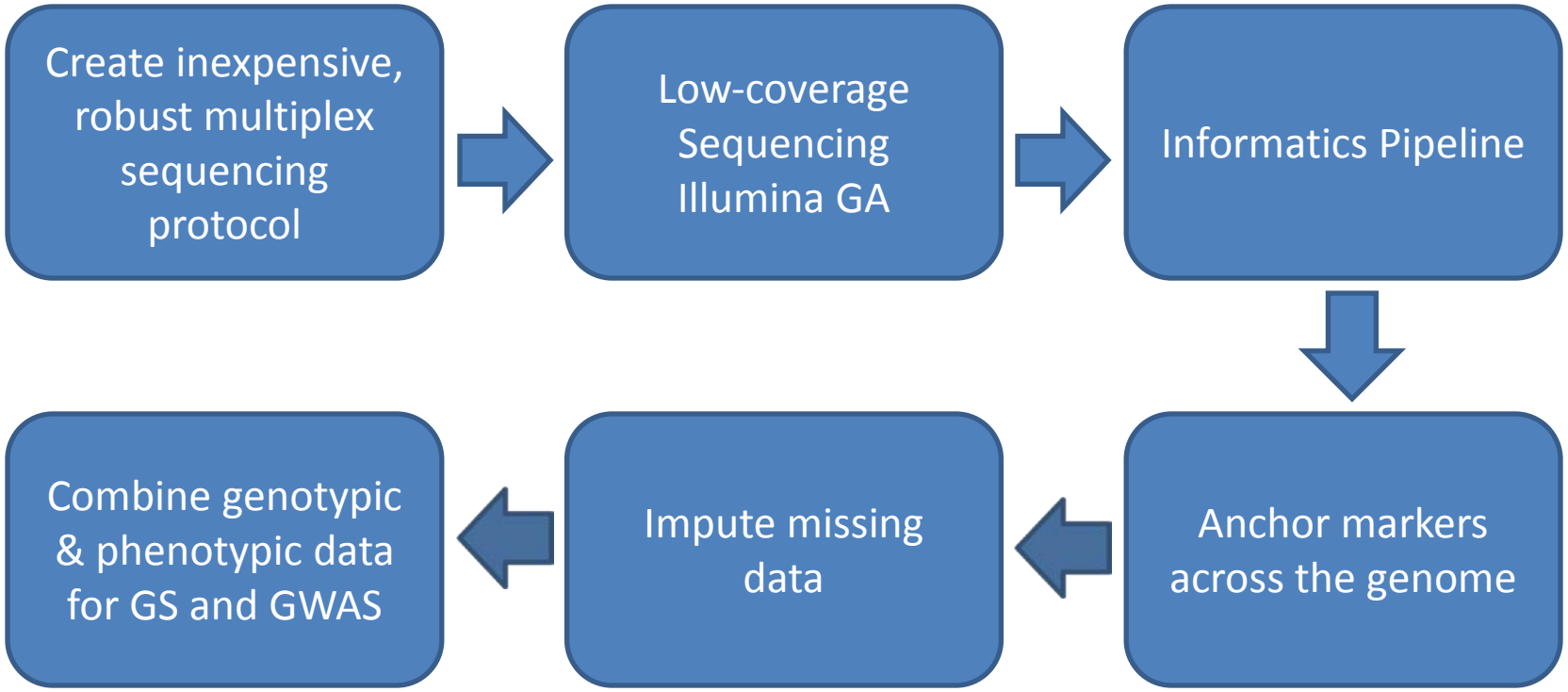
Peter Bradbury
USDA/Cornell University

Genotyping by sequencing (GBS) makes use of high through-put, short-read sequencing to provide low cost genotyping with high information content. To reduce problems in maize caused by a large genome size, reduced representation libraries are produced using a restriction enzyme that targets genomic regions while multiplexing with barcodes reduces the cost for individual samples. Using this technology DNA samples can be genotyped at over 2 million sites. Challenges include the need for a sophisticated bioinformatics pipeline, a relatively high level of missing data and under sampling of heterozygotes. Fortunately, open-source freely available software and imputation methods exist to address these challenges. Genotypes obtained using GBS can be used to examine relationships among lines, perform linkage and genome wide association studies, and perform genomic selection. While GBS has mostly been used to discover and score SNPs, it can be used to study other types of structural variants as well.

Applying Genotyping by Sequencing (GBS) to Corn Genetics and Breeding

Peter Bradbury
USDA-ARS / Cornell Univ
Ithaca, NY

Goal: to create a public genotyping – informatics platform based on next-generation sequencing



Background

Genotyping by sequencing (GBS) in any large genome species requires reduction of genome complexity.

I. Target enrichment

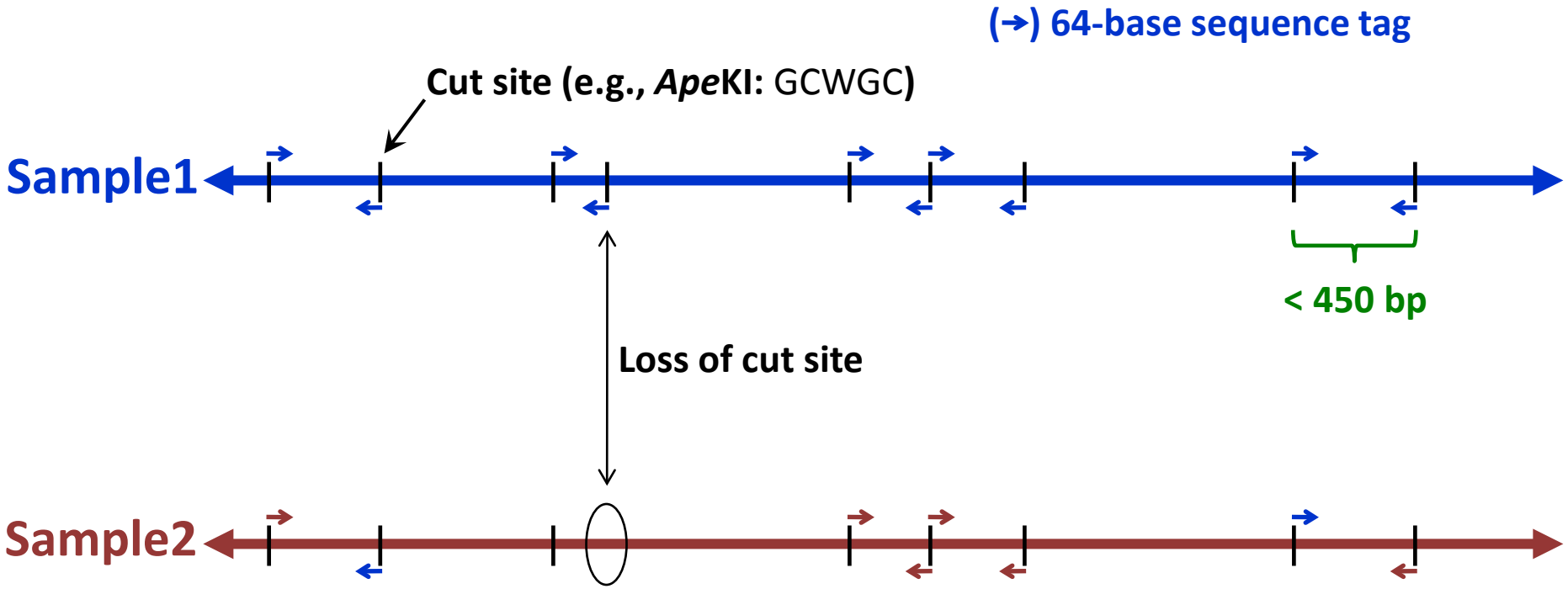
- Long range PCR of specific genes or genomic subsets
- Molecular inversion probes
- Sequence capture approaches hybridization-based (microarrays)

II. Restriction Enzymes (REs)

Technically less challenging

- Methylation sensitive REs filter out repetitive genomic fraction

What is genotyping-by-sequencing (GBS)?



