Genotyping – By – Sequencing (GBS) in Crop Plants

Rajneesh Paliwal

ICRISAT
Science with a human face

CGIAR
RESEARCH PROGRAM ON
DrylandCereals

Topics presented

GBS OVERVIEW AND SNP CALLING

GENOME WIDE ASSOCIATION STUDY (GWAS)

APPLICATION OF GBS-SNP INCLUDING GWAS

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

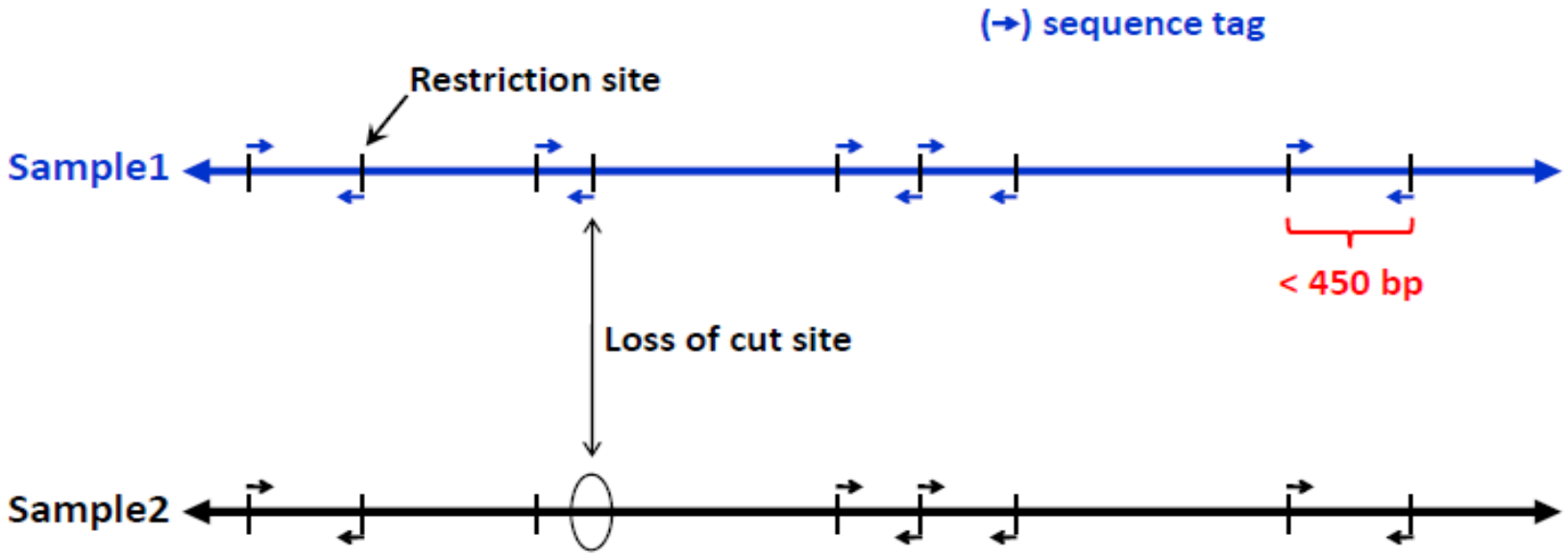
Target enrichment

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

Restriction Enzymes (REs)

- ***Technically less challenging***
- Methylation sensitive Res filter out repetitive genomic fraction

Overview of Genotyping by Sequencing (GBS)

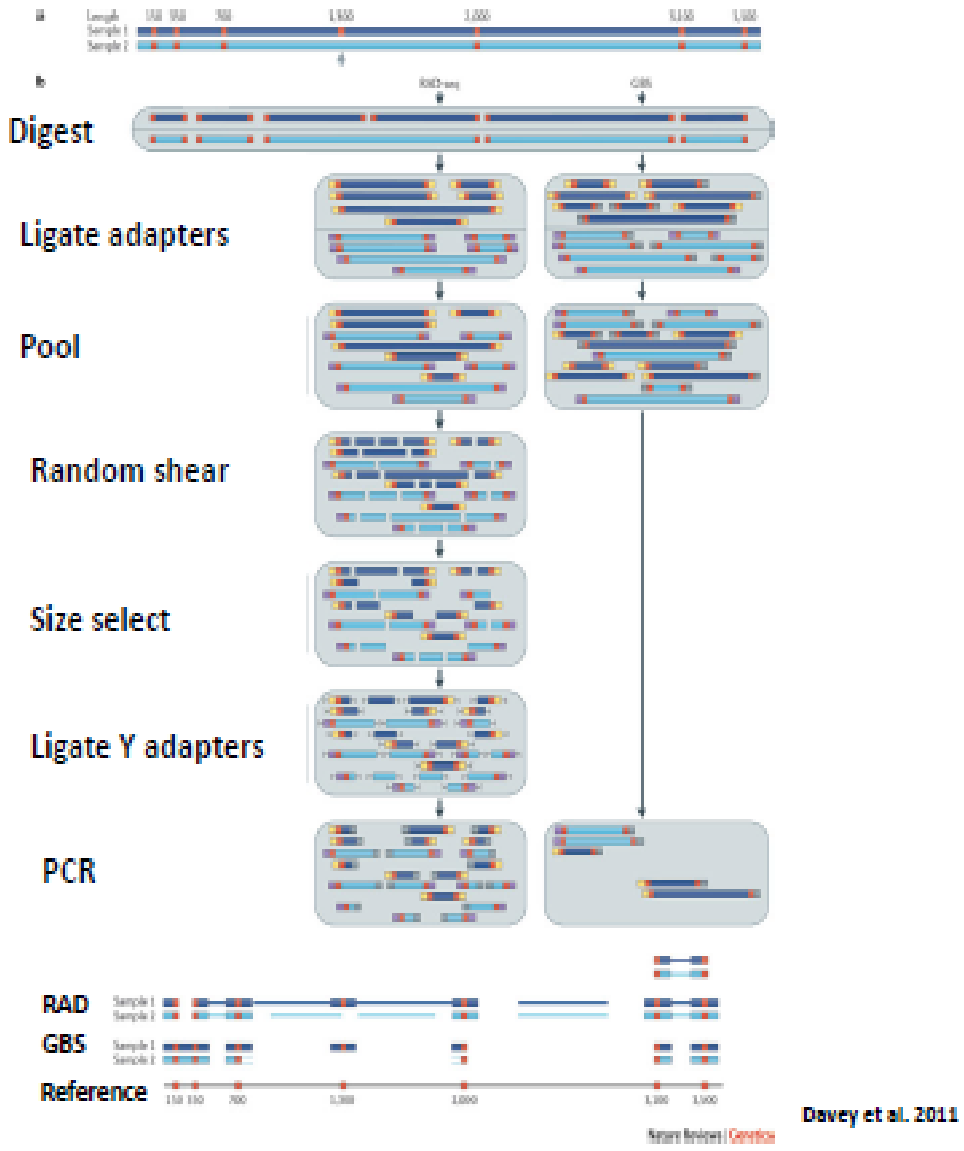


- Focuses NextGen sequencing power to ends of restriction fragments in both reference and non-reference genome plants
- Both SNPs and presence/absence markers can be scored
- Small Indels are identified but are not scored

GBS is a simple, highly multiplexed system for constructing libraries for NGS

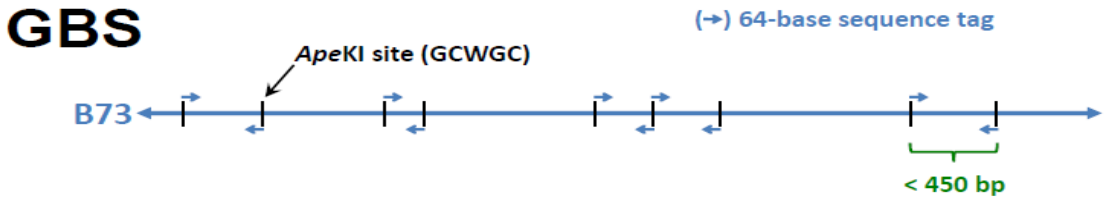
- ❑ Reduced Sample handling
- ❑ Few PCR & purification steps
- ❑ No DNA size fractionation
- ❑ Efficient barcoding system
- ❑ Simultaneous marker discovery and genotyping
- ❑ Scales very well

RADs vs GBS SNPs



Davey et al. 2011

Reduced Genome Representation through GBS



Most frequently asked question for new species

"How many SNP will I get?"

Answer: "It depends……."

