Usage Cases of GBS

Jeff Glaubitz (jcg233@cornell.edu)

Senior Research Associate, Buckler Lab, Cornell University

Panzea Project Manager

Cornell CBSU Workshop

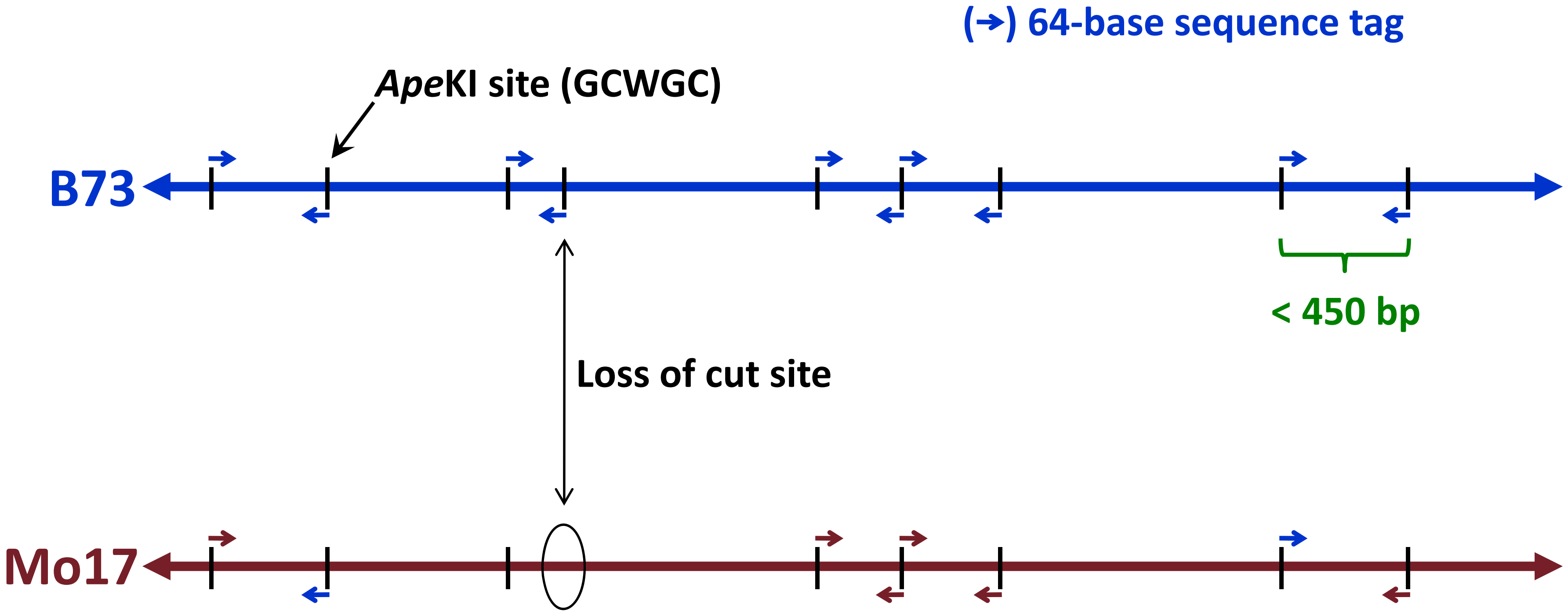
Oct 31-Nov 1, 2011

Some potential applications of GBS Data

- Marker discovery
- Phylogeny/Kinship
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genomic selection
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Improving reference genome assembly

Marker Discovery

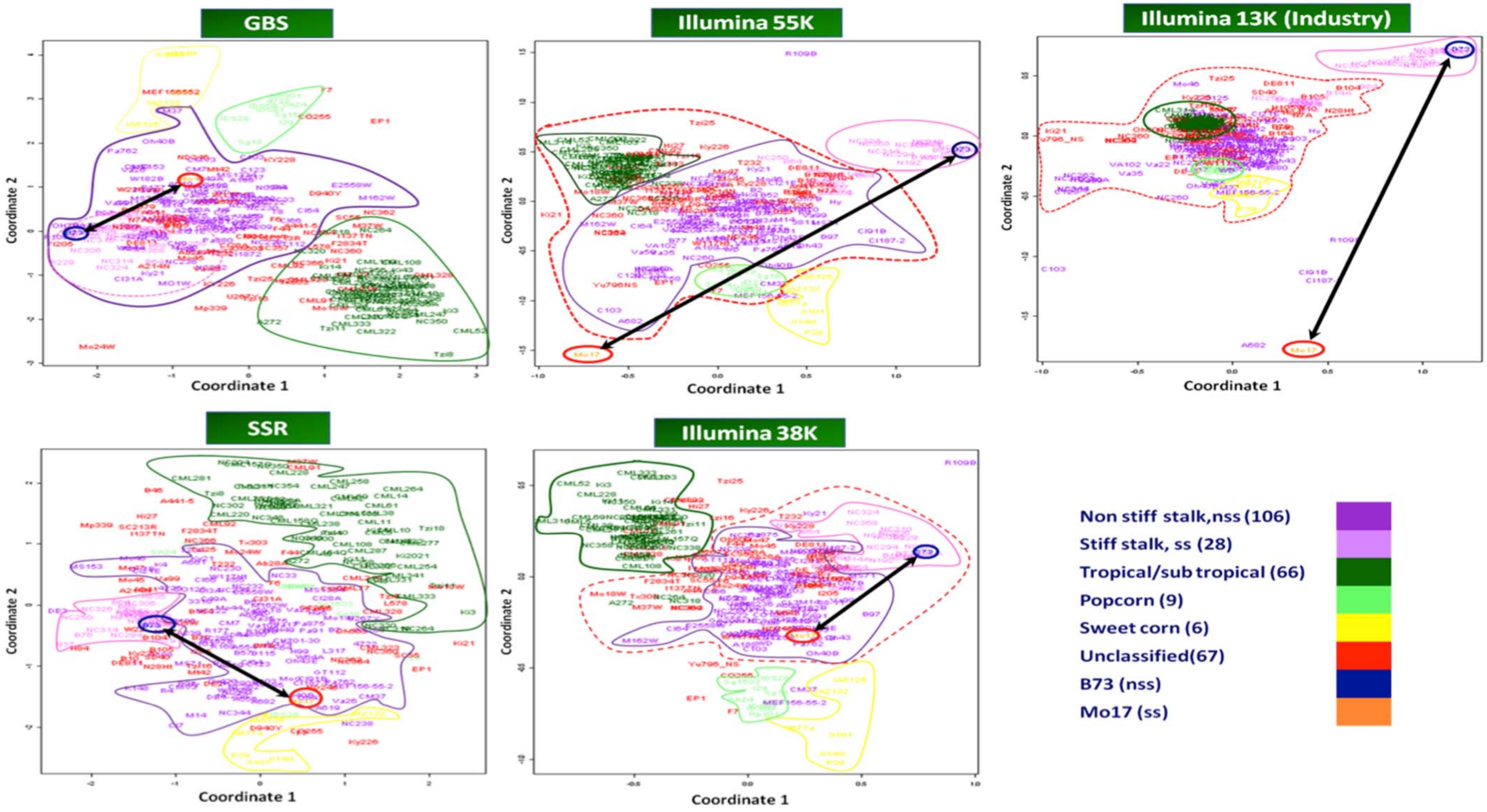
- GBS markers can be converted to SNPs or PCR assays of indels
- Develop SNP assays from polymorphic tags at same location
- Develop PCR primers from adjacent tags & hope for large indels



Phylogeny/Kinship

- Missing data not an issue for estimating pairwise genetic distance or kinship
  - Each pair of individuals has large, “random” sample of markers in common
- Works really well even in non-model organisms
  - Fei Lu’s previous talk on switchgrass
- Principle Coordinates Analysis better than Principle Components Analysis
  - Uses distance matrix rather than every genotype
  - Missing data not an issue for Prin. Coord. Analysis
- SNPs can be strongly affected by ascertainment bias
  - Panel used to discover the SNPs can severely distort estimates of population genetic parameters (e.g., kinship, diversity)
  - Industry SNPs on the Maize 55K SNP chip an extreme example

Less Ascertainment Bias than SNPs?



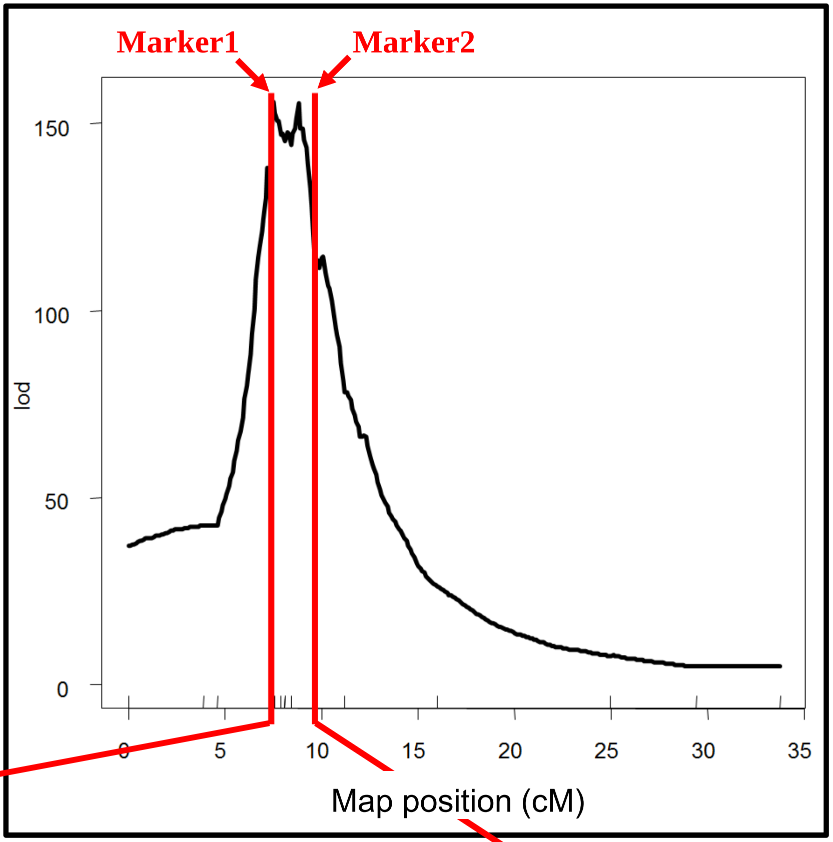
Dr. Ram Sharma – Visiting Scientist, Buckler lab, Cornell (unpublished)

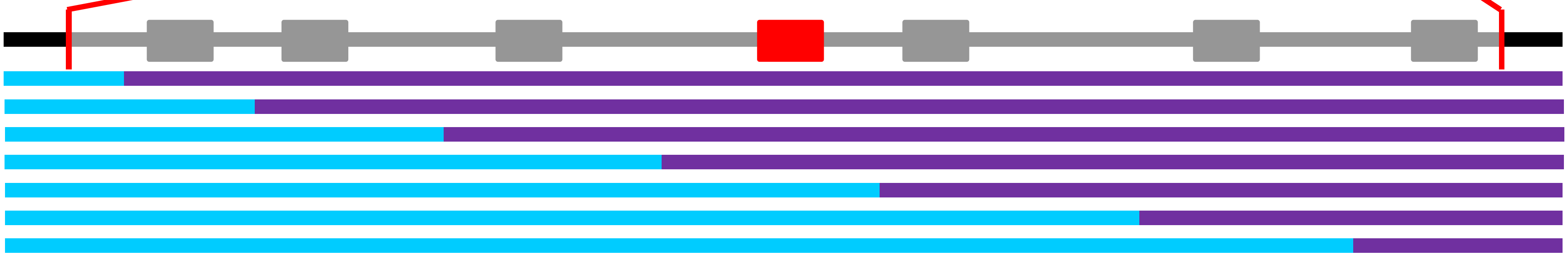
Linkage mapping of QTL in a biparental cross

- In maize, we use the reference genome to order markers
- With *Ape*KI, too many markers for traditional software (MapMaker, JoinMap, R-QTL etc.)
- Filter for a smaller set of markers with high coverage
- Use *Pst*I for fewer markers with higher coverage
- JoinMap can handle at least 3,000 markers
- Newer software?
  - MSTMap claims 10,000 – 100,000 markers
  - Others?