- Reduced representation approach inspired by Altshuler *et al.* (2000)
- Focuses NextGen sequencing power to ends of restriction fragments
- Scores both biallelic markers and presence/absence markers

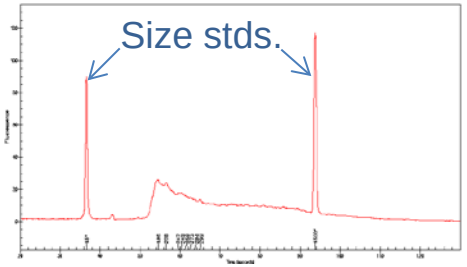
The GBS protocol is simple and robust

Elshire *et al.* 2011. PLoS ONE

Rob Elshire

Barcode adapter

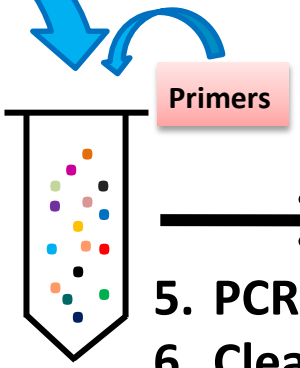
Common adapter



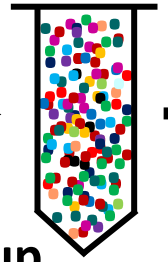
7. Evaluate fragment sizes

1. Plate DNA & adapter pair

4. Pool DNAs



5. PCR
6. Cleanup

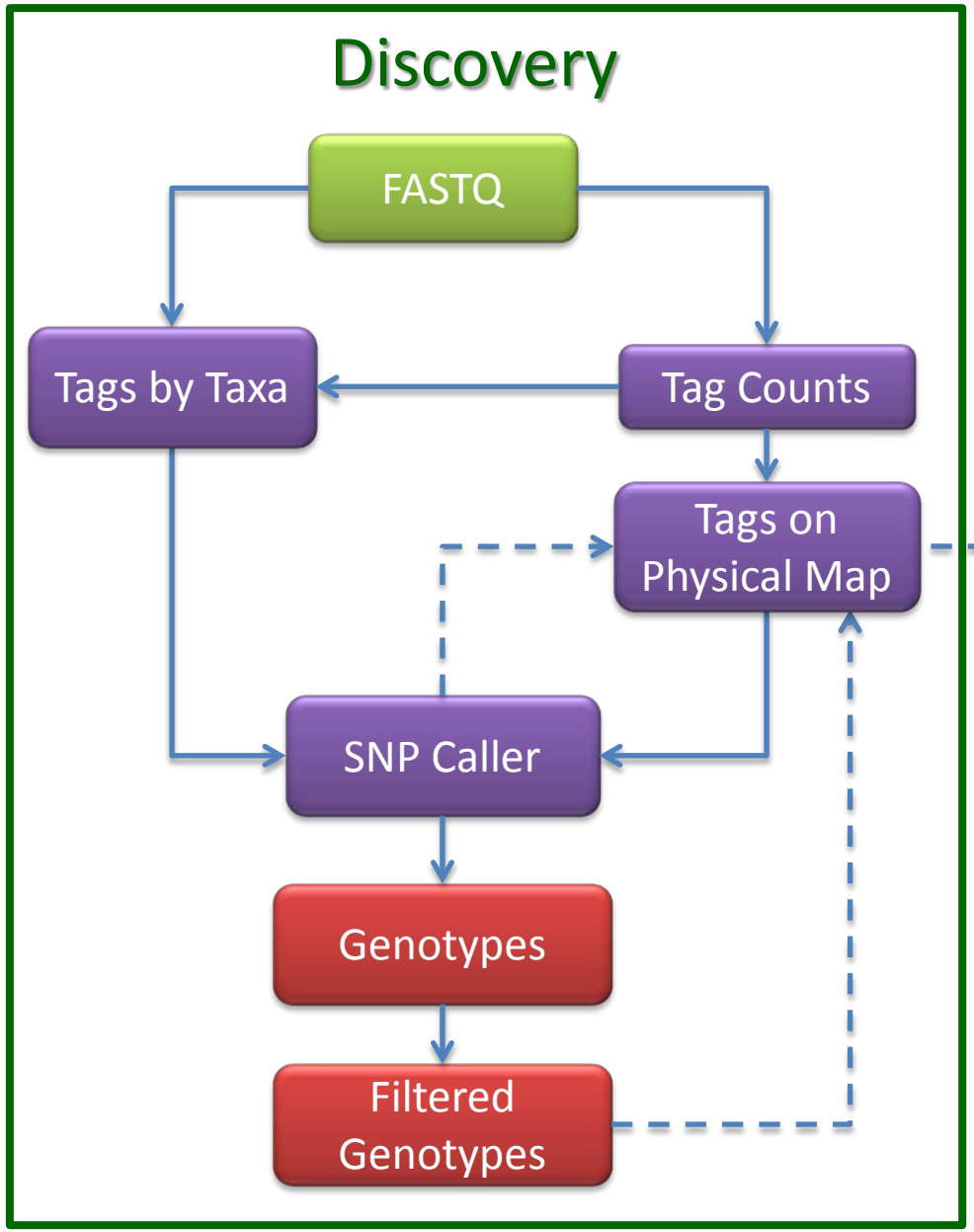


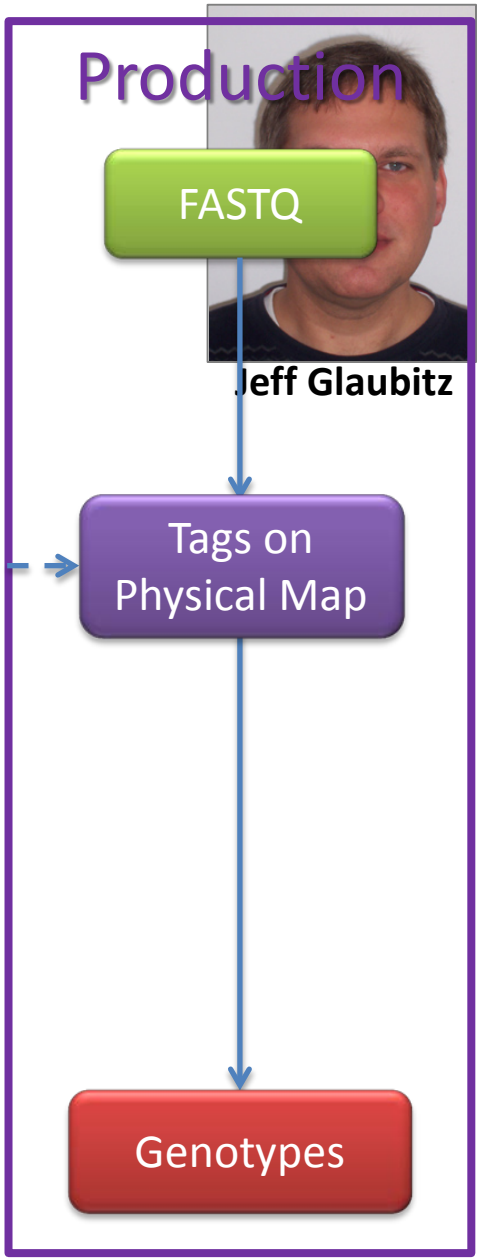
2. Digest DNA with RE (e.g. *Ape*KI)
3. Ligate adapters

8. Run on single lane of Illumina flowcell

Sequence Reads in FASTQ format

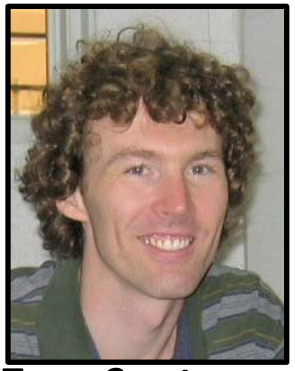
Reference-based GBS bioinformatics pipeline





Jeff Glaubitz


Ed Buckler


Terry Casstevens


James Harriman

TASSEL

Tassel Version 4.0 (*Larger Data Sets / Faster*) (Build: February 14, 2013 Requires: Java 1.6)

Launch TASSEL 4.0
Launch TASSEL 4.0 (950Mb Heap Size) (Use if Error Creating Java Virtual Machine)
TASSEL 4.0 Standalone (Using Git - *Recommended!*) (**Do Not Use for GBS Pipeline - Under Development**.)