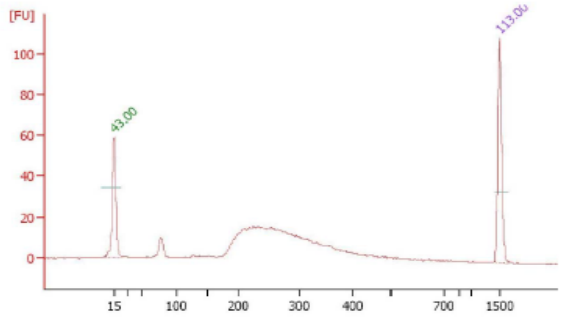
- Genome size and expected hetrozygosity affects size of fragment pool for desired amount of sequence coverage (enzyme choice and multiplex level).
- Amount of extent diversity and how well your samples capture that diversity
- Reference genome sequence? 3-4x more SNPs attained by aligning to a reference sequence

Cont…

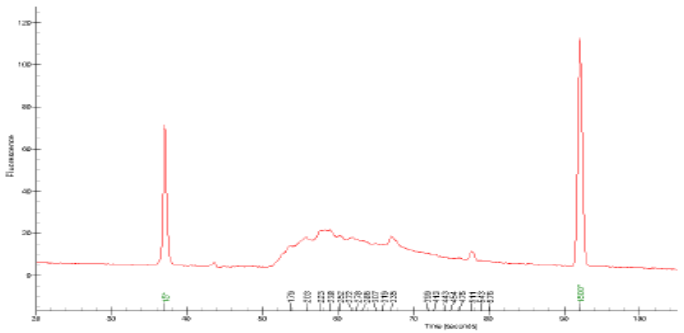
Species	Genome Size (Mb)	Enzyme	Sample Size	No. SNPs
Maize	2,600	*Ape*KI	33,000	2,200K
Rice	400	*Ape*KI	850	60K
Grape	500	*Ape*KI	8,000	1,200K
Willow*	460	*Ape*KI	459	23K
Pine*	16,000	*Ape*KI	12	63K
Vole*	3,400	*Pst*I	283	53K
Fox*	2,400	*EcoT22*I	48	16K
Cow	3,000	*Pst*I	48	64K
Neurospora (fungus isolates)	40	*Ape*KI	384	100K

*No reference genome. UNEAK analysis pipeline used for analysis. To avoid homology/paralogy issues this pipeline calls SNPs very conservatively.

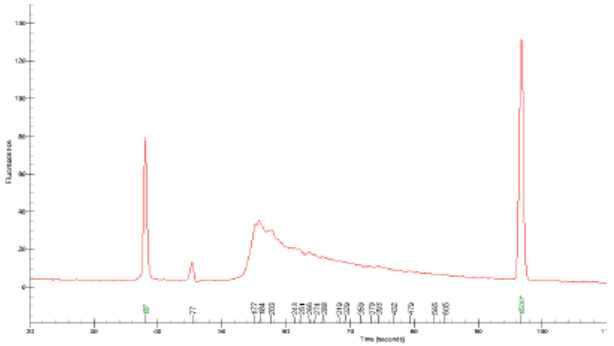
*Pst*I works well for mammals and birds.



EcoT22I works well for fish, amphibians and invertebrates.



*Ape*KI works well for grasses.



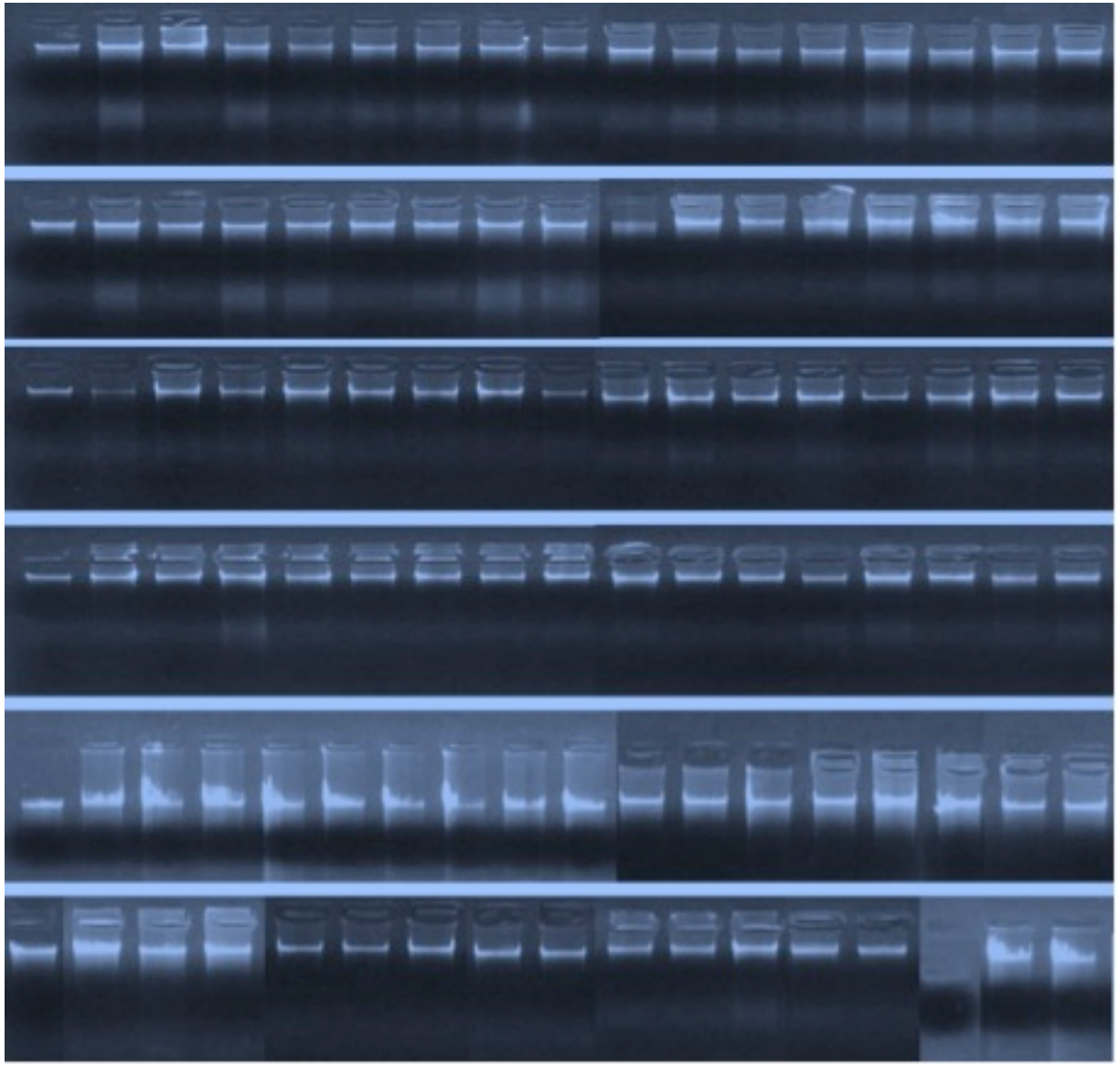
Most significant GBS technical issues?