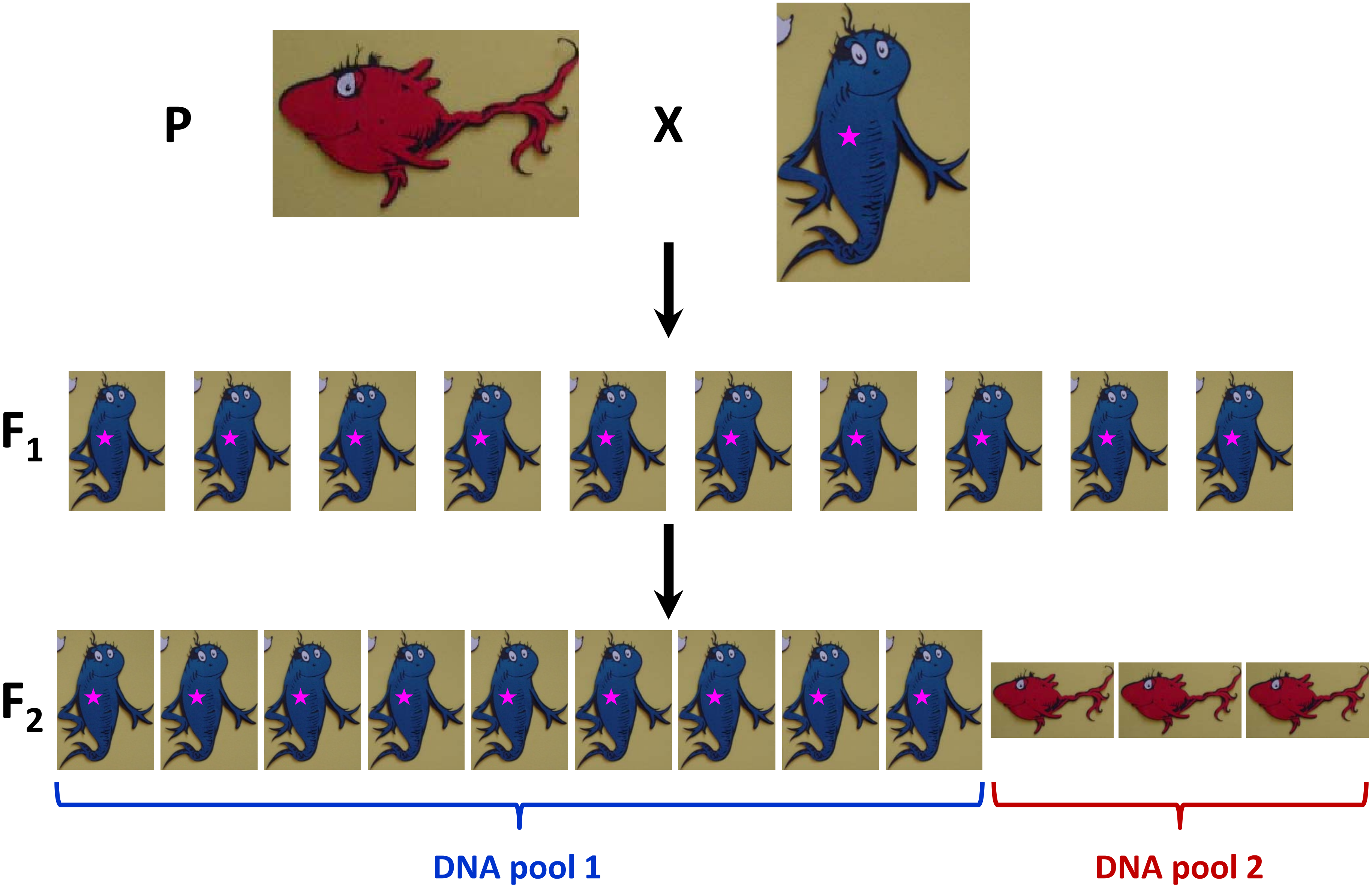
Fine mapping QTL

- Need to saturate interval containing QTL with markers
- GBS a good source of markers
- Also need to collect recombinants in the interval
- Near-isogenic lines (NILs) helpful (Mendelize)
- Good reference genome