Tassel Version 3.0 (Build: January 10, 2013 Requires: Java 1.6)

Launch TASSEL 3.0
Launch TASSEL 3.0 (950Mb Heap Size)
Launch TASSEL 3.0 (512Mb Heap Size)
TASSEL 3.0 Standalone (Using Git - *Recommended!*) (**For GBS Pipeline**)

Tassel Documentation

Tassel 3.0 User's Guide (Updated: December 22, 2011)
Tassel Tutorial Data
Visit Tassel User Group
Tassel FAQs
Tassel 4.0 Change History

Tassel Pipeline Documentation

Tassel 3.0 / 4.0 Pipeline Document (Updated December 13, 2012)
Tassel 3.0 GBS Pipeline Document (Updated September 13, 2012)
Tassel 3.0 UNEAK Pipeline Document (Updated May 11, 2012)

GBS Pipeline functions only available via the command line

Open source, hosted at sourceforge.net

Pros & Cons of GBS

- Pros
  - obtain large amount of data very quickly
  - inexpensive: www.igd.cornell.edu/index.cfm/page/projects/GBS/GBSpricing.htm
  - relatively free of ascertainment bias
  - can assay regions absent from reference genome
- Cons
  - large amount of missing data (~40-80% with *Ape*KI)
  - difficult to call heterozygotes in highly heterozygous, unrelated individuals
  - technically missing confounded with biologically missing

Zea samples genotyped with GBS

- 30,000 *Zea* samples (*Ape*KI)
  - 40% RILs from bi-parental families
  - 35% unrelated inbred lines
  - 25% outcrossed (heterozygous) individuals
    - maize landraces & teosinte

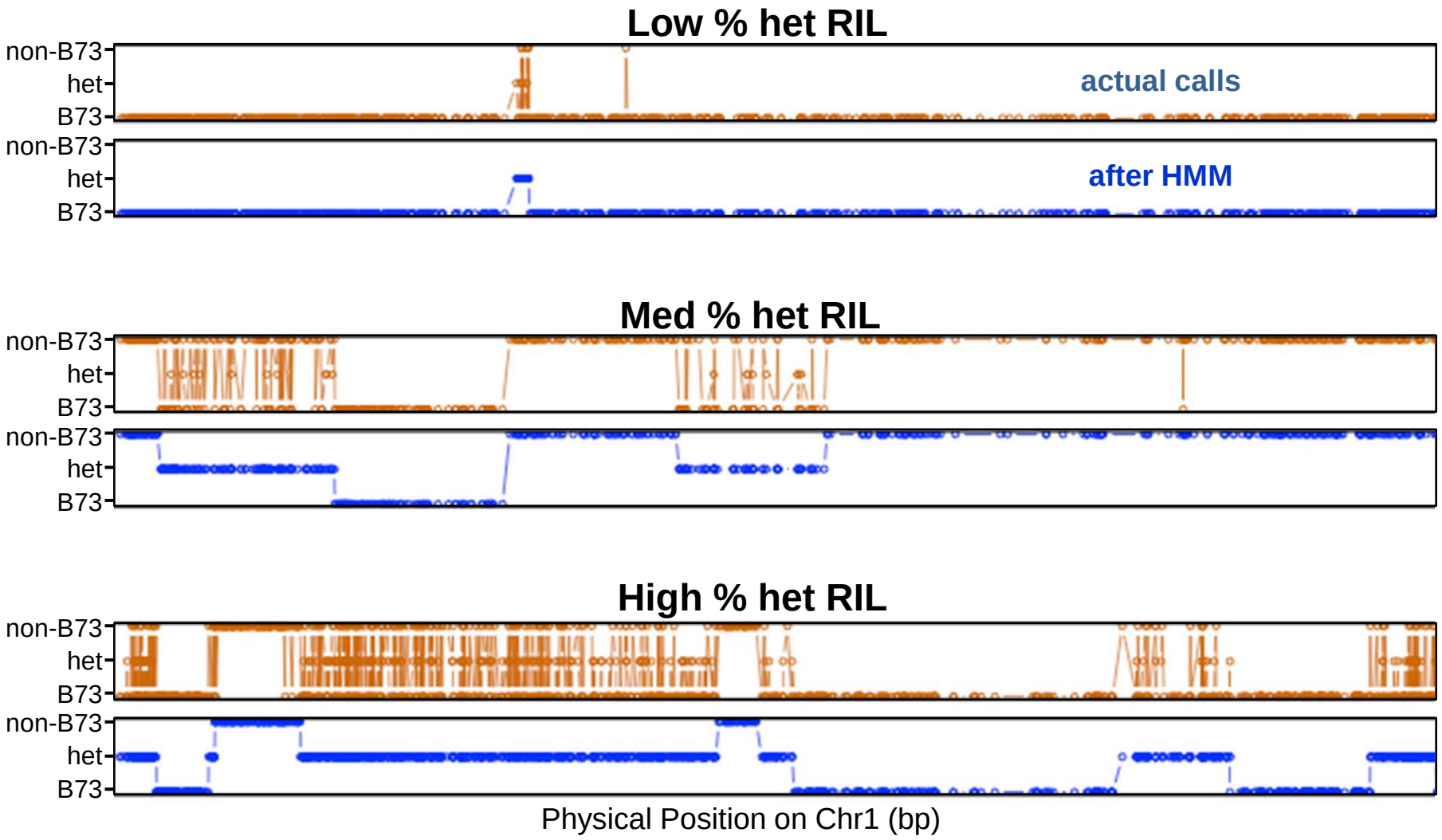
- **2.2 million SNPs (after filtering error-prone)**
- **73% missing data (27% call rate)**

GBS error rates *vs*. Maize 50K SNP Chip

- 7,254 SNPs in common
- 279 maize inbreds in common ("Maize282" panel)

Comparison to 50K SNPs	Mean Error Rate (per SNP)	Median Error Rate (per SNP)
Filtered GBS genotypes		
All genotypic comparisons:	**1.18%**	**0.93%**
Homozygotes only:	**0.58%**	**0.42%**

HMM for calling hets/correcting errors in biparental RIL populations



Simulation Results

S4 RILs, bulk DNA sample of 4 progeny

type of error	hom -> hom	hom -> het	het -> hom
before imputation	.0020	.0020	.8009
after imputation	.00002	.00049	.0146

S1 families, bulk DNA sample of 4 progeny

type of error	hom -> hom	hom -> het	het -> hom
before imputation	.0020	.0020	.8017
after imputation	.0000051	.0036	.0106