- ❑ DNA Quality
- ❑ DNA quantification

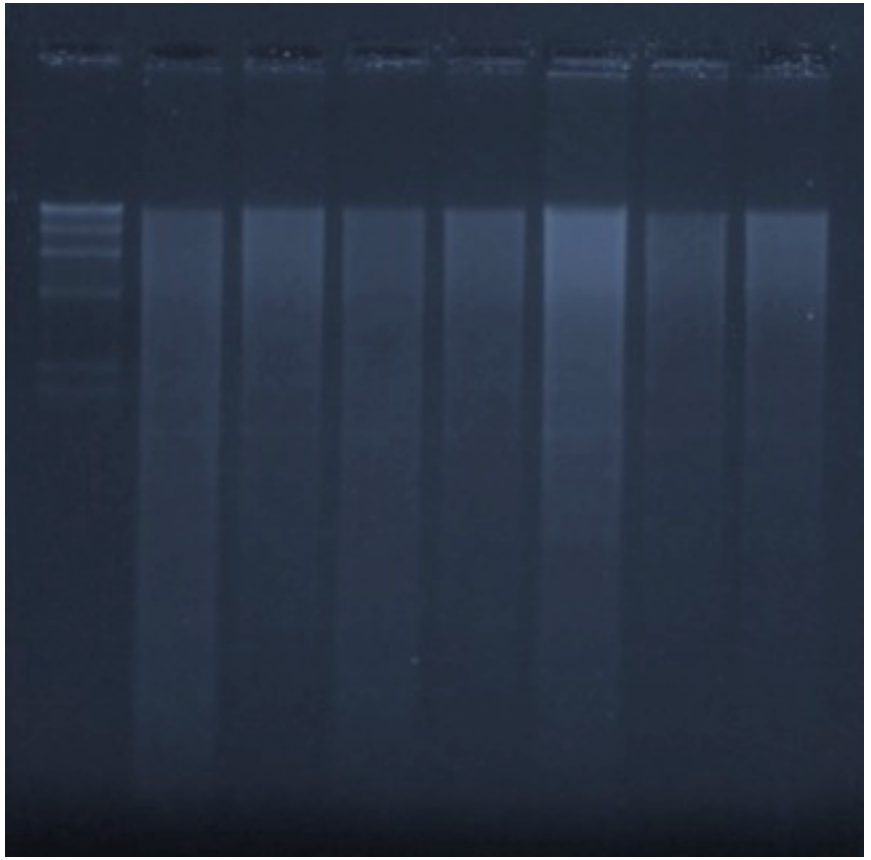
SAMPLE PROCESSING

DNA – Next Generation Sequencing

- High molecular weight
- Good quality

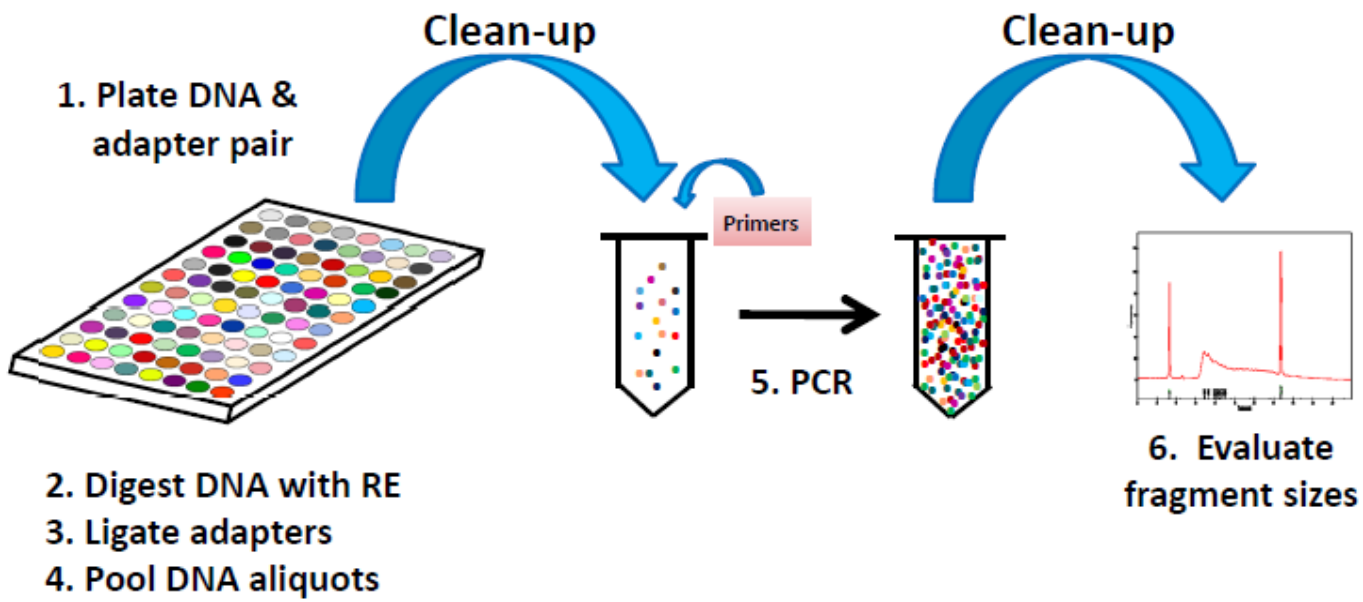


Restriction Digestion

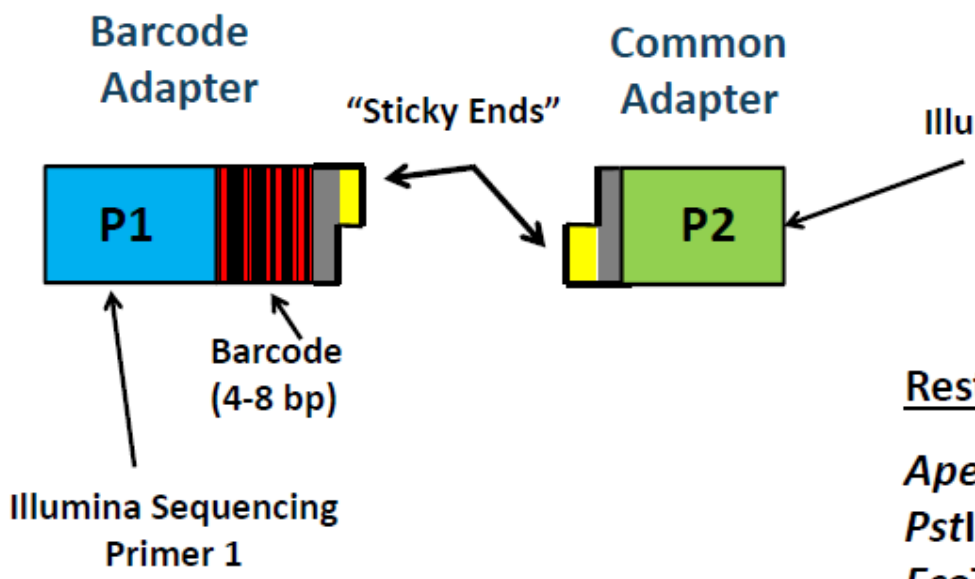


GBS 96- or 384-plex Protocol

(http://www.maizegenetics.net/gbs-overview)