Bulked Segregant Analysis



P

X

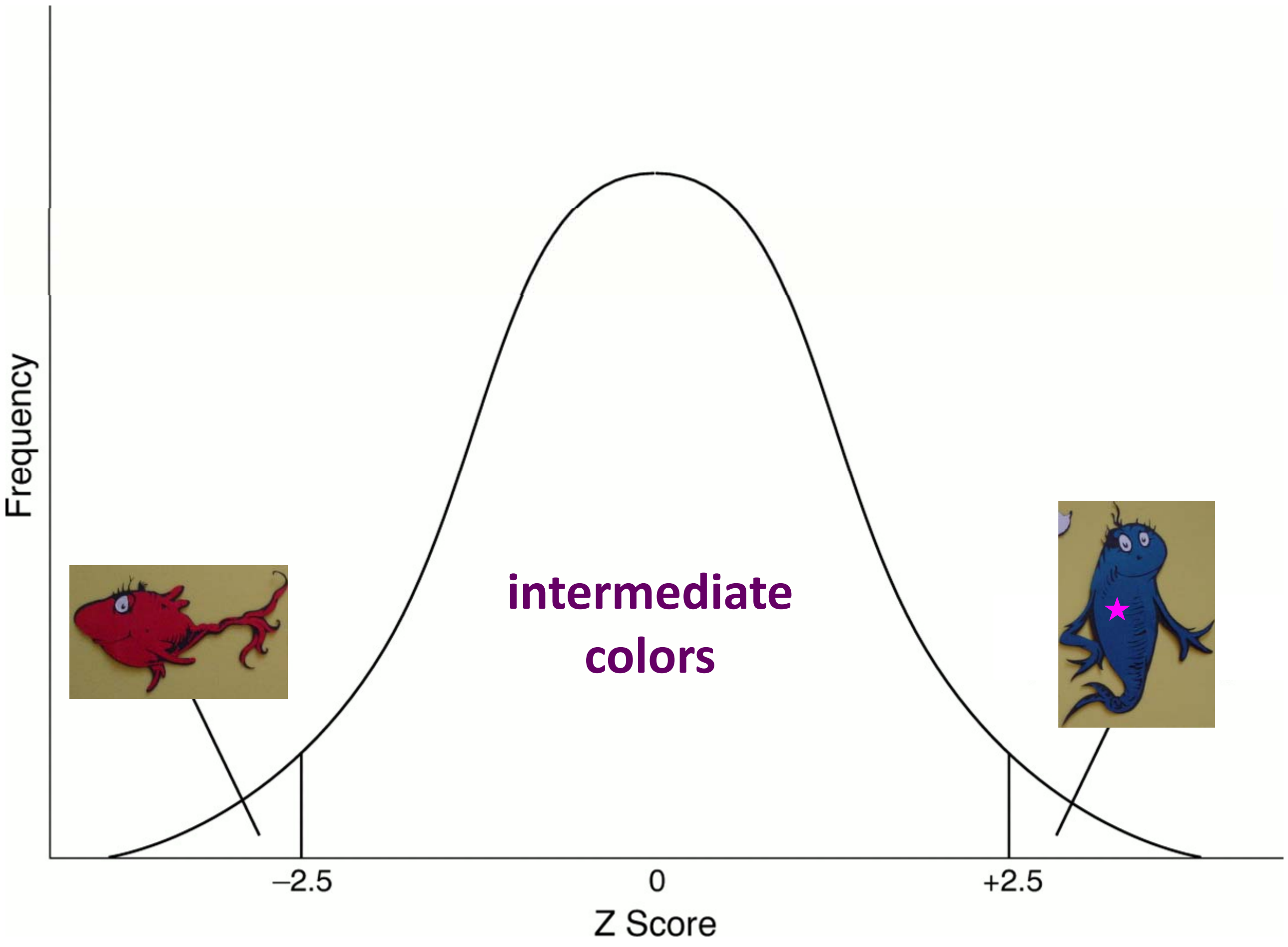
F_1

F_2

DNA pool 1

DNA pool 2

Bulked Segregant Analysis



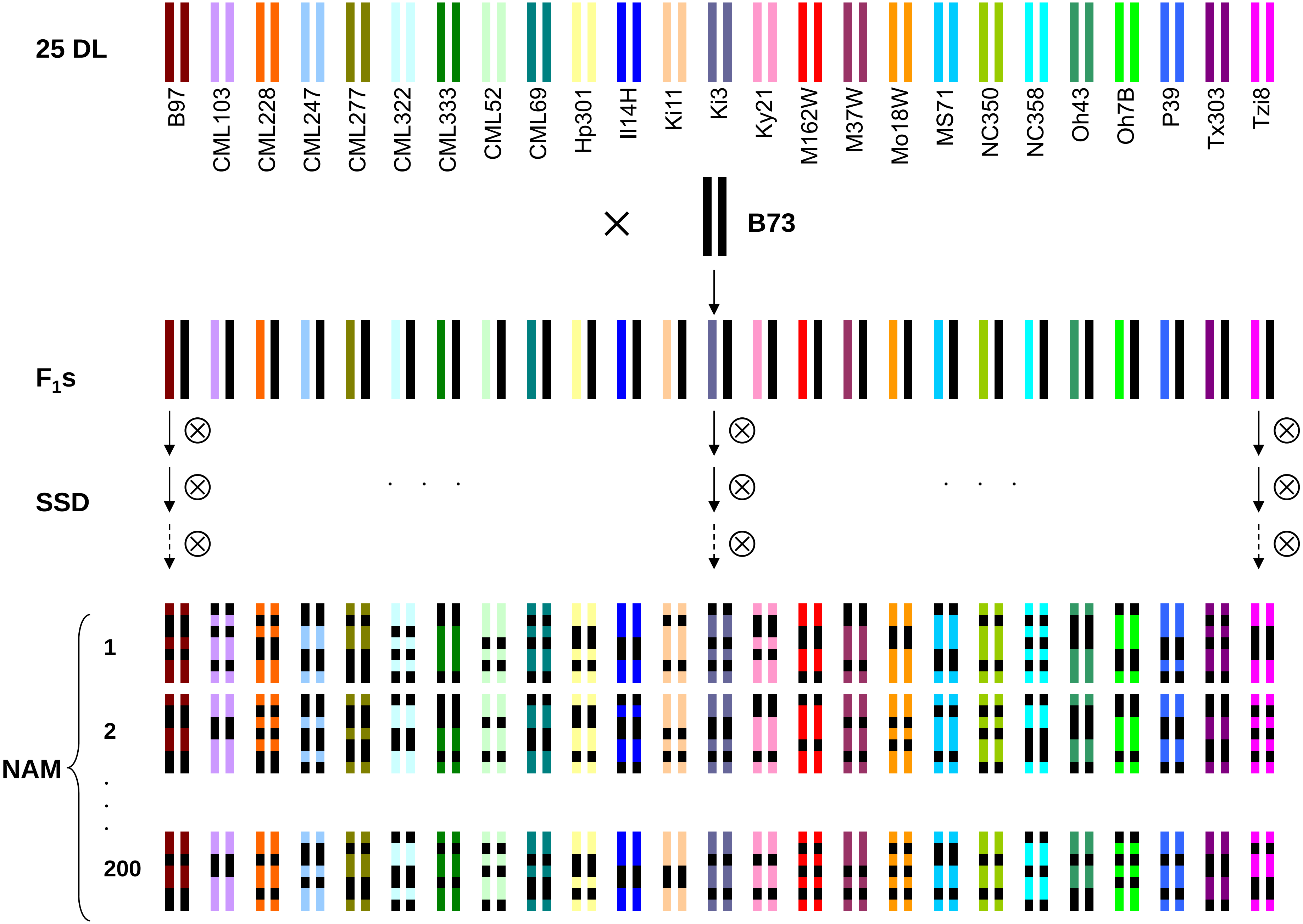
intermediate colors

Color

Genomic Selection & GWAS

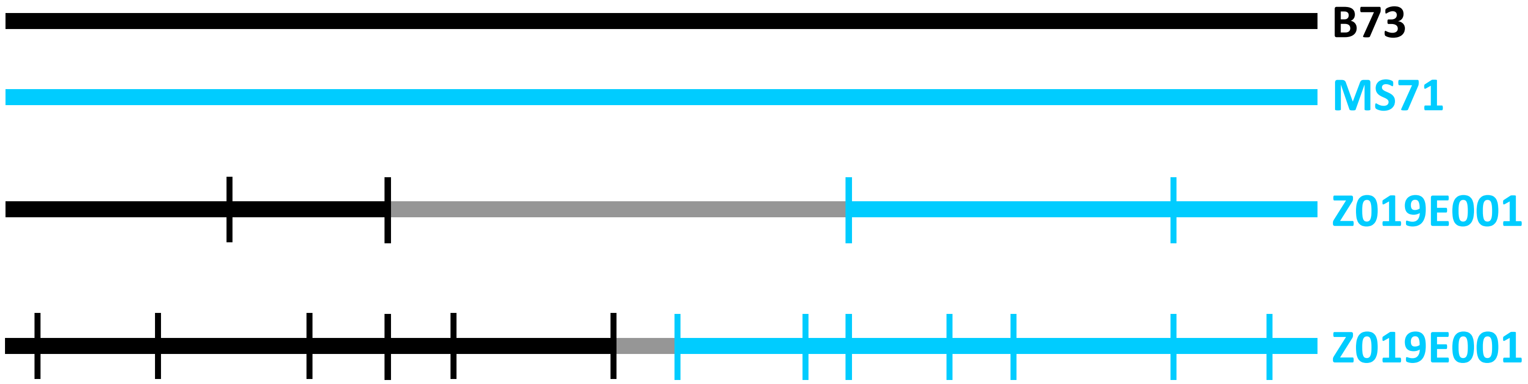
- **Complete data not required for genomic selection**
  - **Closely linked markers in LD cover for each other**
- **In contrast, missing data are more problematic for GWAS**
  - **imputation necessary, but might cause spurious results**
  - **avoid false imputation of biologically missing regions**
  - **area of active research**
- **In NAM-GWAS, imputation is much less of an issue**
  - **NAM = “Nested Association Mapping” population**

The maize NAM population was built for NAM-GWAS

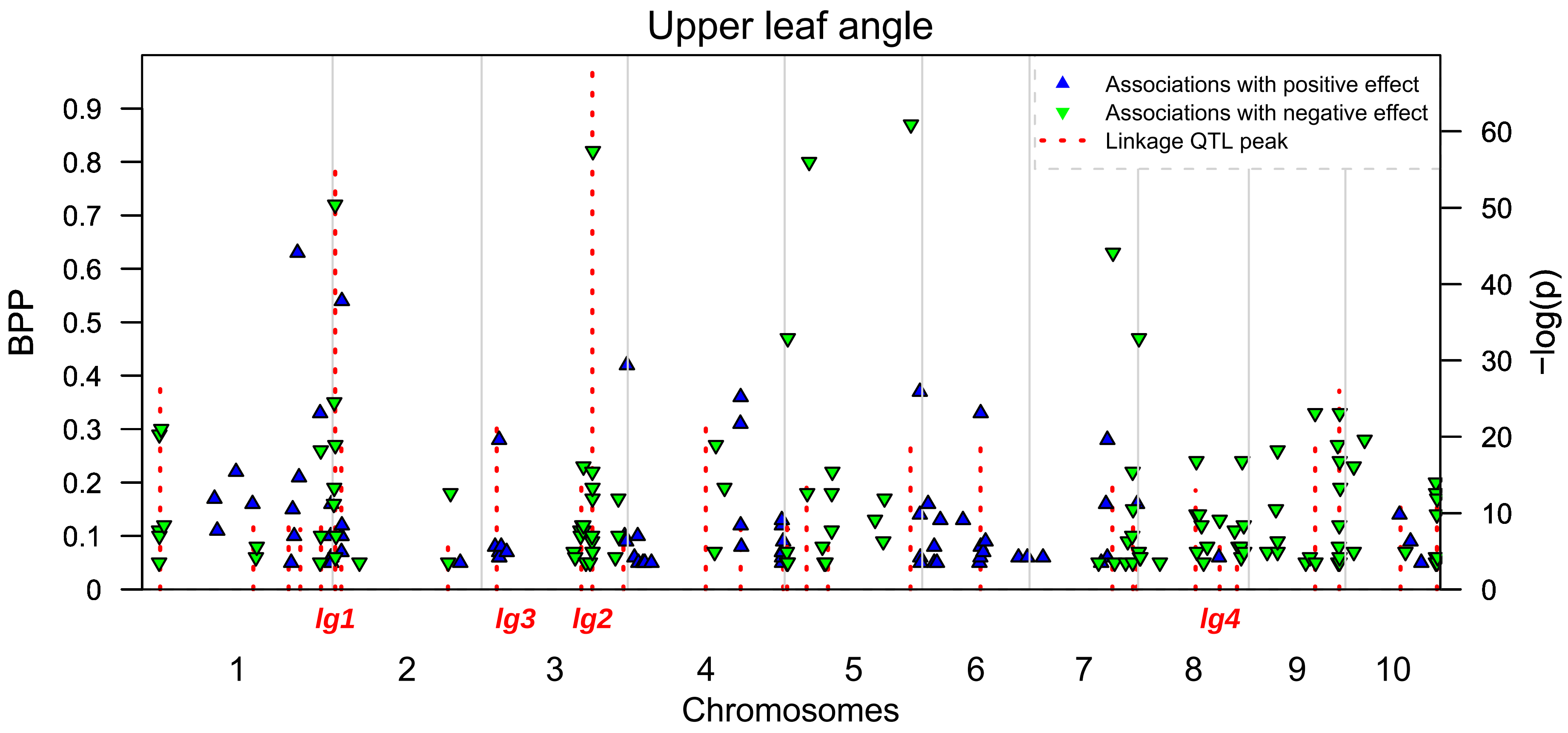


We are using GBS to pinpoint the location of cross overs in the NAM RILs

- B73 is the reference genome: complete knowledge
- Remaining NAM parents whole genome sequenced via Illumina at 4x coverage (paired end random sheared)
    - 26 million high quality SNPs
- Precise knowledge of crossover locations in NAM RILs allows us to more accurately project sequences of parents onto RILs:

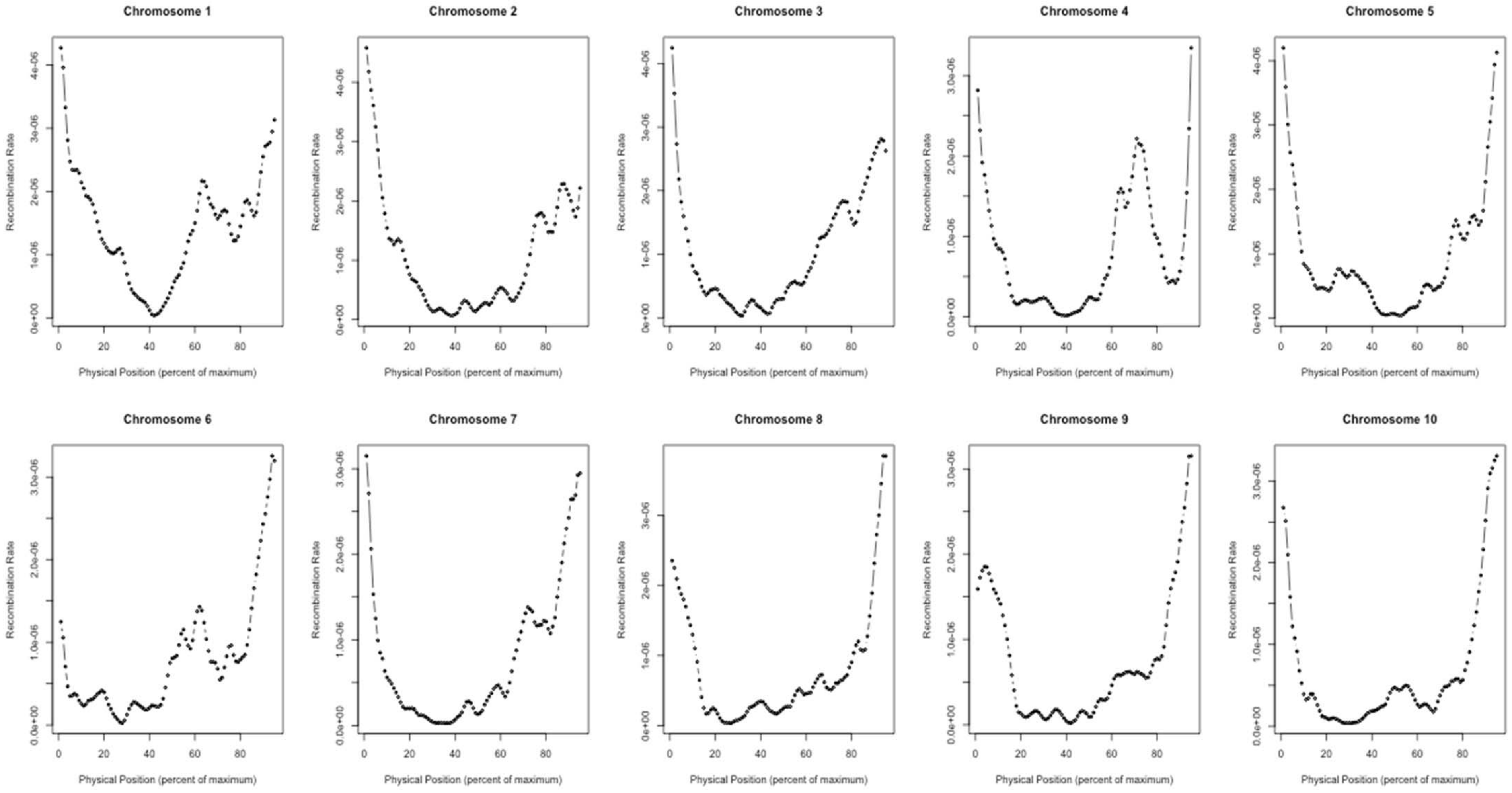


liguleless1 and *liguleless2* explain the two "biggest" leaf angle QTL



Tian, Bradbury, et al 2011 Nature Genetics

Recombination Rates for NAM from GBS Data



Peter Bradbury – USDA Scientist, Buckler lab, Cornell (unpublished)

The maize B73 reference genome: room for improvement?

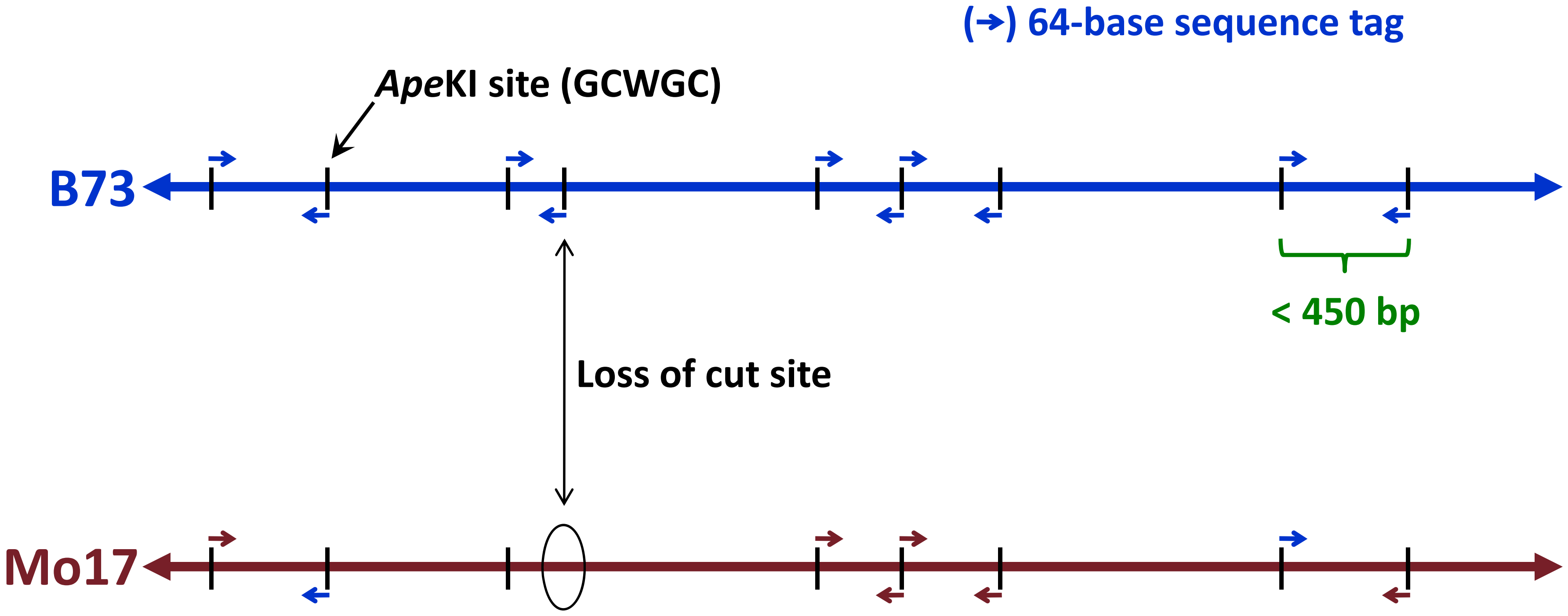
1) The B73 reference genome accurate for B73 but less so for other maize lines (*e.g.*, Mo17)
2) Even for B73, some regions of the genome are in the wrong place
3) Some large (multiple BAC) contigs could not be anchored
    - assigned to "chromosome 0"
    - 30 chr0 contigs in B73 RefGenV1
    - 17 chr0 contigs in B73 RefGenV2
4) Some regions of the genome are missing
    - ≈5% of B73 sequence is not in the B73 reference genome

The maize B73 reference genome: room for improvement?

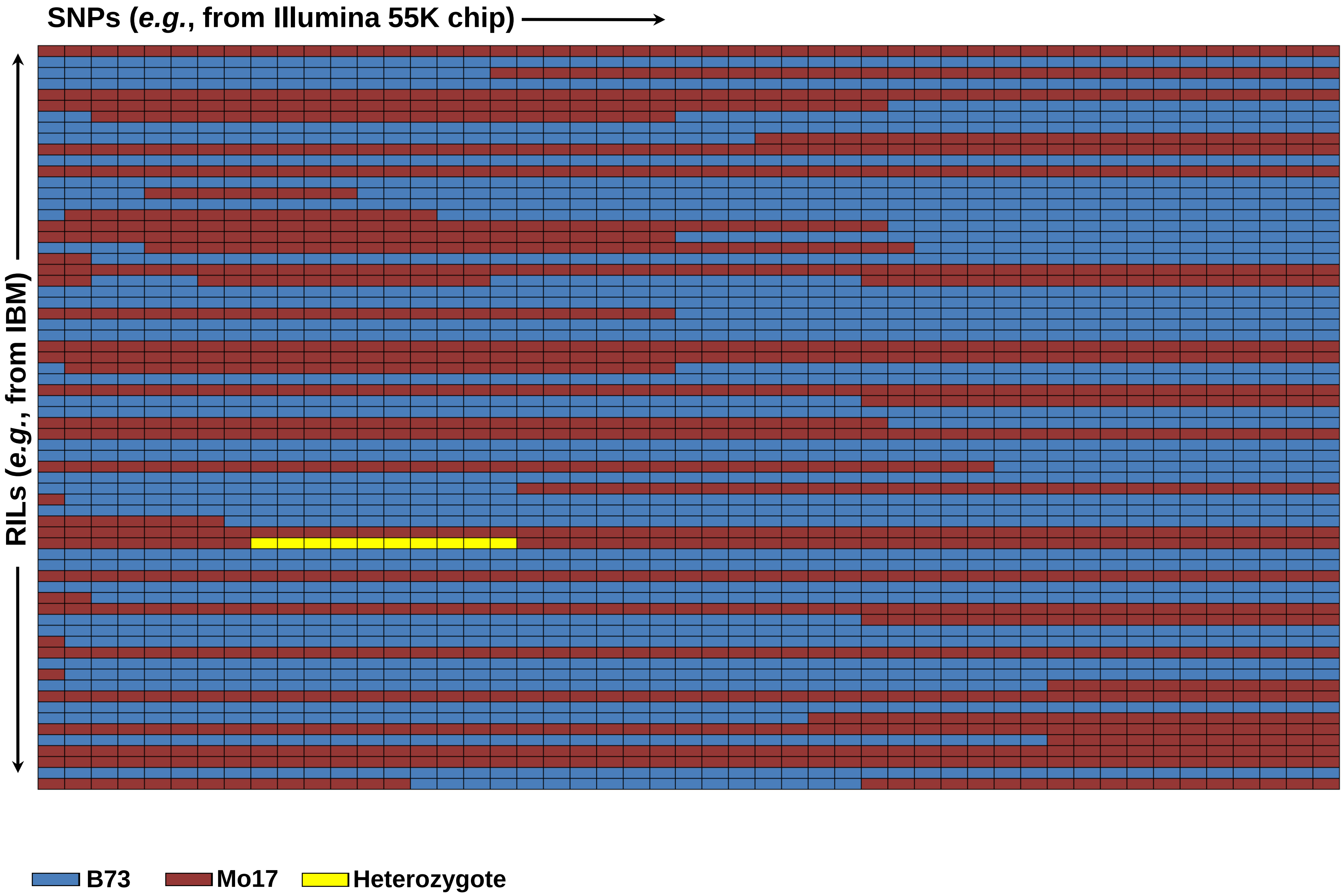
1) The B73 reference genome accurate for B73 but less so for other maize lines (*e.g.*, Mo17)
2) Even for B73, some regions of the genome are in the wrong place
3) Some large (multiple BAC) contigs could not be anchored
    - assigned to “chromosome 0”
    - 30 chr0 contigs in B73 RefGenV1
    - 17 chr0 contigs in B73 RefGenV2
4) Some regions of the genome are missing
    - ≈5% of B73 sequence is not in the B73 reference genome