Use of imputed markers for joint linkage analysis of NAM

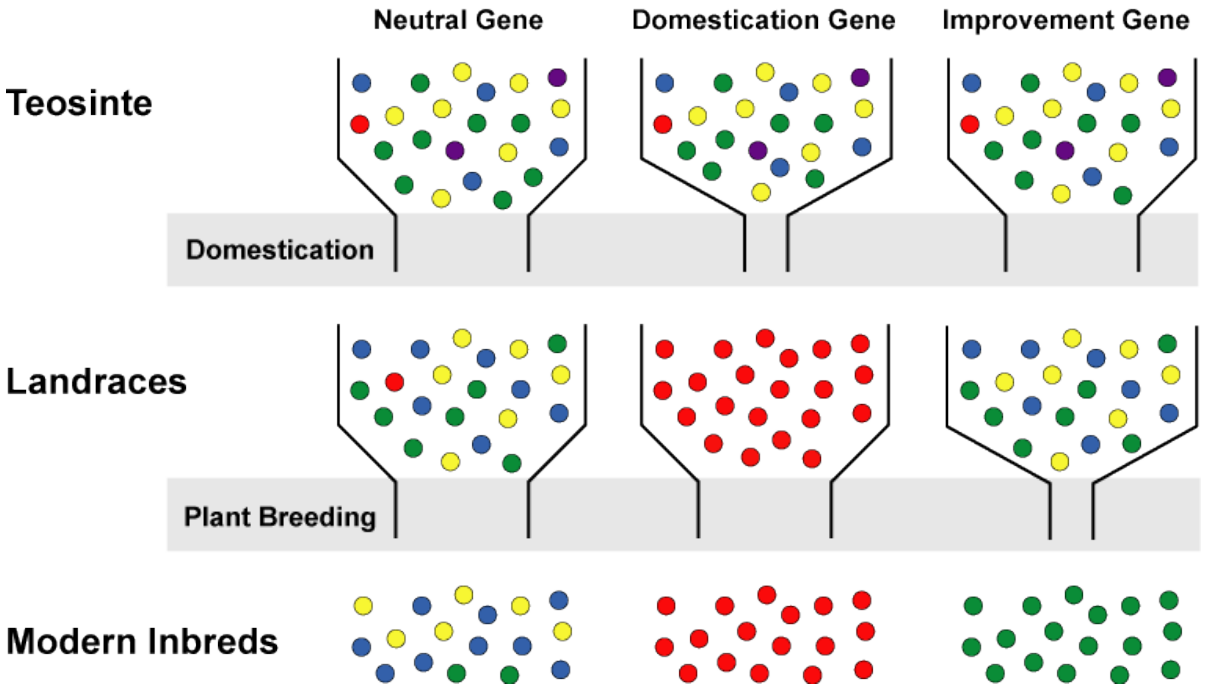
- Impute markers every 0.2 cm using
- 1106 SNPs Illumina GoldenGate Array
- GBS build 1, 600K SNPs
- GBS build 2, 2.2 million SNPs

Average support interval from regression model

Data	**20 terms**	**30 terms**
Array	5.03 cM	4.95 cM
Build 1	2.97 cM	3.49 cM
Build 2	2.95 cM	2.9 cM

Imputation in unrelated inbred lines

- Aided by limited number of ancestral haplotypes in modern maize



Imputation in unrelated inbred lines


Ed Buckler

- Aided by limited number of ancestral haplotypes in modern maize
- **Performed high coverage GBS on 195 diverse inbreds**
- **Imputation via a nearest neighbor approach**
  - **based on bit arithmetic, so extremely fast**
  - **sliding windows of 4096, 2048, 1024 & 512 sites**
    - disagreements between window sizes set to missing
  - **nearest neighbors with identity-by-state >= 95%**
  - **minimum of 2 nearest neighbors**
  - **missing SNP imputed to consensus of nearest neighbors**
- **Median imputation error rate (by masking): 0.4%**

GBS error rates *vs*. Maize 50K SNP Chip

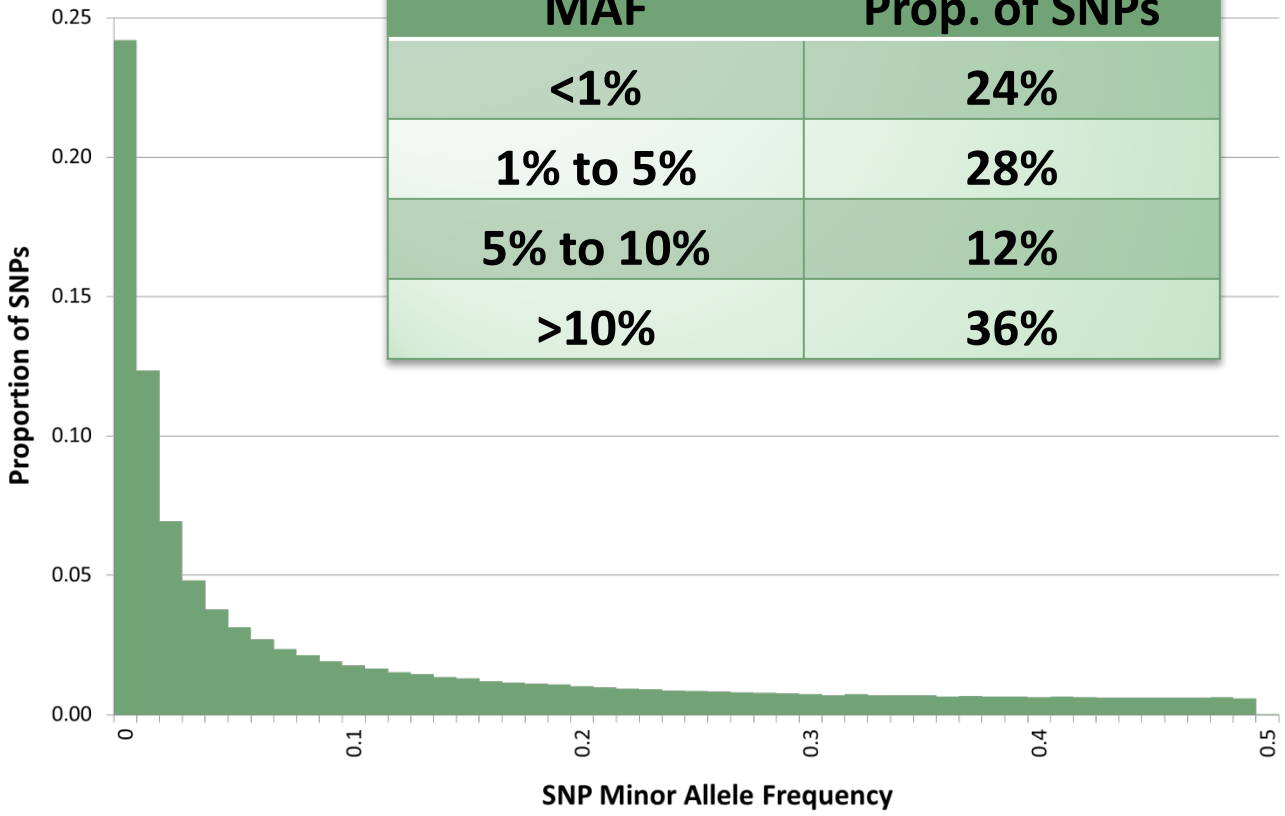
- 7,254 SNPs in common
- 279 maize inbreds in common ("Maize282" panel)

Comparison to 50K SNPs	Mean Error Rate (per SNP)	Median Error Rate (per SNP)
Filtered GBS genotypes		
All genotypic comparisons:	**1.18%**	**0.93%**
Homozygotes only:	**0.58%**	**0.42%**
Imputed GBS genotypes:		
All genotypic comparisons:	**6.94%**	**4.07%**
Homozygotes only:	**4.62%**	**1.83%**