GBS Adapters and Enzymes

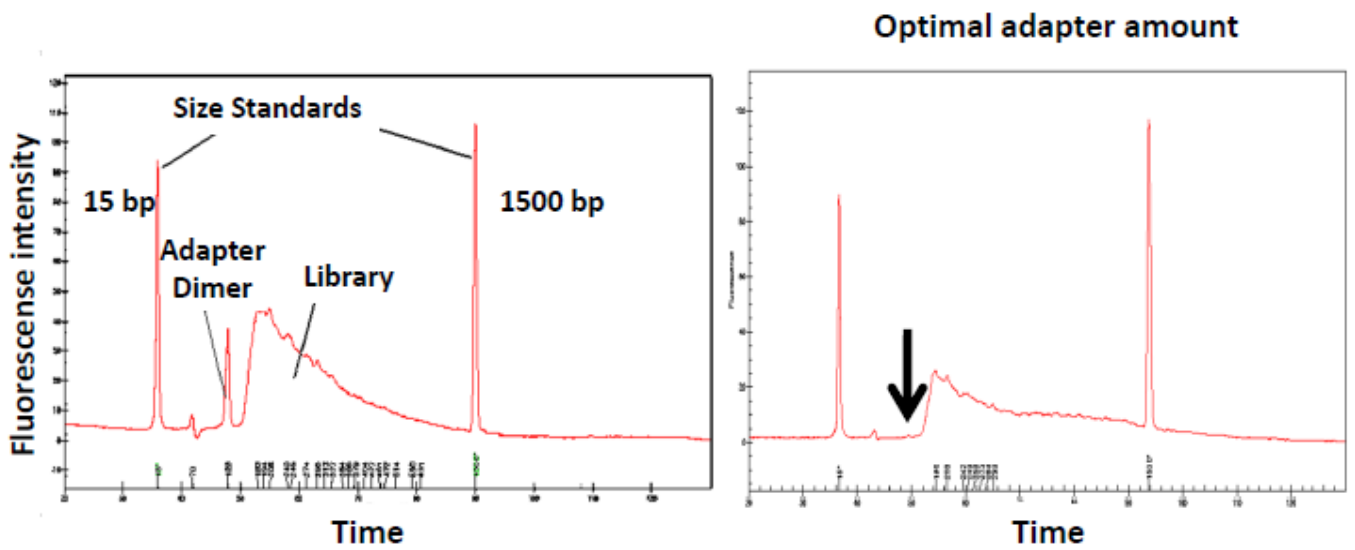


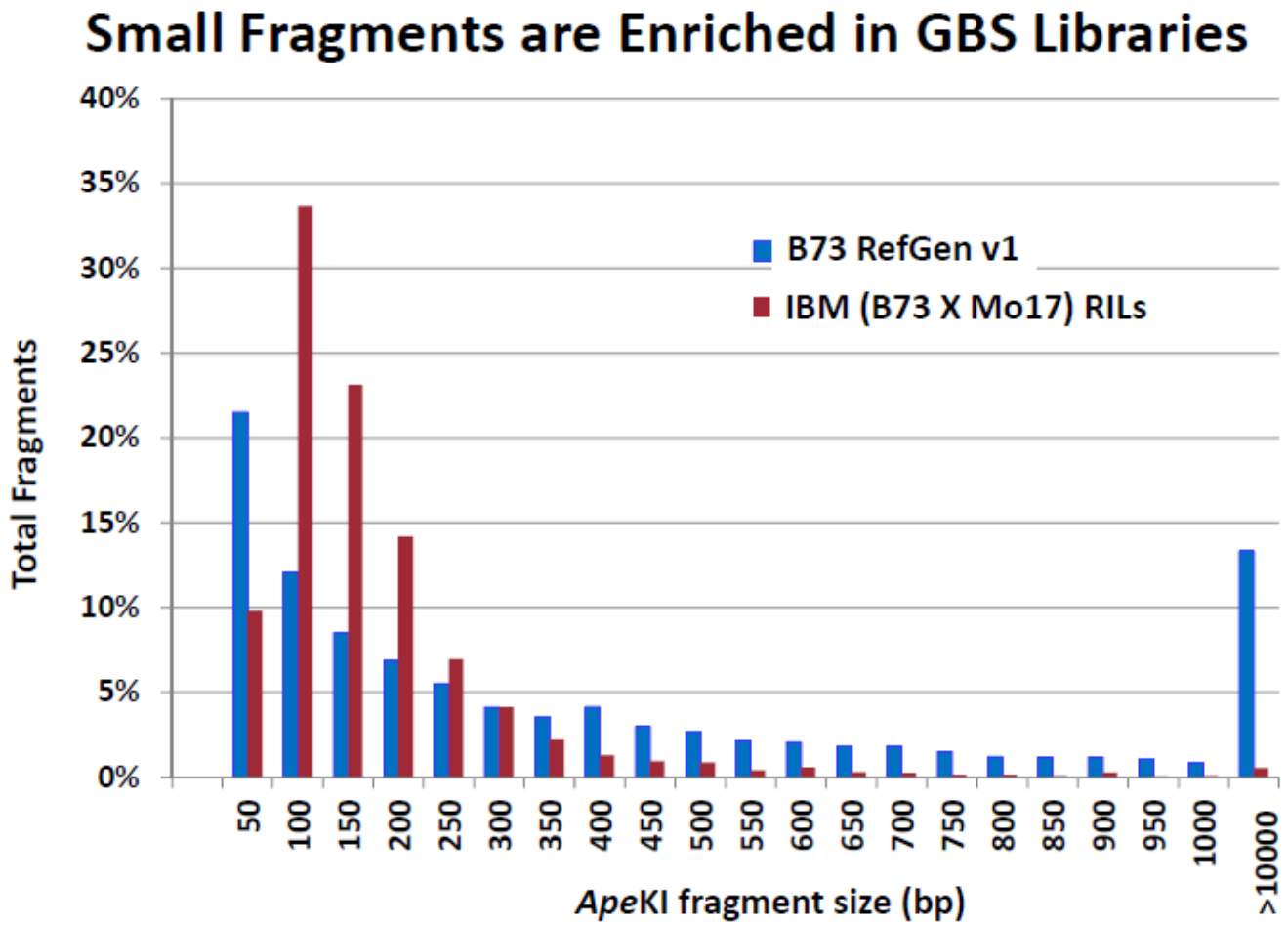
Restriction Enzymes

*Ape*KI	5'G CWGC 3'
*Pst*I	CTGCA G
*Eco*T22I	ATGCA T

Perform Titration to Minimize Adapter Dimers Before Sequencing

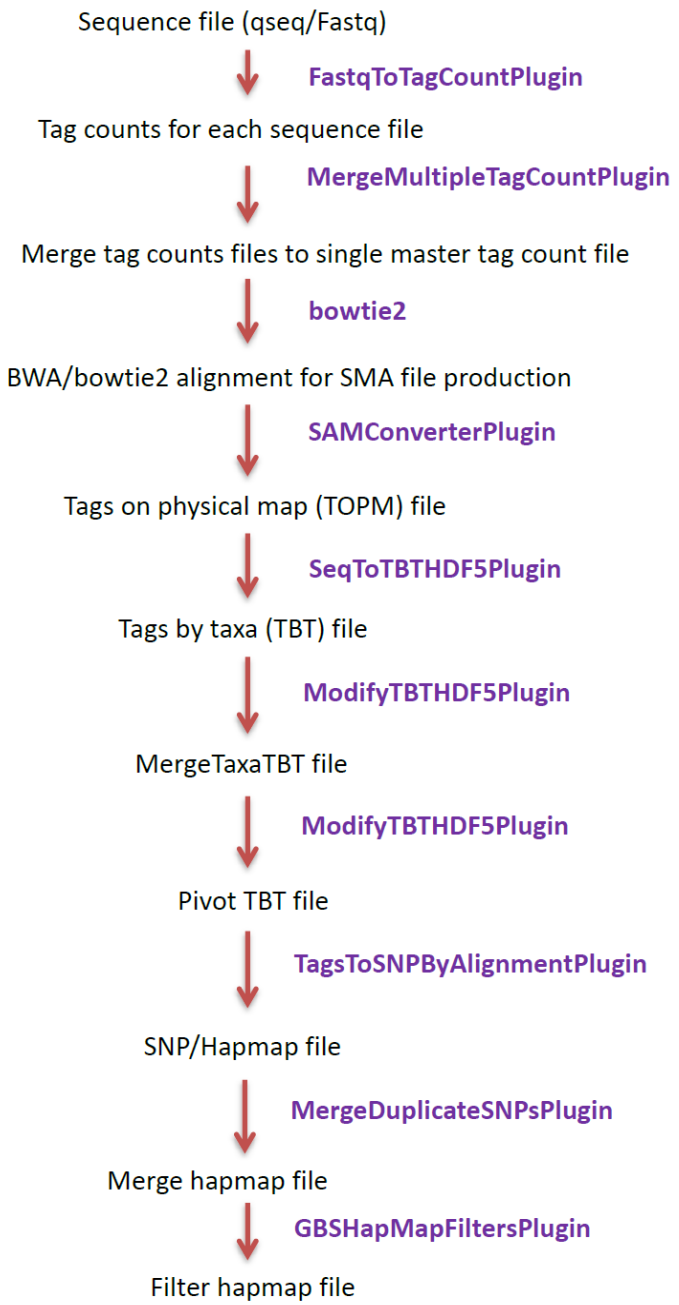
NOTE: Done once with a small number of samples.
Adapter dimers constitute only 0.05% of raw sequence reads




Small Fragments are Enriched in GBS Libraries
40%
35%
30%
25%
20%
15%
10%
5%
0%
Total Fragments
B73 RefGen v1
IBM (B73 X Mo17) RILs
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750
800
850
900
950
1000
>10000
ApeKI fragment size (bp)

Bioinformatics (GBS Pipelines)

GBS Work flow:



GBS pipelines

TASSEL – tassel4.0standalone
http://www.maizegenetics.net/index.php?option=com_content&task=view&id=89&Itemid=119

User manual: Reference genome
http://www.maizegenetics.net/tassel/docs/TasselPipelineGBS.pdf (**Elshire et al. 2011**)

User manual: without reference genome
http://www.maizegenetics.net/images/stories/bioinformatics/TASSEL/uneak_pipeline_documentation.pdf (**Lu et al. 2013**)

For more information on GBS
http://www.maizegenetics.net/gbs-bioinformatics

Computationally intensive

JAVA environment
PERL + JAVA

GBS Vocabulary

Taxa — Individual sample

Key file — Text file containing

- Sample information
- Barcode
- Flowcell and lane number
- Sample prepID

Barcode — Unique DNA sequence associated with each taxa

Sequence file — Text file containing DNA sequences information from Illumina

- Qseq or Fastq file

Read — DNA Sequence produced in sequencing..

GBS Tag — DNA sequence starts with cut site remnant and having additional sequence without barcode

TagsByTaxa (TBT) — Matrix of GBS tags (row) with taxa (columns)

GBS pipelines

Types

Discovery pipeline

- Requires a reference genome
- Multiple steps to process the data into genotypes