Most tags can be mapped as individual alleles

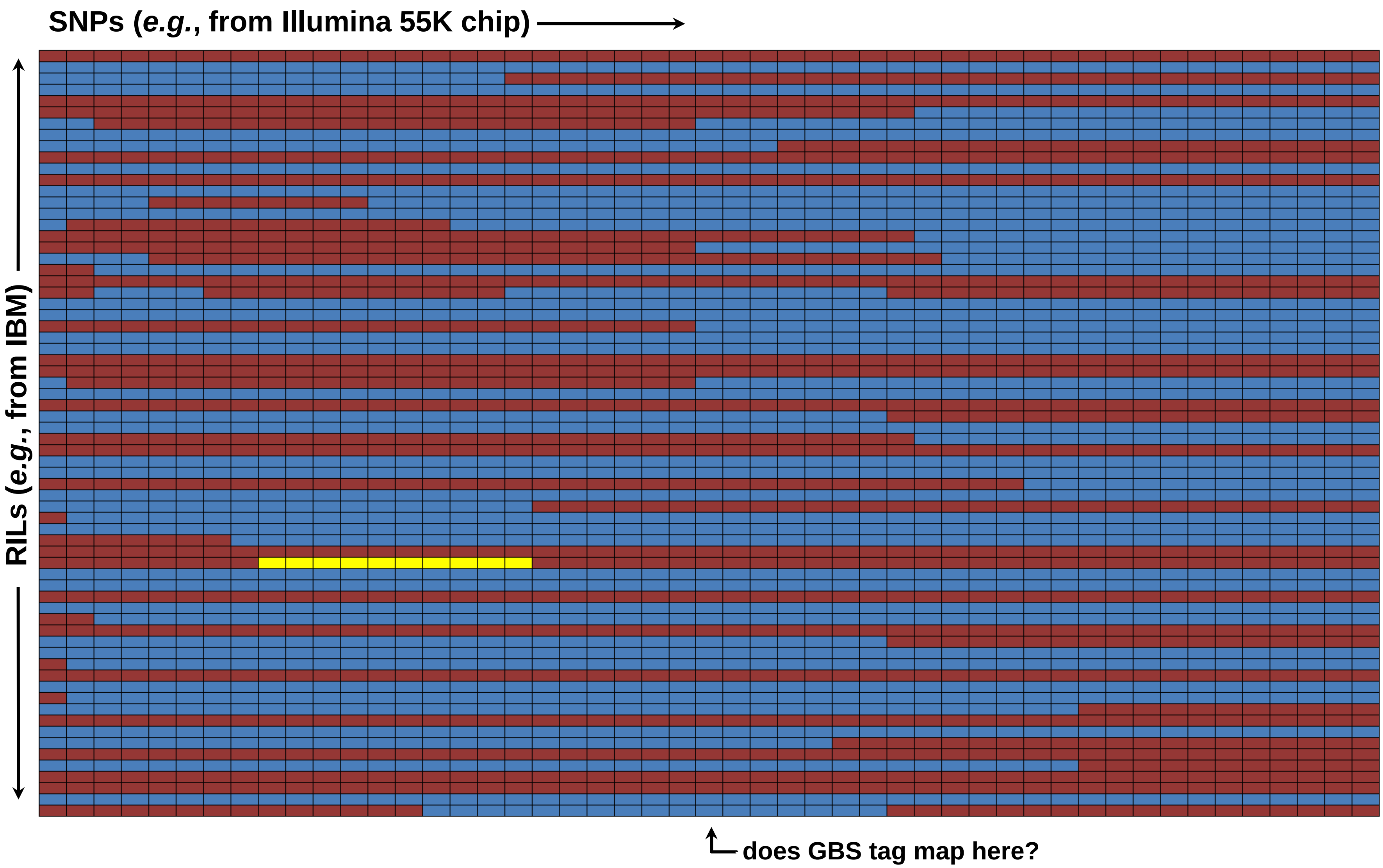
- In a biparental cross such as maize IBM (B73 x Mo17)
- Provided that they are polymorphic between the parents



Genetically mapping individual GBS alleles



Genetically mapping individual GBS alleles

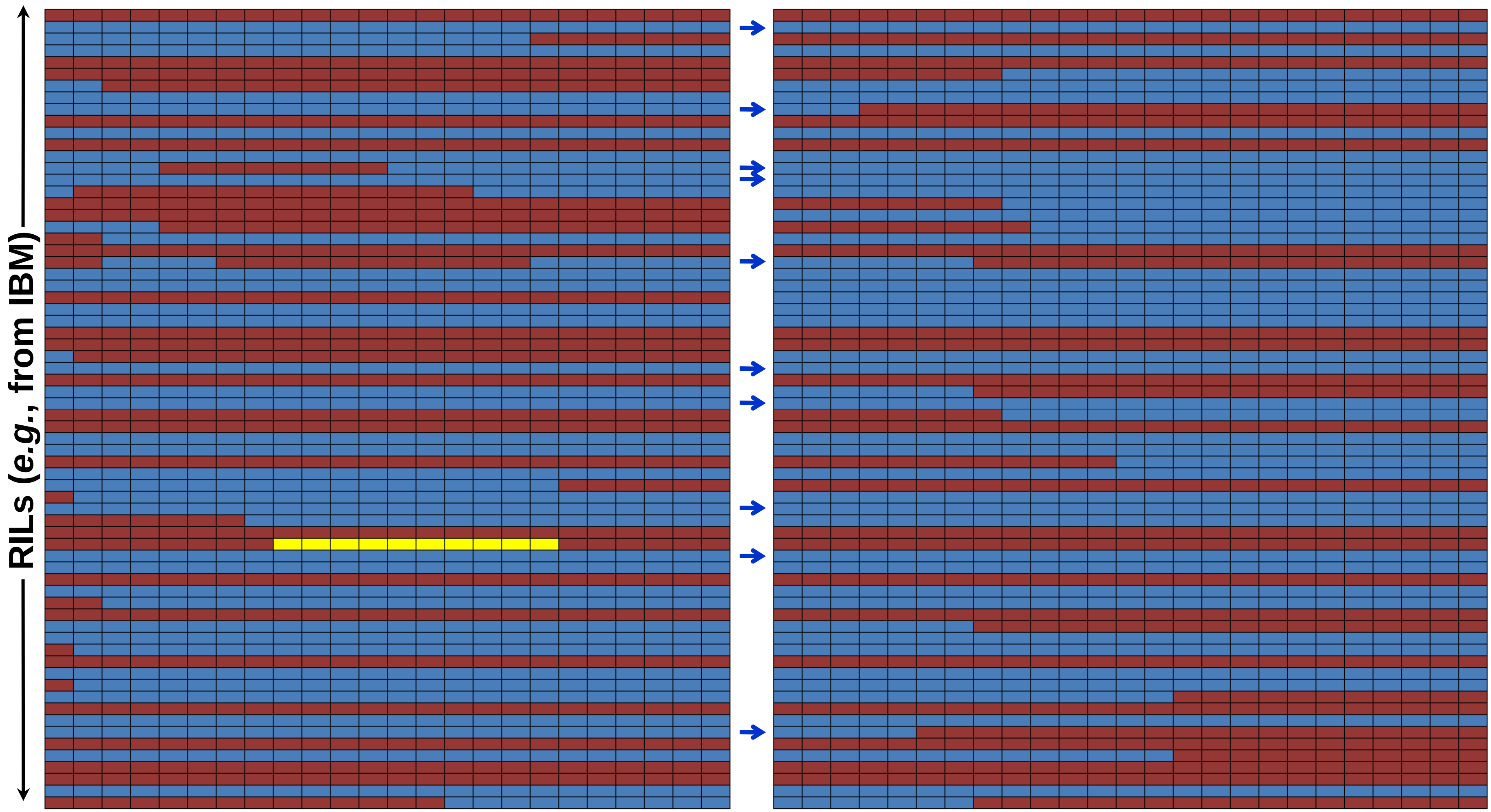


🟦 B73 🟥 Mo17 🟨 Heterozygote

Genetically mapping individual GBS alleles

(→) 64-base sequence tag (GBS coverage ~0.4x)

SNPs (*e.g.*, from Illumina 55K chip) →

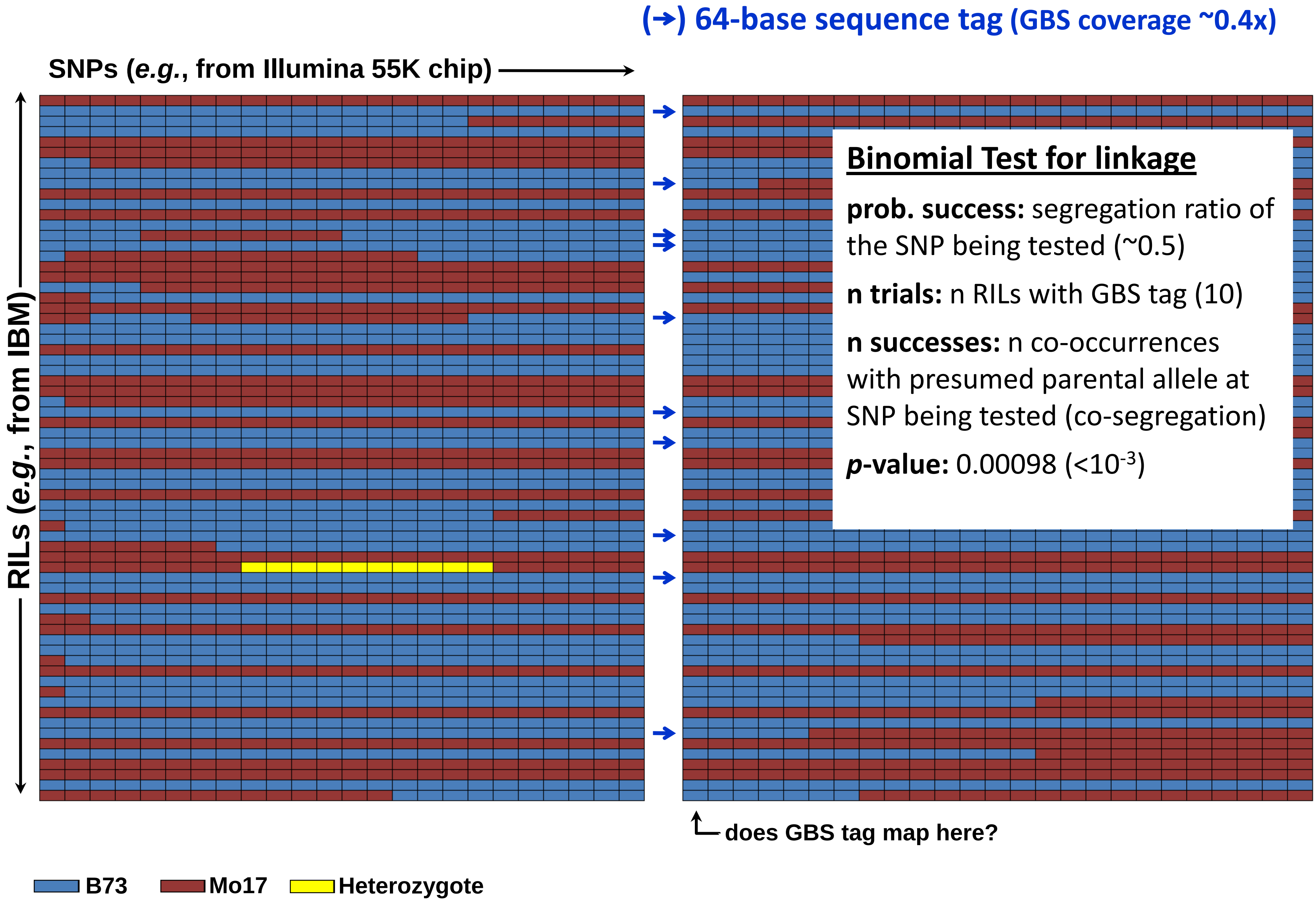


RILs (*e.g.*, from IBM)

↑ does GBS tag map here?