Most GBS SNPs in maize inbreds are rare

Cinta Romay

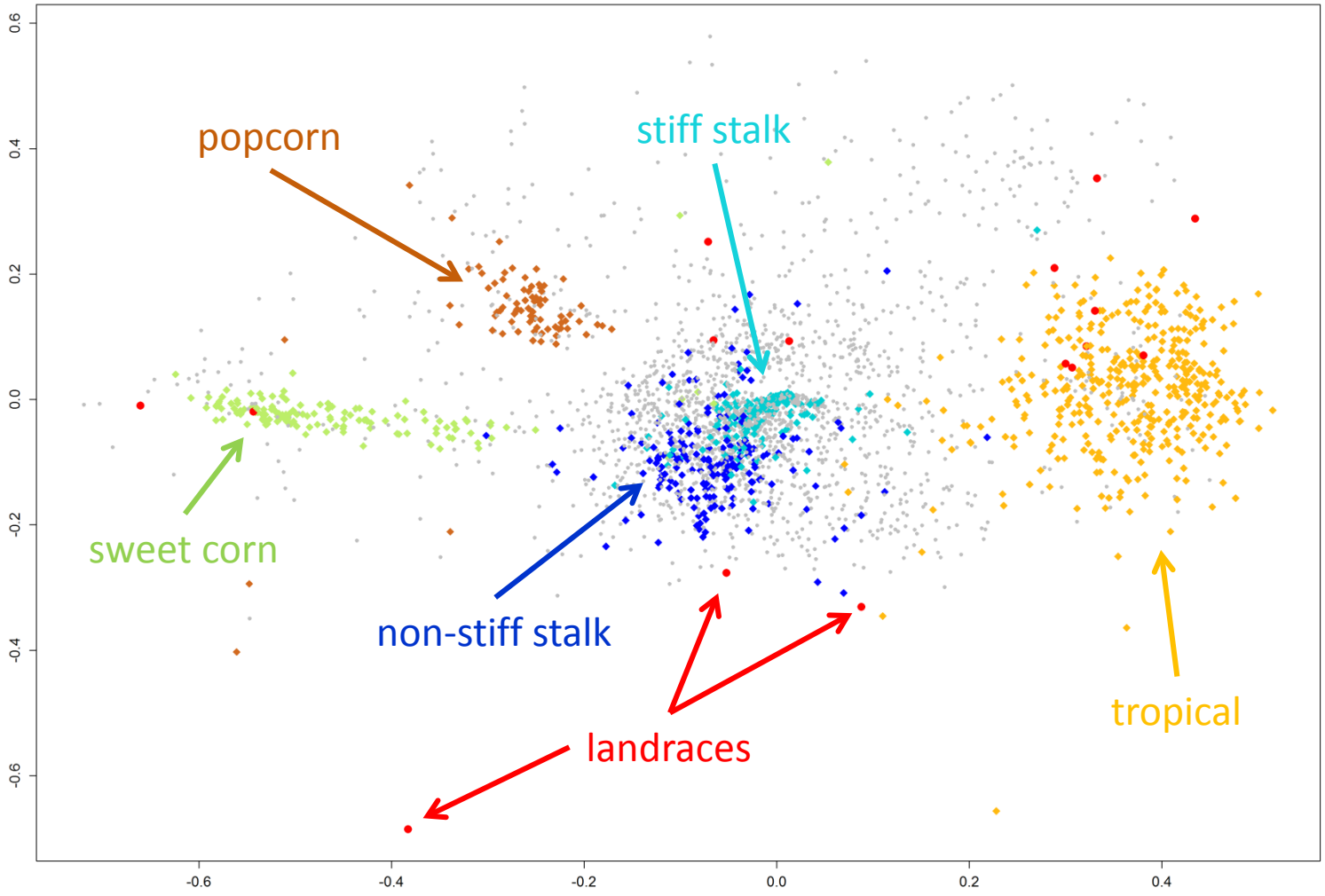
MAF	Prop. of SNPs
<1%	24%
1% to 5%	28%
5% to 10%	12%
>10%	36%



Distribution of >570K SNPs across 2,709 lines

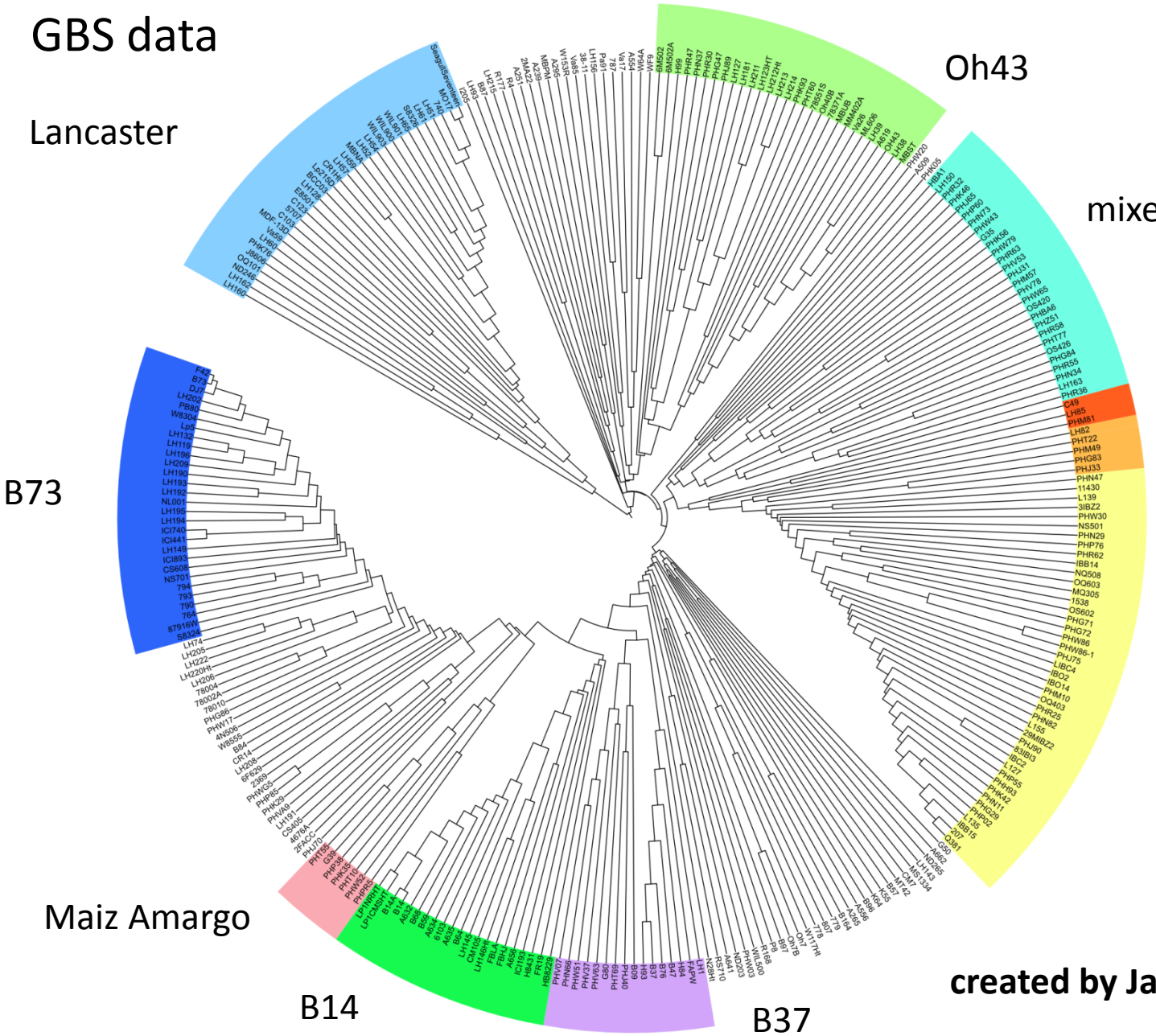
GBS explains genetic relationships among maize inbreds

Multidimensional Scaling (MDS) analysis of USDA maize inbreds using 660K GBS markers



Cinta Romay

Phylogenetic Tree of public and ex-PVP inbreds generated using GBS data



created by Jason Morales

Less ascertainment bias than array-based SNPs?