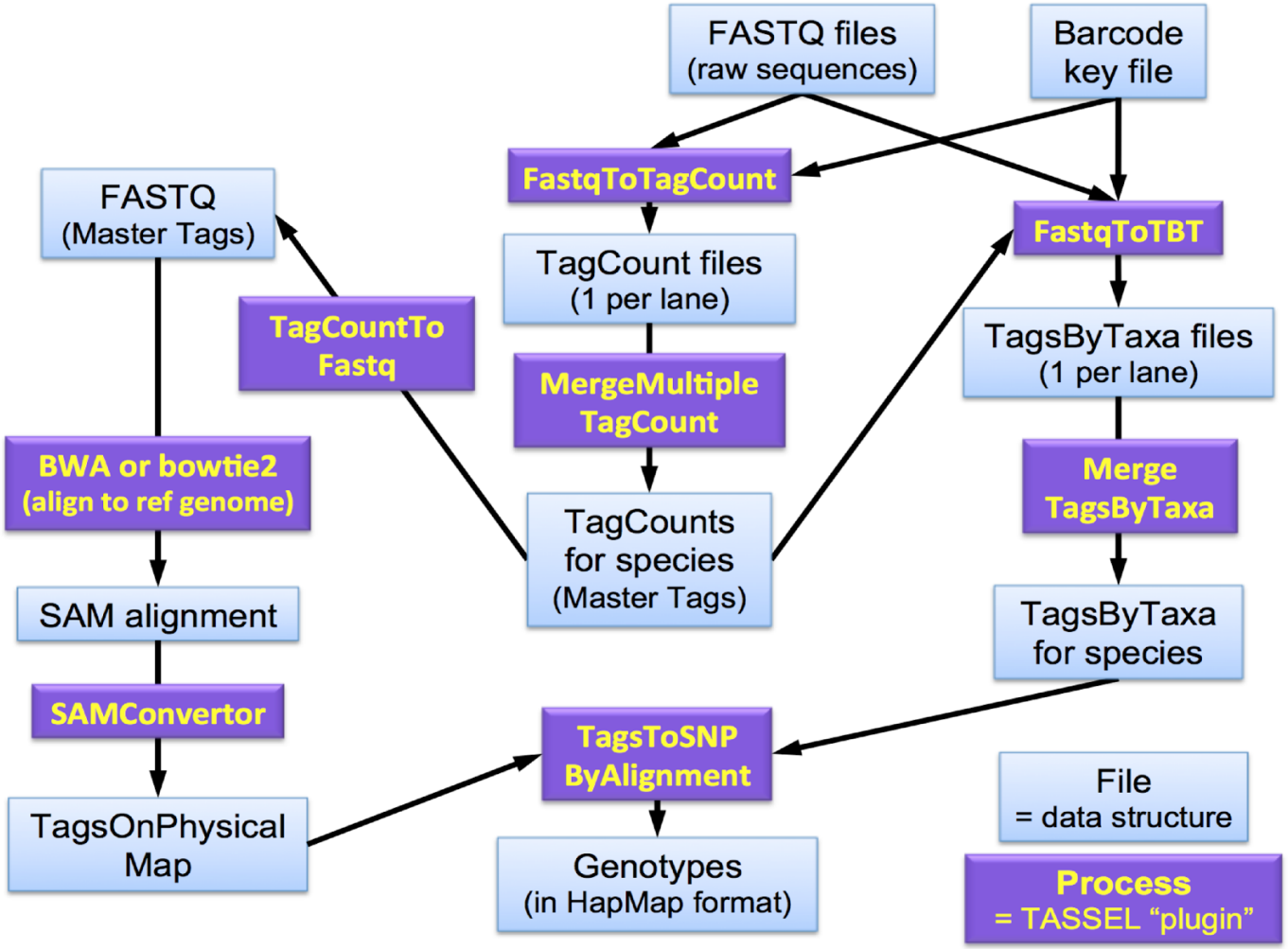
Production pipeline

- Uses info from discovery pipeline
- Only one step from sequence to genotypes

UNEAK pipeline

- Species without reference genome

Discovery GBS Pipeline



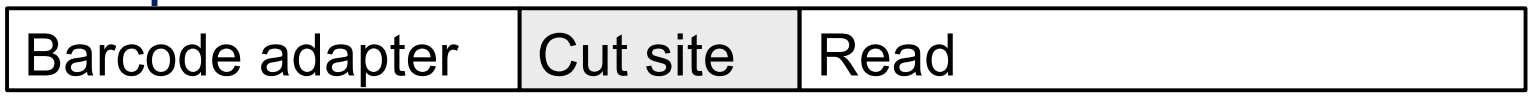
Discovery pipeline

Key file format

Flowcell	Lane	Barcode	DNA Sample	LibraryPlate	Row	Col	Library PrepID	Library PlateID	Enzyme	FullSampleName
628NVAAXX	5	TGCA	IS2787	ICRISAT_Plate_1A	A	2	250006891	450013296	ApeKI	IS2787:628NVAAXX:5:250006891
628NVAAXX	5	ACTA	IS3957	ICRISAT_Plate_1A	A	3	250006893	450013296	ApeKI	IS3957:628NVAAXX:5:250006893
628NVAAXX	5	GTCT	IS6193	ICRISAT_Plate_1A	A	4	250006895	450013296	ApeKI	IS6193:628NVAAXX:5:250006895
628NVAAXX	5	GAAT	IS9303	ICRISAT_Plate_1A	A	5	250006897	450013296	ApeKI	IS9303:628NVAAXX:5:250006897
628NVAAXX	5	GCGT	IS11119	ICRISAT_Plate_1A	A	6	250006899	450013296	ApeKI	IS11119:628NVAAXX:5:250006899
628NVAAXX	5	TGGC	IS13845	ICRISAT_Plate_1A	A	7	250006901	450013296	ApeKI	IS13845:628NVAAXX:5:250006901
628NVAAXX	5	CGAT	IS15752	ICRISAT_Plate_1A	A	8	250006903	450013296	ApeKI	IS15752:628NVAAXX:5:250006903
628NVAAXX	5	CTTGA	IS19132	ICRISAT_Plate_1A	A	9	250006905	450013296	ApeKI	IS19132:628NVAAXX:5:250006905
628NVAAXX	5	TCACC	IS20351	ICRISAT_Plate_1A	A	10	250006907	450013296	ApeKI	IS20351:628NVAAXX:5:250006907
628NVAAXX	5	CTAGC	IS22330	ICRISAT_Plate_1A	A	11	250006909	450013296	ApeKI	IS22330:628NVAAXX:5:250006909
628NVAAXX	5	ACAAA	IS23669	ICRISAT_Plate_1A	A	12	250006911	450013296	ApeKI	IS23669:628NVAAXX:5:250006911
628NVAAXX	5	TTCTC	IS303	ICRISAT_Plate_1A	B	1	250006913	450013296	ApeKI	IS303:628NVAAXX:5:250006913
628NVAAXX	5	AGCCC	IS2807	ICRISAT_Plate_1A	B	2	250006915	450013296	ApeKI	IS2807:628NVAAXX:5:250006915
628NVAAXX	5	GTATT	IS3971	ICRISAT_Plate_1A	B	3	250006917	450013296	ApeKI	IS3971:628NVAAXX:5:250006917
628NVAAXX	5	CTGTA	IS6351	ICRISAT_Plate_1A	B	4	250006919	450013296	ApeKI	IS6351:628NVAAXX:5:250006919
628NVAAXX	5	AGCAT	IS9468	ICRISAT_Plate_1A	B	5	250006921	450013296	ApeKI	IS9468:628NVAAXX:5:250006921
628NVAAXX	5	ACTAT	IS12169	ICRISAT_Plate_1A	B	6	250006923	450013296	ApeKI	IS12169:628NVAAXX:5:250006923
628NVAAXX	5	GAGAAT	IS13926	ICRISAT_Plate_1A	B	7	250006925	450013296	ApeKI	IS13926:628NVAAXX:5:250006925
628NVAAXX	5	CCAGCT	IS16044	ICRISAT_Plate_1A	B	8	250006927	450013296	ApeKI	IS16044:628NVAAXX:5:250006927

GBS Restriction Fragment Structure

Accepted read



ATGCCTCTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTG

Rejected or Trimmed reads

Potential chimeric sequence

Short sequence

Adapter dimer

Raw sequence data

Read

ATGCCTCTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTGAGATCGGAAGAGCGGT

↓

GBS pipelines

↓

ATGCCTCTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTGAGATCGGAAGAGCGGT
CGATCAGCAGTTGACTGGACATCTAGGGGCAAAGCACTGTTTCGGTGCGGGCTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA
CCGGATATCAGCAGGCCGTGGTACATGTAATGGAGCATGGATTTGAAGGTGGATGCCTTCATGTCCTGGACGGCGATCGAGCTCGC

↓

TagsCounts

CTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTG
CAGCAGTTGACTGGACATCTAGGGGCAAAGCACTGTTTCGGTGCGGGCTGAAAAAAAAAA
CAGCAGGCCGTGGTACATGTAATGGAGCATGGATTTGAAGGTGGATGCCTGCAAAAAAAAA

Tag count > 5/10

↓

MergeMultipleTagCountPlugin

Merge Tags counts

MasterTagCounts

Total # Tags

1453006 2	Tags size	
CTGCCCTCGTTTAAGTCCGATTGCTGATAG[illegible]GTTTCTTACCAAAATCCACATGGCTG	64	409
CAGCAAAAAAAAAAAGGACATGGGTATCCGGTAACACTGCAGAAAATTTGAGTAAAAGGCGGTA	64	314
CAGCAAAAAAAAAAATCTCAGAAAATGCAGTGCAGAGTGTATATTTTGCTTGATCCTGGCTCAC	64	235
CAGCAAAAAAAAAACAAAAAAACAAAACAAAAGGATTTCTTTTAGGGAAAAAACATAAGGTGCG	64	338
CAGCAAAAAAAAAACAAATTGTCAATTATTACAATCCCAAAGGCGAATGACGAAACTAAAATCT	64	
CAGCAAAAAAAAAACAAGACAGGAAGAACGGTGGTGGAGTTGAGTAGAGAGAGGCCAGCAAAGC	64	
CAGCAAAAAAAAAACCCTCTGCTCTGAAGGTGGAAGGTAGGATAATGCACCTGGATATCAGAAA	64	281
CAGCTTACCCTTGTTTGGTTGCCCGCATACGCCTTTATGCAAGTTTGCACAGTGCTTAAAGCAG	64	171
CTGCAACAACAAGCTCGTCGGCGCCAAGTTCTTCGGCCTGGGGTACGAGGCCGCGCACGGCGGG	64	290
CTGCAACAACGTGTCCTTCGCGCCATATGGCGGGAACTGGCGCCGGGGCAAGAAGATCGCGGTG	64	382
CTGCAACAAGACAGAATTTGTTAAAATGTAATATGAAACAGACACATCAAGTTAAATTCTGGTA	64	239
CTGCAACAAGCAGAAGGTCCGGCGGGGCCTGTGGTCGCCGGAGGAGGACGAGAAGCTCATCAAG	64	197
CTGCAACAAGGAAATTGCAGGCGCGCGCGATACAATGTACCAATACAACGAAGATGCTCTCTGC	64	125
CTGCAACAAGTTTTTTGCGTGGGCAAGACTTCCATGCTTCTAGAGCCTGATCTCTCAAAATCTC	64	141

Tags

Reads

MergeTagsByTaxa

Tags

CTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTG
CAGCAGTTGACTGGACATCTAGGGGCAAAGCACTGTTTCGGTGCGGGCTGAGATCGGAAGAGCG
CAGCAGGCCGTGGTACATGTAATGGAGCATGGATTTGAAGGTGGATGCCTTCATGTCCTGGACG

ATGCCTCTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTGAGATCGGAAGAGCGGT
GCGTCTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTGAGATCGGAAGAGCGGT
CGATCAGCAGTTGACTGGACATCTAGGGGCAAAGCACTGTTTCGGTGCGGGCTGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGA
CCGGATATCAGCAGGCCGTGGTACATGTAATGGAGCATGGATTTGAAGGTGGATGCCTTCATGTCCTGGACGGCGATCGAGCTCGC