🟦 B73 🟥 Mo17 🟨 Heterozygote

Genetically mapping individual GBS alleles


(➔) 64-base sequence tag (GBS coverage ~0.4x)
SNPs (*e.g.*, from Illumina 55K chip)
RILs (*e.g.*, from IBM)
does GBS tag map here?
🟦 B73
🟥 Mo17
🟨 Heterozygote

Binomial Test for linkage

prob. success: segregation ratio of the SNP being tested (~0.5)

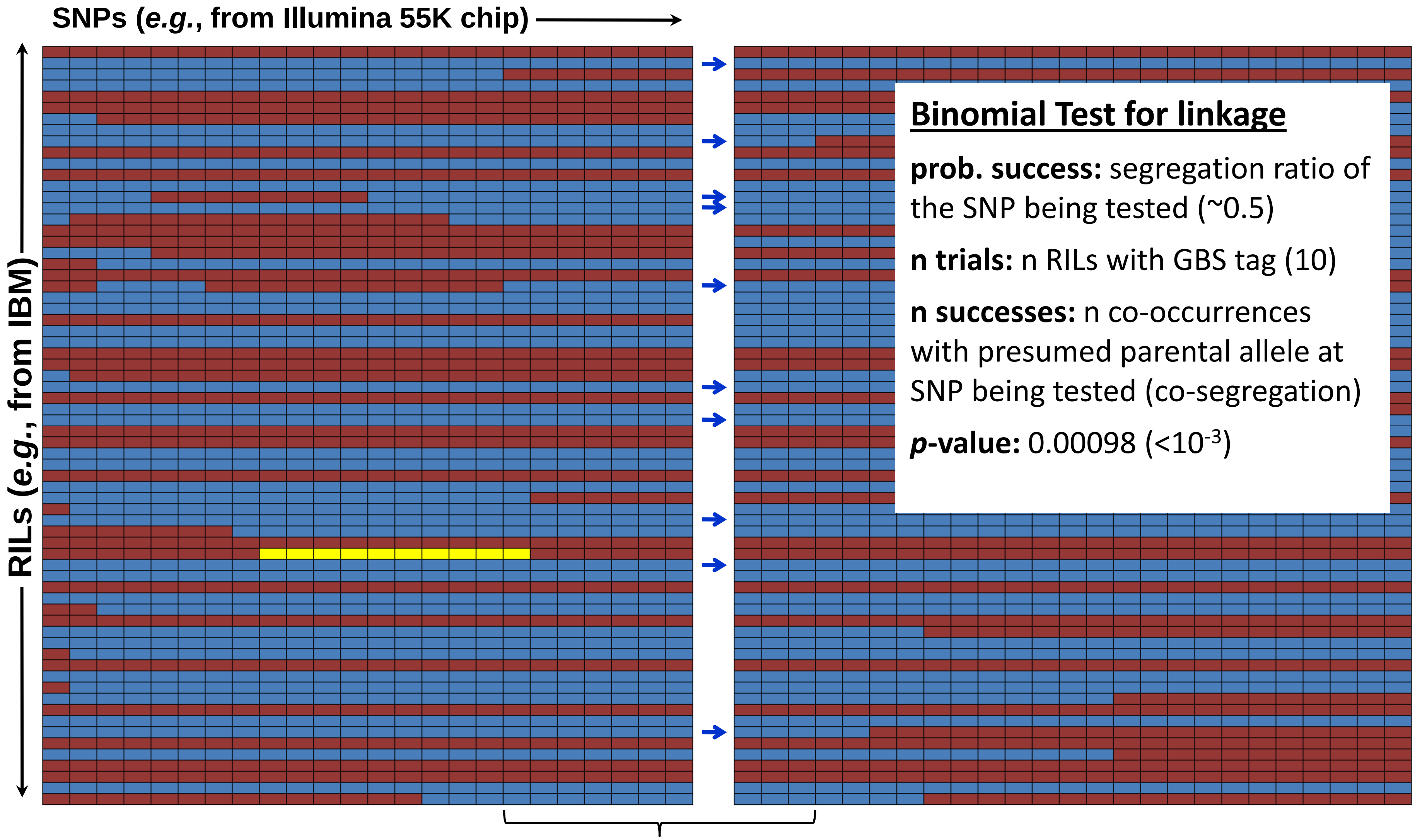
n trials: n RILs with GBS tag (10)

n successes: n co-occurrences with presumed parental allele at SNP being tested (co-segregation)

***p*-value:** 0.00098 ($<10^{-3}$)

Genetically mapping individual GBS alleles

(→) 64-base sequence tag (GBS coverage ~0.4x)



Binomial Test for linkage

prob. success: segregation ratio of the SNP being tested (~0.5)

n trials: n RILs with GBS tag (10)

n successes: n co-occurrences with presumed parental allele at SNP being tested (co-segregation)

***p*-value:** 0.00098 ($<10^{-3}$)

These 10 SNPs all tie (p = 9.8 x 10^{-4})

🟦 **B73** 🟥 **Mo17** 🟨 **Heterozygote**

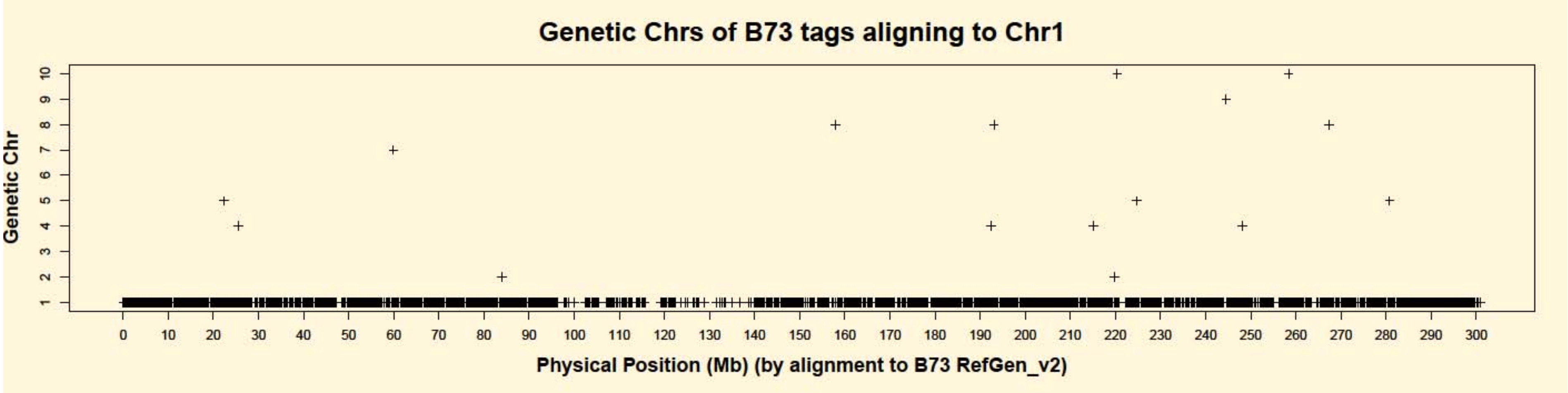
Genetically mapping individual GBS alleles in IBM

Min # Successes	*p*-value	max Recomb.	Total # GBS tags mapped	# B73 tags mapped	# Mo17 tags mapped
10	$<10^{-3}$	<5%	485,860	266,192	219,668
20	$<10^{-6}$	<5%	235,531	123,094	112,437
30	$<10^{-7}$	<5%	140,713	73,829	66,884

Genetically mapping individual GBS alleles in IBM

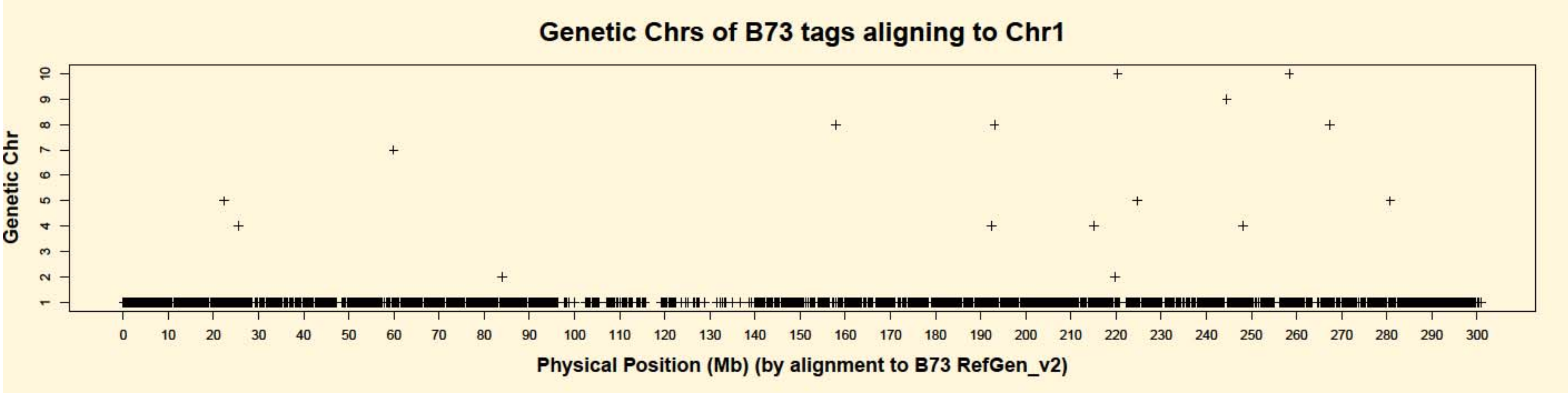
Min # Successes	*p*-value	max Recomb.	Total # GBS tags mapped	# B73 tags mapped	# Mo17 tags mapped
10	$<10^{-3}$	<5%	485,860	266,192	219,668
20	$<10^{-6}$	<5%	235,531	123,094	112,437
30	$<10^{-7}$	<5%	140,713	73,829	66,884