Ram Sharma

Non stiff stalk,nss (106)

Stiff stalk, ss (28)

Tropical/sub tropical (66)

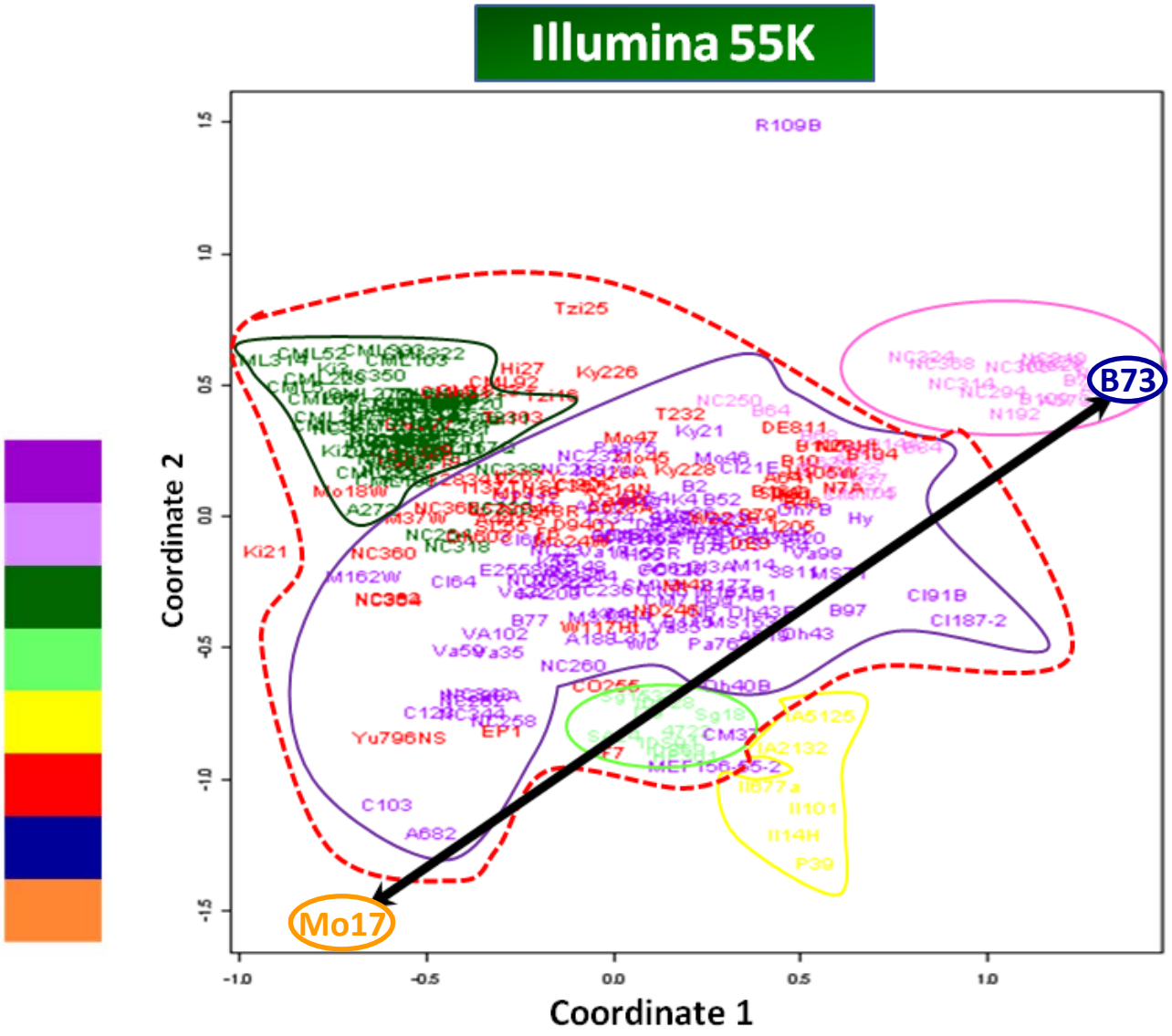
Popcorn (9)

Sweet corn (6)

Unclassified(67)

B73 (ss)

Mo17 (nss)



Less ascertainment bias than array-based SNPs?

Ram Sharma

Non stiff stalk,nss (106)

Stiff stalk, ss (28)

Tropical/sub tropical (66)

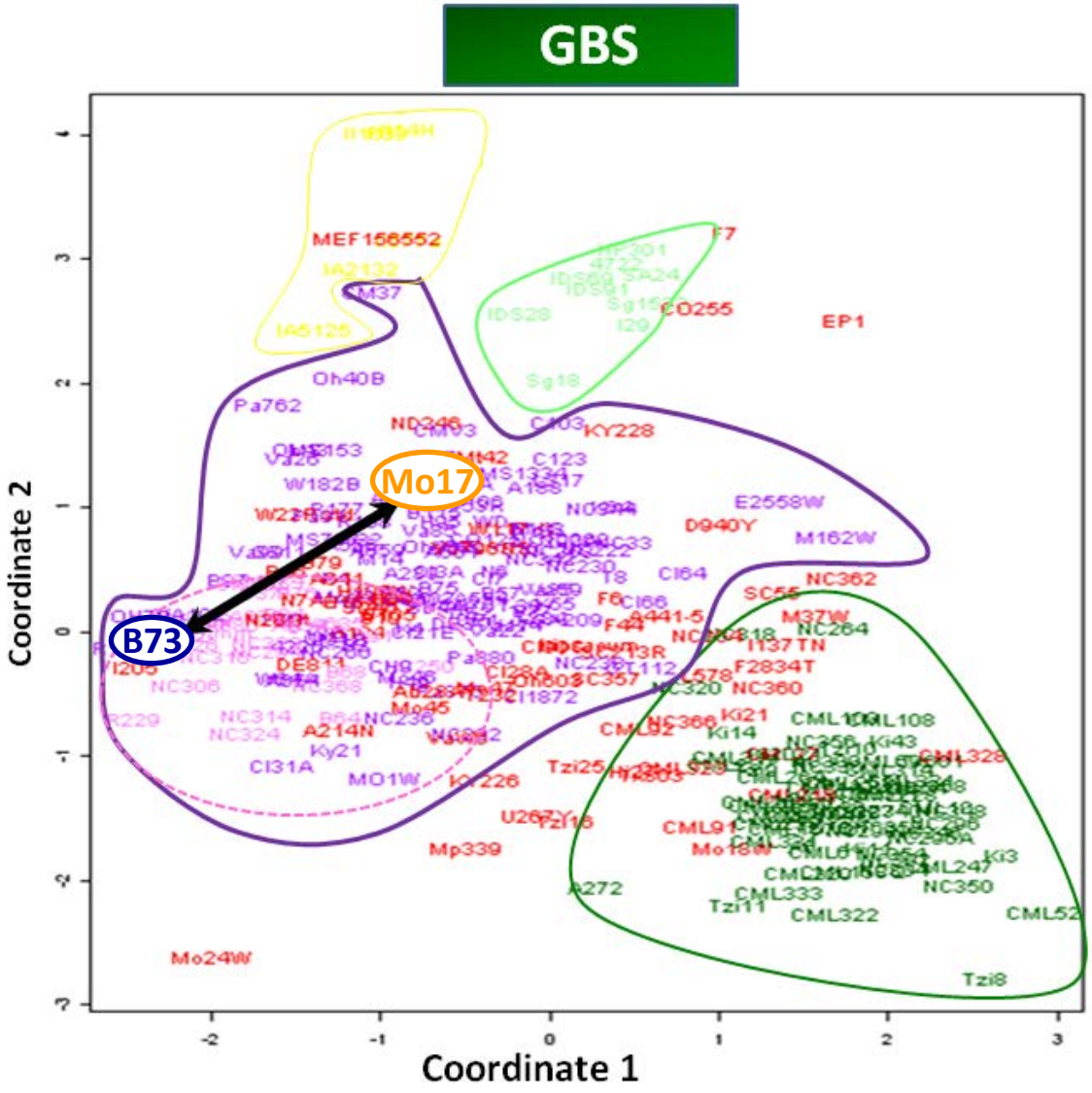
Popcorn (9)

Sweet corn (6)

Unclassified(67)