SeqToTBTHDF5Plugin

ATGCCT	BTx623
CGAT	N13
CCGGATAT	E36-1
CAGA	ICSL1118
AACT	ICSL1126
GCGT	ICSL1134

Qseq file

Tags file

key file

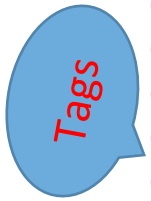
TagsByTaxa (TBT)

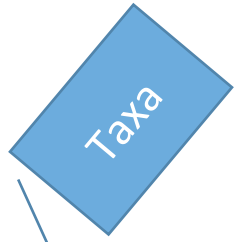
(Presence/Absence matrix)

MergeTBTFiles

Merge TBT files

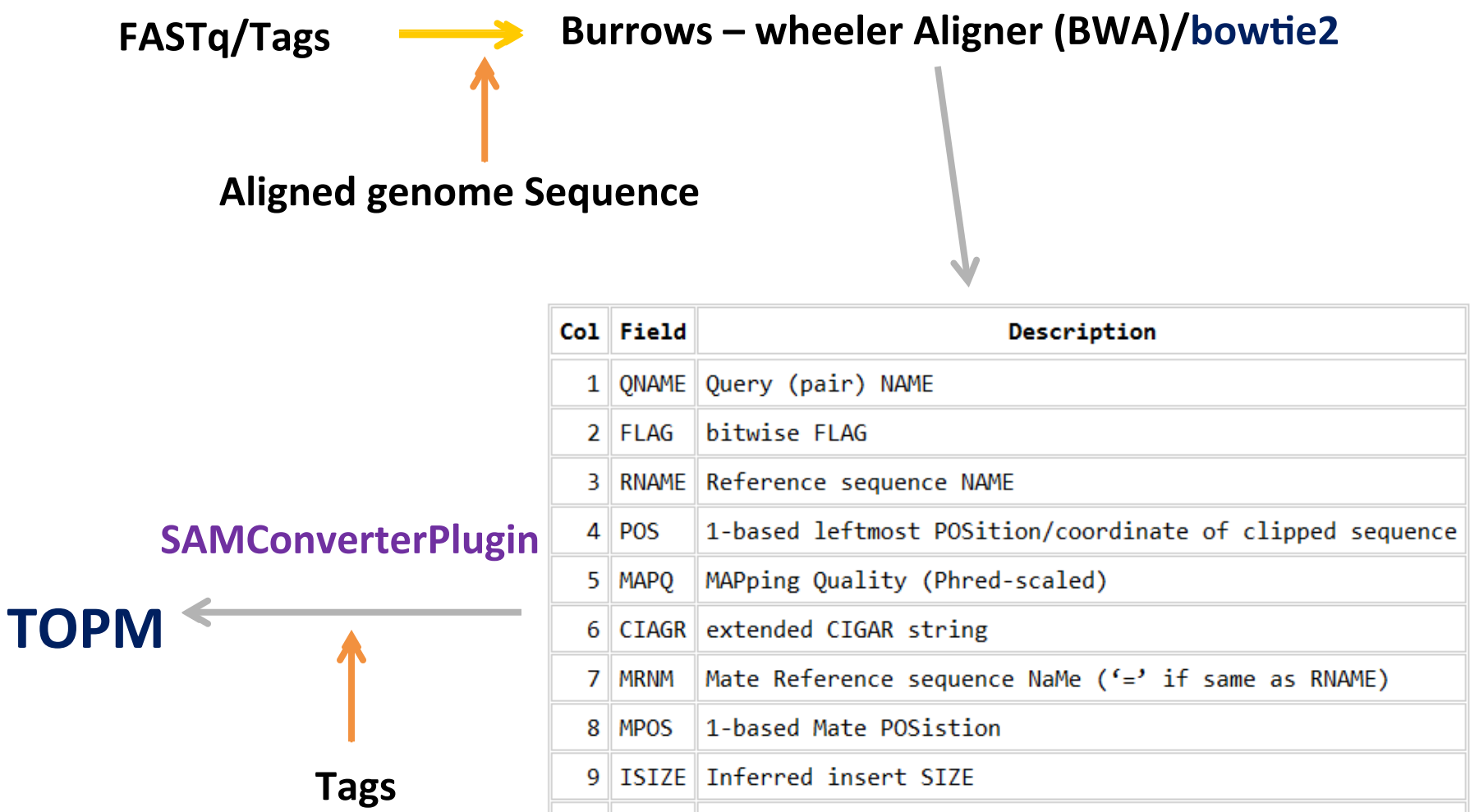
TBT file





		BTx623	N13	E36-1	ICSL1118	ICSL1126	ICSL1134
CTGCCCTCGTTTAAGTCCGATTGCTGATAGTTATCTTTGTTTCTTACCAAAATCCACATGGCTG	64	1	0	0	0	0	1
CAGCAAAAAAAAAAAGGACATGGGTATCCGGTAACACTGCAGAAAATTTGAGTAAAAGGCGGTA	64	0	0	0	0	0	0
CAGCAAAAAAAAAAATCTCAGAAAATGCAGTGCAGAGTGTATATTTTGCTTGATCCTGGCTCAC	64	0	0	1	0	0	0
CAGCAAAAAAAAAACAAAAAAACAAAACAAAAGGATTTCTTTTAGGGAAAAAACATAAGGTGCG	64	1	1	0	1	0	0
CAGCAAAAAAAAAACAAATTGTCAATTATTACAATCCCAAAGGCGAATGACGAAACTAAAATCT	64	0	0	0	0	0	0
CAGCAAAAAAAAAACAAGACAGGAAGAACGGTGGTGGAGTTGAGTAGAGAGAGGCCAGCAAAGC	64	1	0	0	0	0	1
CAGCAAAAAAAAAACCCTCTGCTCTGAAGGTGGAAGGTAGGATAATGCACCTGGATATCAGAAA	64	0	0	0	0	0	0
CAGCTTACCCTTGTTTGGTTGCCCGCATACGCCTTTATGCAAGTTTGCACAGTGCTTAAAGCAG	64	0	0	0	1	0	0
CTGCAACAACAAGCTCGTCGGCGCCAAGTTCTTCGGCCTGGGGTACGAGGCCGCGCACGGCGGG	64	0	0	0	0	0	1
CTGCAACAACGTGTCCTTCGCGCCATATGGCGGGAACTGGCGCCGGGGCAAGAAGATCGCGGTG	64	0	0	0	0	0	0
CTGCAACAAGACAGAATTTGTTAAAATGTAATATGAAACAGACACATCAAGTTAAATTCTGGTA	64	0	0	0	0	0	0
CTGCAACAAGCAGAAGGTCCGGCGGGGCCTGTGGTCGCCGGAGGAGGACGAGAAGCTCATCAAG	64	1	1	1	0	0	1
CTGCAACAAGGAAATTGCAGGCGCGCGCGATACAATGTACCAATACAACGAAGATGCTCTCTGC	64	0	0	0	0	0	0
CTGCAACAAGTTTTTTGCGTGGGCAAGACTTCCATGCTTCTAGAGCCTGATCTCTCAAAATCTC	64	0	0	0	0	0	0

TagsOnPhysicalMap (TOPM)



Col	Field	Description
1	QNAME	Query (pair) NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSistion
9	ISIZE	Inferred insert SIZE
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