B73 reference genome highly accurate for B73…



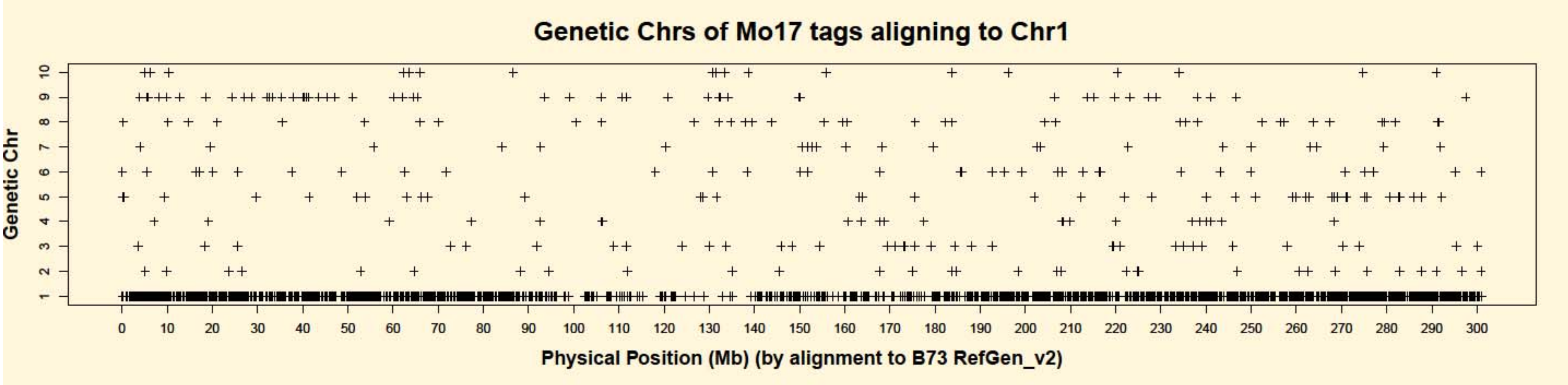
- 0.4% of B73 tags genetically map to different chromosome than they align to

B73 reference genome highly accurate for B73…



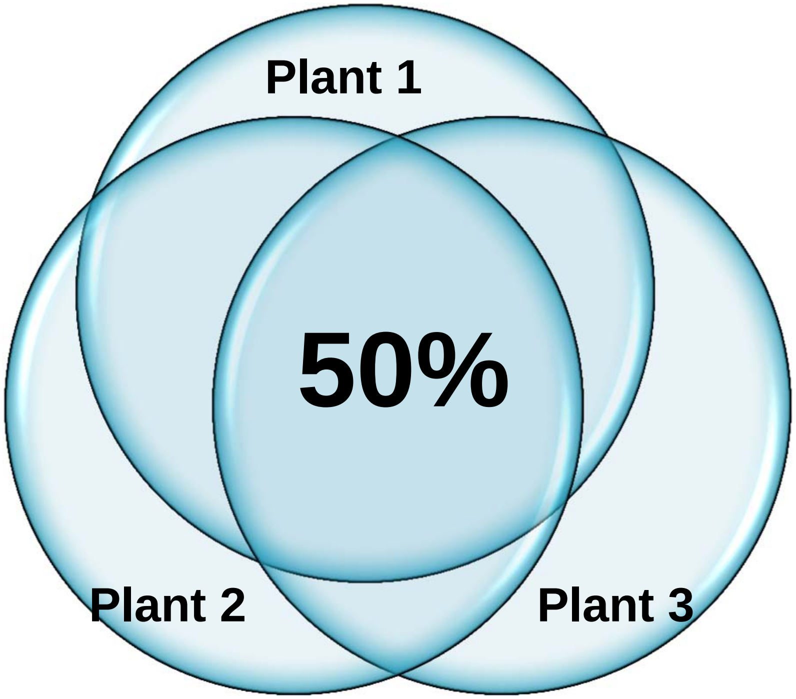
- 0.4% of B73 tags genetically map to different chromosome than they align to

…but far less so for other maize lines

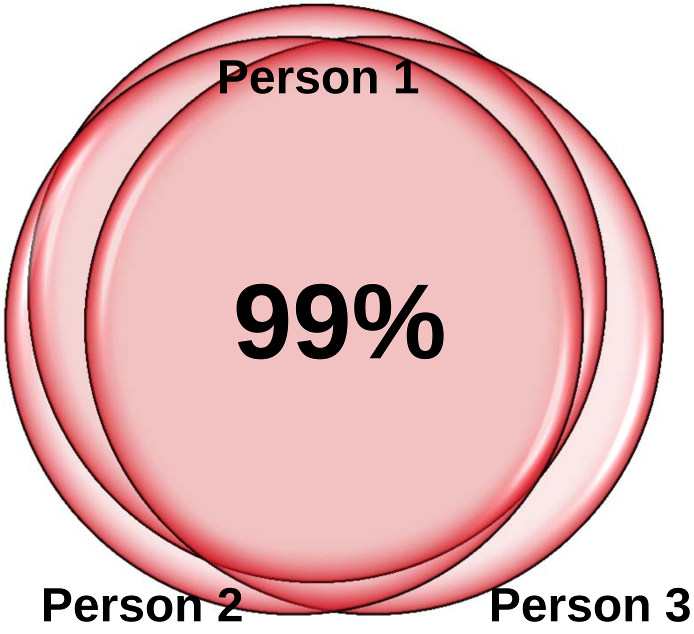


- 9.3% of Mo17 tags genetically map to different chromosome than they align to

Only 50% of the maize genome is shared between two varieties



Maize



Humans

Fu & Dooner 2002, Morgante et al. 2005, Brunner et al 2005

Numerous PAVs and CNVs - Springer, Lai, Schnable in 2010

Some chunks of the B73 reference genome are in the wrong place

Physical Chr	Start (Mb)	End (Mb)	Genetic Chr	Approx. Genetic Location (Mb)	# Tags
10	139.3	139.8	2	16.5–16.8	49
9	102.5	106.9	9	15–32	49
7	150.1	161.8	5	192–214	13
10	0.2	0.4	4	83–151	12
8	48.4	50	2	61–127	12
10	0.07	0.2	7	47–100	9
2	231.2	231.2	7	18–26	8
3	228.1	230.5	5	194–212	6

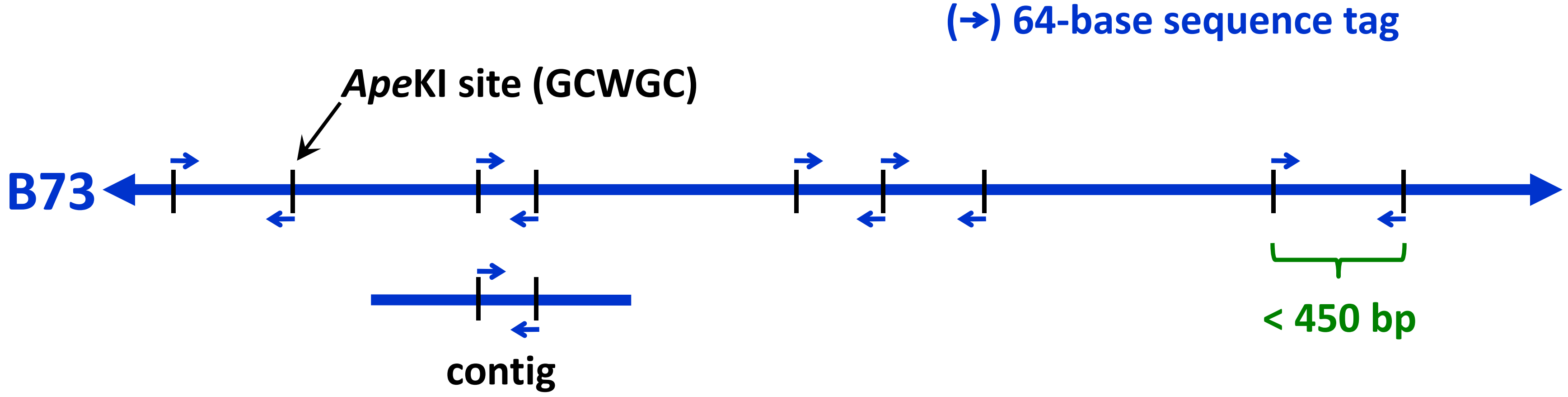
The maize B73 reference genome: room for improvement?

1) The B73 reference genome accurate for B73 but less so for other maize lines (*e.g.*, Mo17)
2) Even for B73, some regions of the genome are in the wrong place
3) Some large (multiple BAC) contigs could not be anchored
   - assigned to "chromosome 0"
   - 30 chr0 contigs in B73 RefGenV1
   - 17 chr0 contigs in B73 RefGenV2
4) Some regions of the genome are missing
   - ≈5% of B73 sequence is not in the B73 reference genome

Mapping Chr0 and *de novo* contigs via GBS

- **The sequences of Chr0 contigs are known**
  - **so we know which *Ape*KI GBS tags are present**
- ***De novo* contigs constructed from 454 whole genome sequencing**
  - **by collaborators at CSHL (Ware *et al.*)**
  - **can predict *Ape*KI GBS tags from these**
- **Created a pipeline to genetically map novel contigs using linkage populations**
- **Used IBM GBS data for proof of concept**

Adjacent tags on a Chr0 or *de novo* contig can be merged into haplotypes



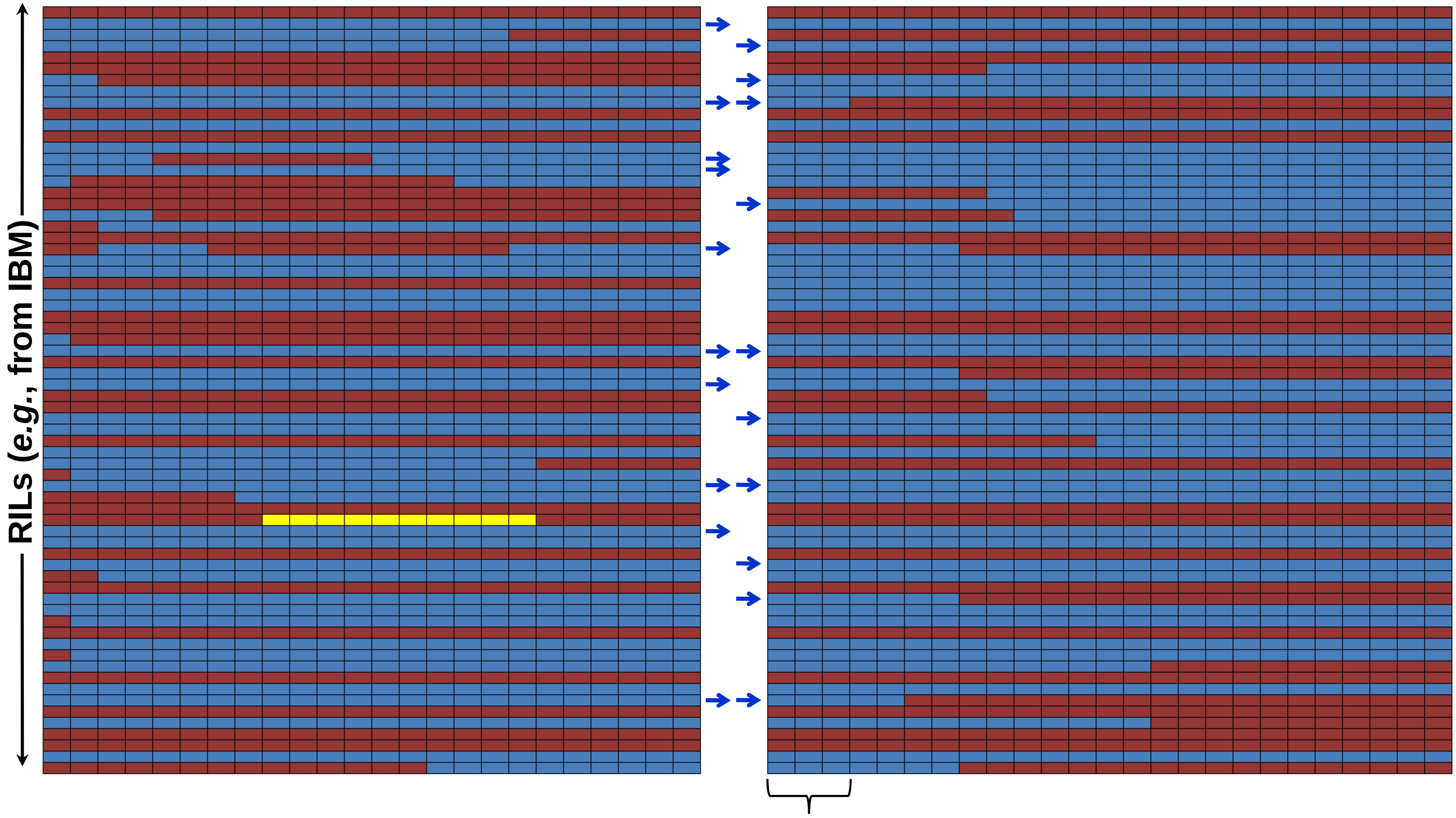
de novo (*e.g.*, from 454 or Illumina sequence)