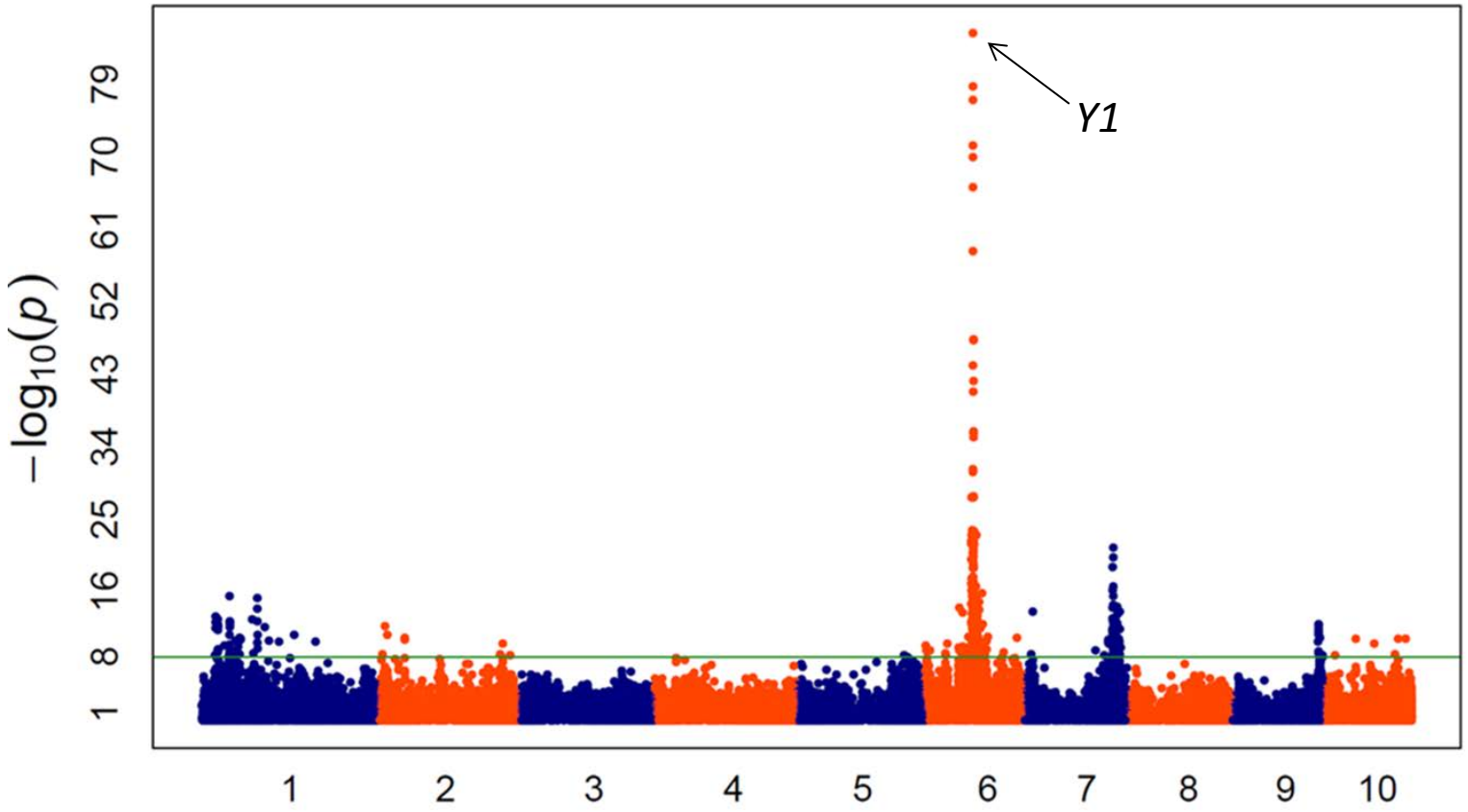
B73 (ss)

Mo17 (nss)