TOPM file

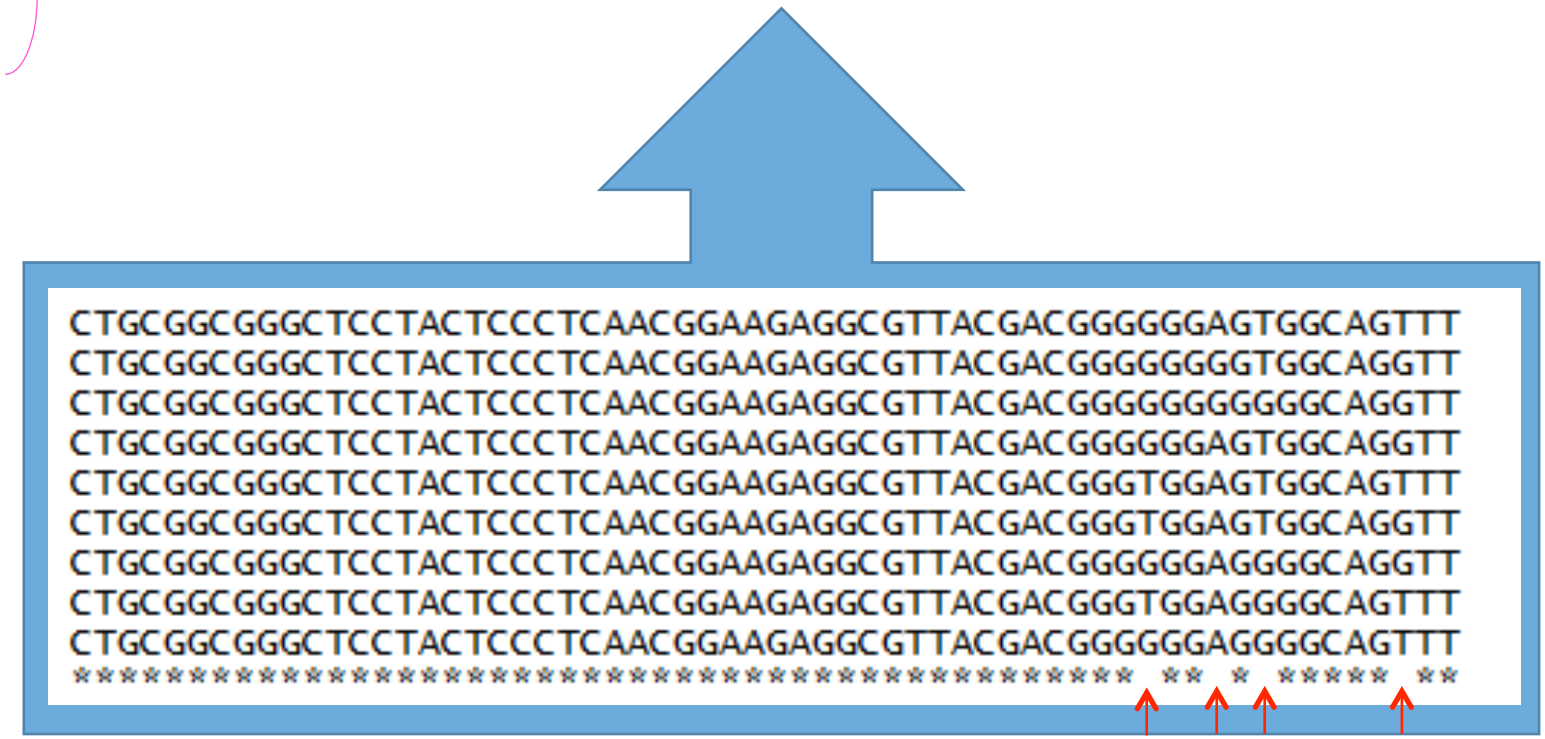
Tag	length	chr	start	end
CACCCCCGCCAAGTGTTTCTGTGTTGAGGAGTCAACAAGAACCAATTGCAGGATGTAAAAGCAG	64	1	12982	13046
CTGCTGGTTAATACGTACTCCAGTTGGTTACCATGCACCAAGCAGGGCAAGGCACTCCTGCTCC	64	1	2132	2196
CTGCTGGTTAATACGTACTCCAGTTGGTTACTATGCACCAAGCAGGGCAAGGCACTCCTGCTCC	64	1	2132	2196
CAGCCCTACGGCCGCAACTGCAAAAATCAGATGTGAGCATAGTTCTCTAATGTATCATAAAAGT	64	2	2266316	2266380
CTGCTACCTCTGCCCTGCGTGCGTGACCCATACATACGGTGGGGCCGGGGTGTAAAATCTTGTG	64	2	2926106	2926170
CAGCGGCTGAACACTGACCAGGTTGACGGCGTCGCGCGCGGCGAGGGCGACGATGTCGGCGCAG	64	3	2129422	2129486
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGTGGAGGGGCAGTTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGGGGAGGGGCAGGTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGGGGAGGGGCAGTTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGGGGAGTGGCAGGTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGGGGAGTGGCAGTTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGGGGGGGGGCAGGTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGGGGGGGTGGCAGGTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGTGGAGGGGCAGTTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGTGGAGTGGCAGGTTCTCT	64	3	2140833	2140897
CTGCGGCGGGCTCCTACTCCCTCAACGGAAGAGGCGTTACGACGGGTGGAGTGGCAGTTTCTCT	64	3	2140833	2140897
CAGCTGTCCCTTGTAACCTGCGGACATGACCGAACATCCAAGTACTCGAGAGCCTTGAAGGCTT	64	4	56916960	56917024
CTGCTTGTTGTTGTTGATTGCTCGTTGGCCGATCGGTGGTCGTTGTCGTCGTTGAGCTGGTAGC	64	4	56942951	56943015
CAGCGCCCAAACCTTCTTTCCCGCTCTTGTTGATTCCTCAGTCTCTCCAGTTCTTGGAACATCG	64	5	54508281	54508345
CTGCCTCAGCACAGCCGGTGCTCTTGGATCCTCGAGCACCTGCACTCTTTCAGGAGGCCAAGGA	64	5	13747204	13747268
CAGCACGCGTATCCTTGGCCCTTTATTGTTTGAAAATGACCTGTGACGGTGCCTTCTGCGTAGT	64	6	52338128	52338192
CTGCGGCGGCGGCATGACCGACGGGCGGCGTGGGTGCCGCAGTGCCGGCAACTTCCGCCGTTCT	64	6	58276481	58276545
CTGCCACAGACGGAGTTACGACAACCTTGCAAAACTCCGCCCGGCTTAAGTTAATTTAATCAGG	64	7	18492751	18492815
CAGCCAAAAAAATGGCAAAAGGTGTAGGACTTGGGTACGTCAAAATTCATTCATGTGAGAATGA	64	7	19393217	19393281

Hapmap file

Key file
Qseq file
Tags file
TOPM file
TBT file

TagsToSNPByAlignmentPlugin

Hapmap file
(Sequence alignment file)



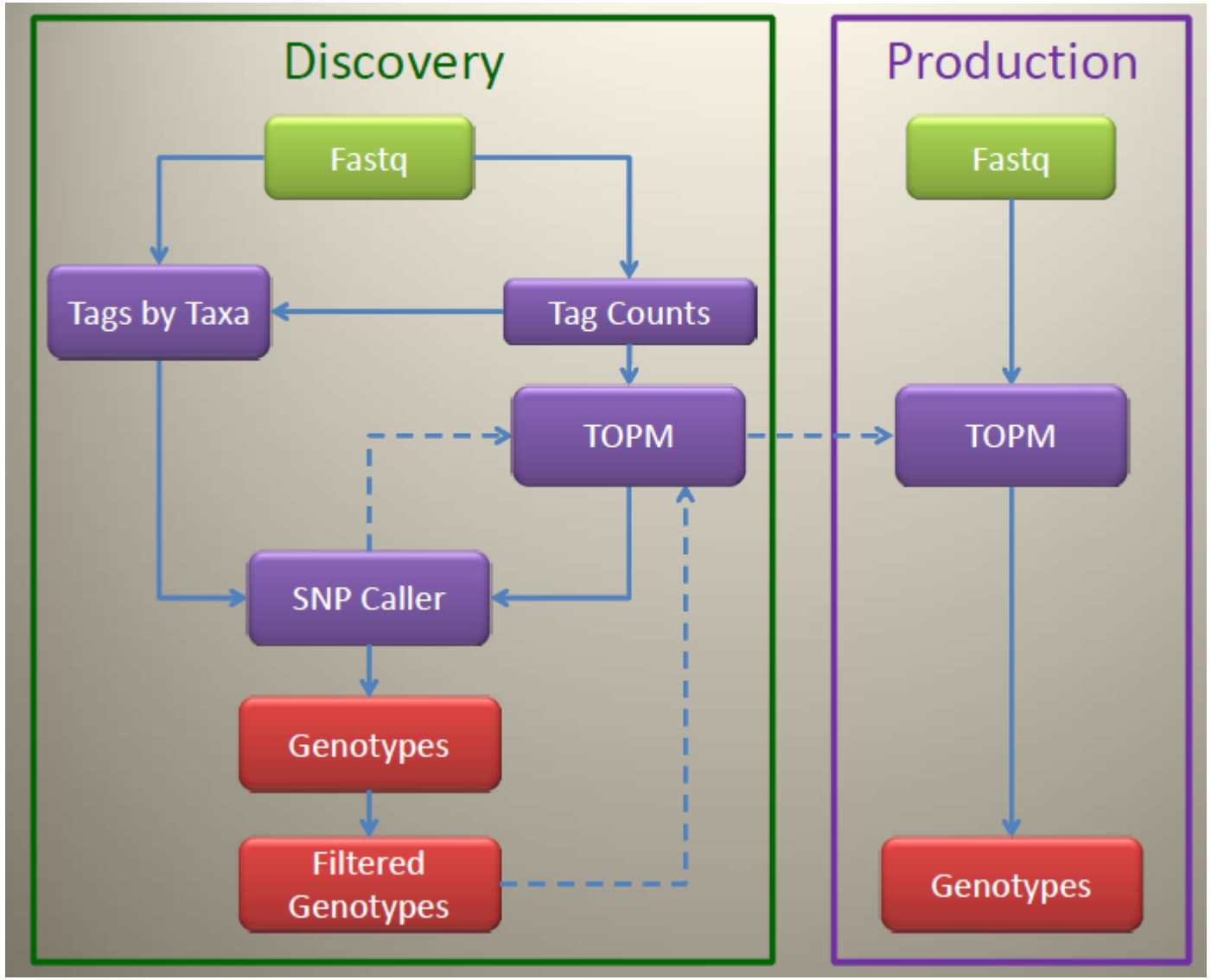
Why production pipeline?