Novel? (not included in B73 RefGen_v2)

Genetically mapping GBS *haplotypes*

(→) 64-base sequence tag (GBS coverage ~0.4x)

SNPs (*e.g.*, from Illumina 55K chip) →



RILs (*e.g.*, from IBM)

These 3 SNPs all tie (p = 1.5 x 10^{-5})

🟦 **B73** 🟥 **Mo17** 🟨 **Heterozygote**

Genetically mapping contigs via GBS

			# contigs genetically mapped	
Contigs	Total #	Source	novel	non-novel
Chr0	17	B73 RefGen_v2	8	---
B73 454 (k96)	3,964,387	CSHL	3,408	36,041
FLcDNA	61,477	CSHL	407	10,776

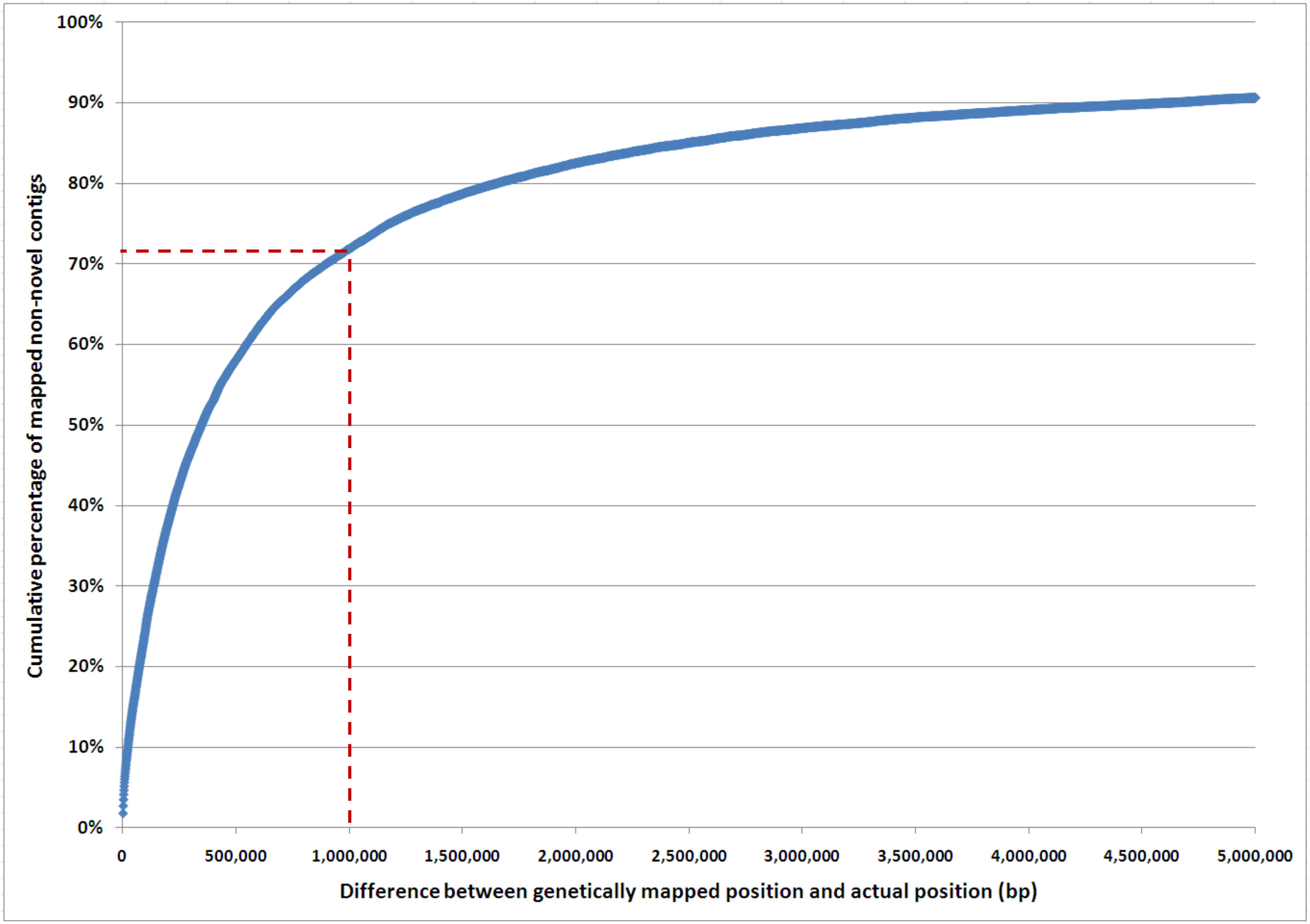
$p < 10^{-7}$

Genetically mapping contigs via GBS

			# contigs genetically mapped	
Contigs	**Total #**	**Source**	**novel**	**non-novel**
Chr0	17	B73 RefGen_v2	8	---
B73 454 (k96)	3,964,387	CSHL	3,408	36,041
FLcDNA	61,477	CSHL	407	10,776

$p < 10^{-7}$

>70% contigs genetically map to within 1 Mb of true position



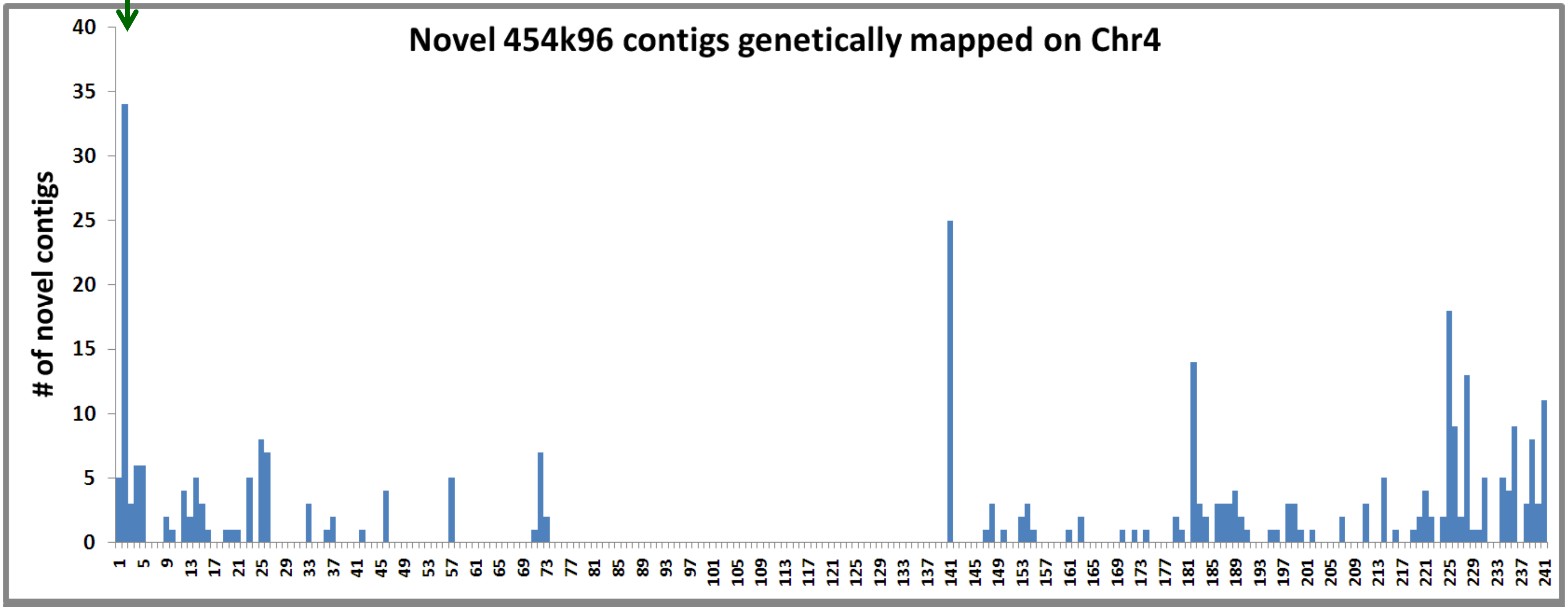
Genetically mapping contigs via GBS

			# contigs genetically mapped	
Contigs	Total #	Source	novel	non-novel
Chr0	17	B73 RefGen_v2	8	---
B73 454 (k96)	3,964,387	CSHL	3,408	36,041
FLcDNA	61,477	CSHL	407	10,776

$p < 10^{-7}$

Some regions of reference genome are missing large chunks

Telomere of Chr4 is a prime target for future improvement



Conclusions – Improving the genome

- GBS data from a mapping population where one of the parents is the reference genome can help improve that reference genome
- Can help place:
  - unanchored contigs (chromosome 0)
  - contigs/BACS that have been misplaced (wrong chromosome)
  - novel contigs from *de novo* sequencing (missing from the reference)
- These improvements incorporated into B73 RefGenV3
- Can uncover major structural differences between lines

This coming year – Improving the genome

- **Add in GBS data from NAM for much higher resolution**
  - **Currently constructing a GBS framework map of NAM**
  - **Anchor as many novel genes & contigs as possible**

- **Use GBS SNP calls in NAM plus >10,000 additional maize lines and map tags by LD (association mapping)**
  - **Further improve genetic mapping resolution?**
  - **Preliminary results: Median resolution = 90Kb**

Some potential applications of GBS Data

- Marker discovery
- Phylogeny/Kinship
- Linkage mapping of QTL in a biparental cross
- Fine-mapping QTL
- Bulked segregant analysis
- Genomic selection
- Genome Wide Association Studies (GWAS)
- NAM-GWAS
- Improving reference genome assembly