GWAS directly hits known Mendelian traits

GWAS of white vs. yellow kernels in 1,595 USDA Ames inbreds

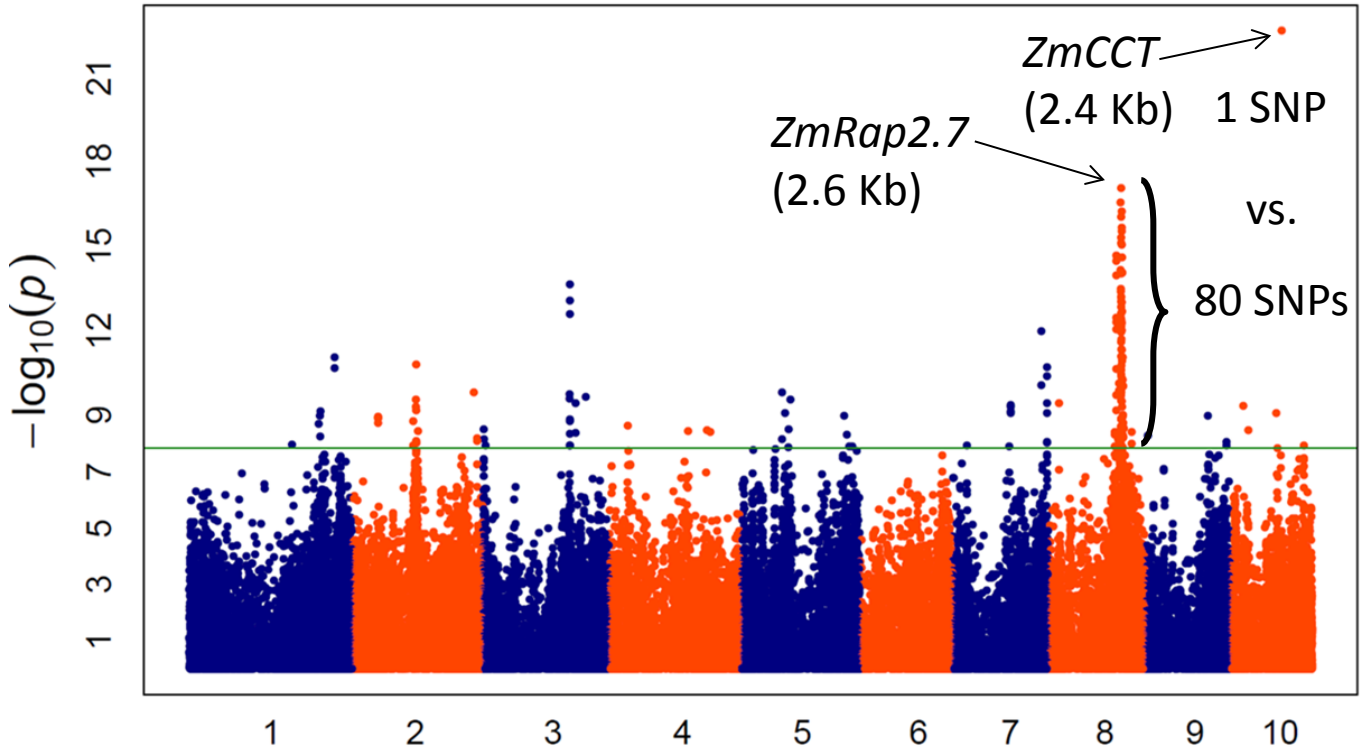


Cinta Romay

The best hit for kernel color lies within *Y1*

GWAS of a more complex trait directly hits known flowering time genes

GWAS of growing degree days to silking in 2,279 inbreds

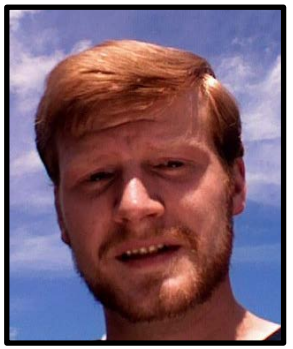


Genome Association Prediction Integrated Tool

Alex Lipka

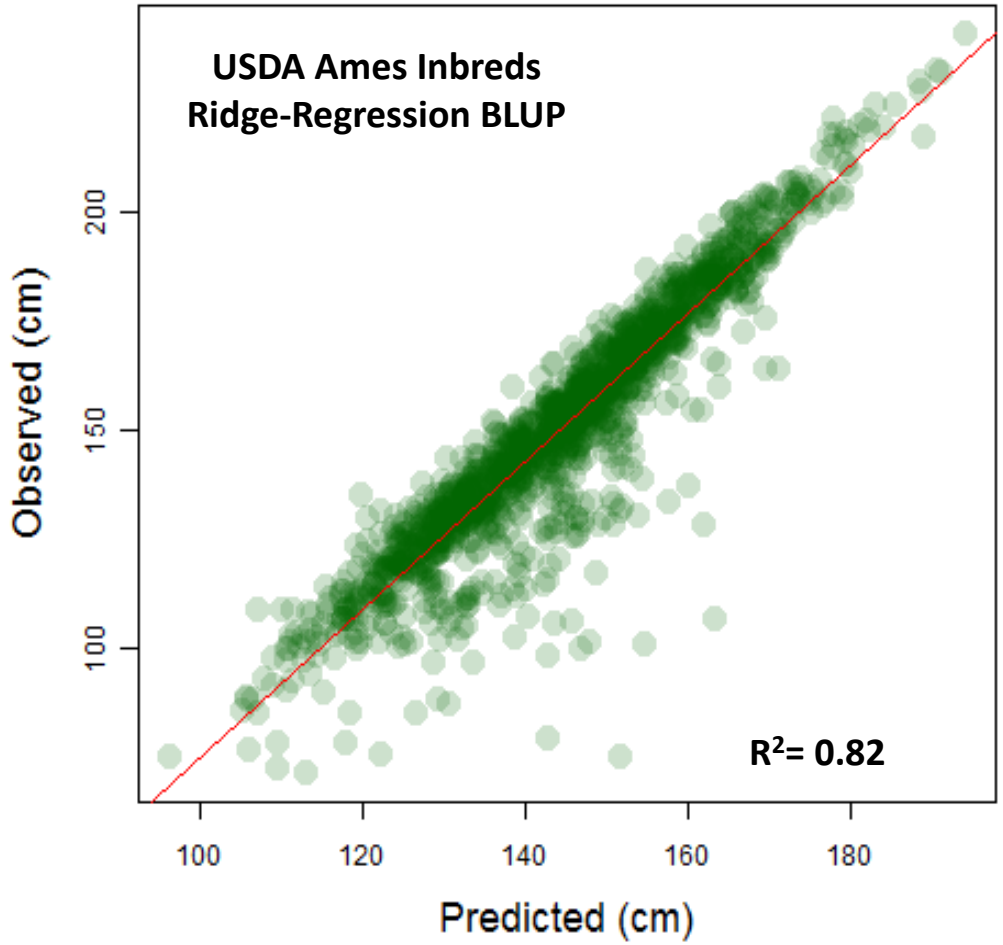
Zhiwu Zhang

Accurate genomic prediction of height based on GBS data



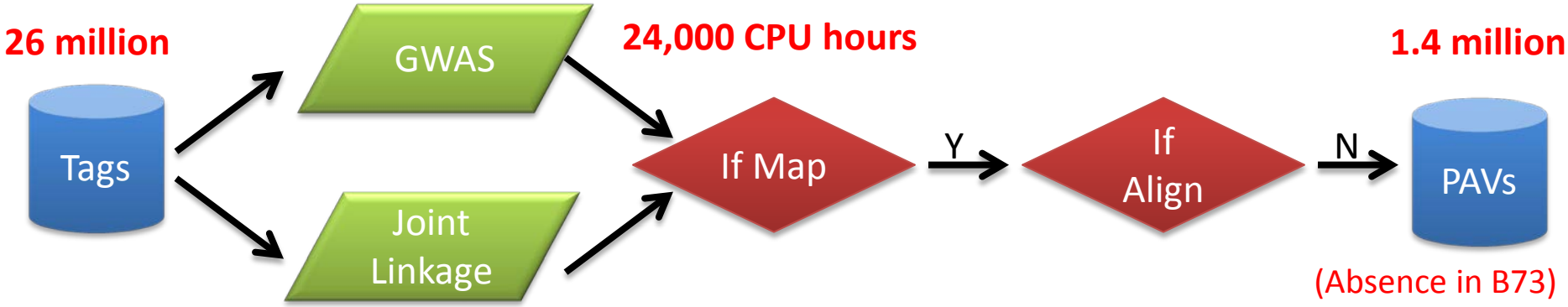
Jason Peiffer

2800 diverse breeding lines

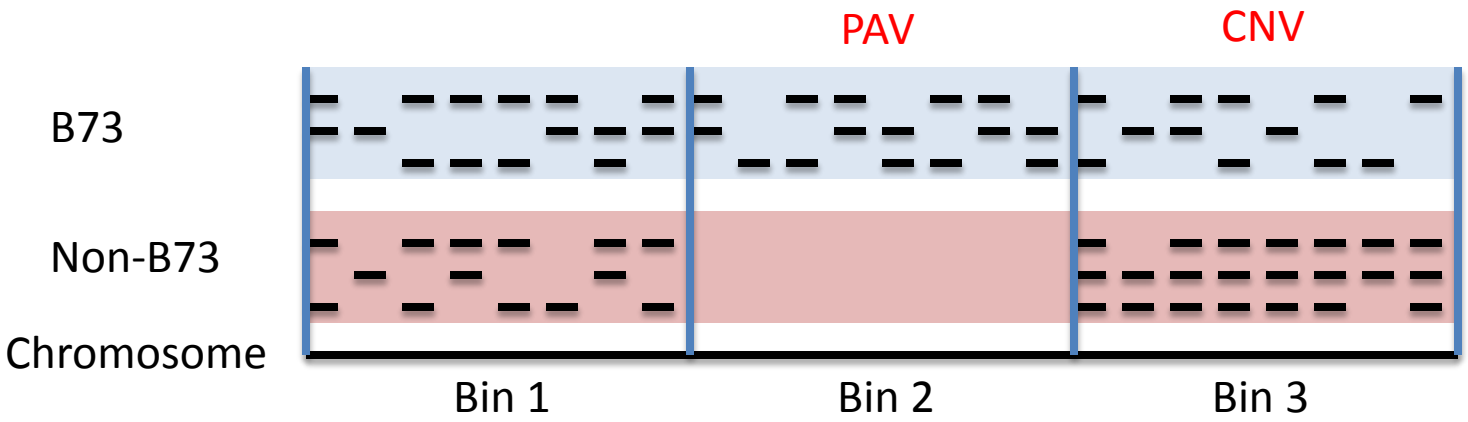


Characterize structural variations in 19,101 maize inbred lines

Mapping tags that do not align to reference



Read depth compared to B73 reference



Fei Lu

Acknowledgements

GBS Protocol & Lab Work:
Rob Elshire (Cornell)
Sharon Mitchell (Cornell)
Charlotte Acharya (Cornell)
Wenyan Zhu (Cornell)
Lisa Blanchard (Cornell)

GBS Bioinformatics:
Ed Buckler (USDA/Cornell)
Peter Bradbury (USDA/Cornell)
Terry Casstevens (Cornell)
Rob Elshire (Cornell)
Fei Lu (Cornell)
Yang Zhang (Cornell)
Dallas Kroon (USDA/Cornell)
Robert Bukowski (Cornell)
Jaroslaw Pillardy (Cornell)
Qi Sun (Cornell)

Statistical Genetics:
Zhiwu Zhang (Cornell)
Alex Lipka (USDA/Cornell)
Peter Bradbury (USDA/Cornell)
Ed Buckler (USDA/Cornell)

Analyses:
Cinta Romay (Cornell)
Jason Peiffer (Cornell)
Ram Sharma (IHBT)
Jason Morales (Purdue)

Maize NAM population & phenotypes:
Ed Buckler (USDA/Cornell)
Nick Lepak (USDA/Cornell)
NSF Maize Diversity Project (Panzea)