- Discovery pipeline takes more time

  Eg: Last maize build took > 3 months

- Most common alleles have been identified in discovery pipelines
- Use of available information from discovery pipelines to call SNPs in new run

GBS Production pipeline



UNEAK pipeline

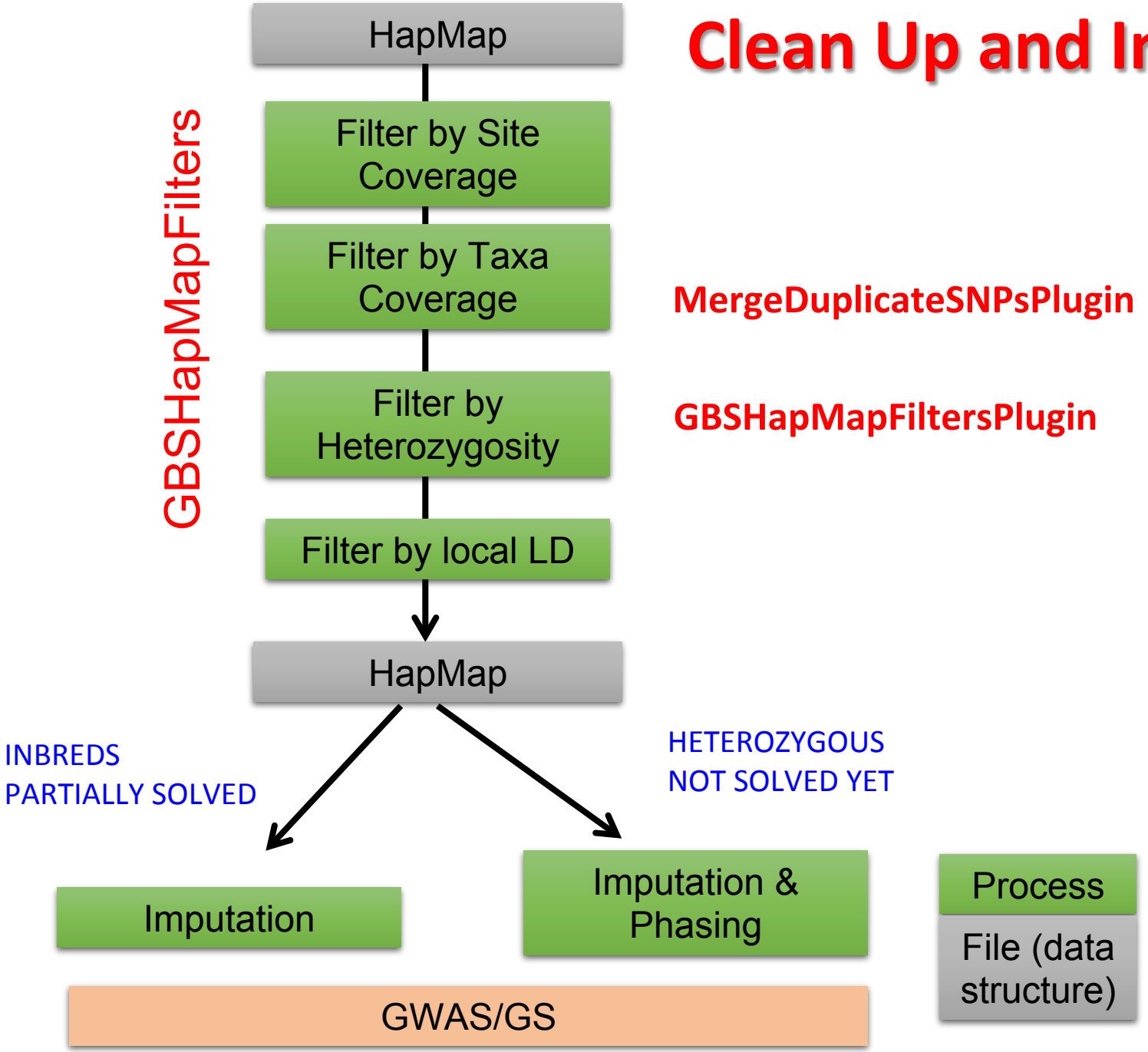
(Universal Network Enabled Analysis Kit)
A reference free SNP calling pipeline

Designed for species that…..

- lack a reference genome
- diploid or polyploid
- inbreeders or outcrosses
- having limited genetic or genomic resources

Lu et al. 2013; PLoS Genetics

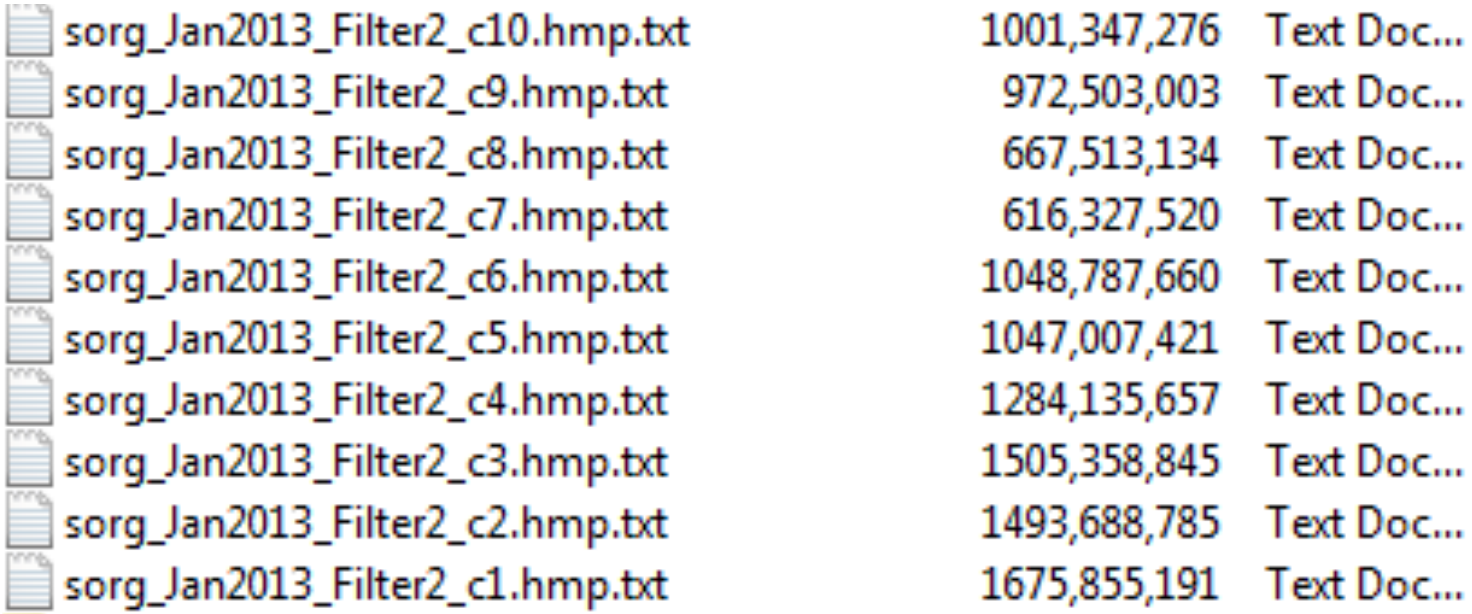
Clean Up and Imputation



Hapmap file

rs#	alleles	chrom	pos	ICSB617	ICSB212	ICSB536	ICSB79	ICSB538	ICSB639	ICSB417	ICSB748	ICSB563
S9_20491	G/A	9	20491	G	N	N	N	N	N	N	N	N
S9_23432	C/Y	9	23432	N	N	Y	C	N	N	C	C	N
S9_23452	T/K	9	23452	N	N	T	T	N	N	T	T	N
S9_26213	G/A	9	26213	N	N	N	W	N	N	N	G	N
S9_27701	T/C	9	27701	T	N	T	N	N	Y	C	T	N
S9_27720	T/K	9	27720	T	N	T	N	N	T	T	T	N
S9_30044	C/T	9	30044	C	N	N	N	N	N	N	N	N
S9_36986	A/G	9	36986	N	N	A	N	N	N	N	N	N
S9_39902	C/T	9	39902	N	N	N	N	T	N	T	N	T
S9_39912	A/G	9	39912	N	N	N	N	G	N	G	N	G
S9_40450	A/C	9	40450	N	K	N	N	N	N	N	N	N
S9_40452	G/T	9	40452	N	N	N	N	N	N	N	N	N
S9_40459	A/G	9	40459	N	N	N	N	N	N	N	N	N
S9_43958	T/C	9	43958	N	N	N	Y	N	M	N	N	N
S9_43982	C/T	9	43982	N	N	N	N	N	N	N	N	N
S9_44191	G/C	9	44191	N	M	N	N	N	N	N	N	N
S9_45394	T/A	9	45394	N	N	N	N	N	N	N	N	N
S9_56733	C/M	9	56733	N	N	N	N	N	N	C	N	N
S9_56734	C/S	9	56734	N	N	N	N	N	N	C	N	N
S9_56735	T/W	9	56735	N	N	N	N	N	N	T	N	N
S9_69001	C/Y	9	69001	C	N	N	N	N	C	N	C	N
S9_72163	C/T	9	72163	N	N	N	N	N	N	T	N	N
S9_74270	C/T	9	74270	C	N	T	N	N	N	C	N	N
S9_76335	A/G	9	76335	N	A	N	N	N	N	R	A	N
S9_76341	T/G	9	76341	N	T	N	N	N	N	K	T	N
S9_78501	A/C	9	78501	N	A	N	N	N	N	N	N	N

Hapmap Files



sorg_Jan2013_Filter2_c10.hmp.txt	1001,347,276	Text Doc...
sorg_Jan2013_Filter2_c9.hmp.txt	972,503,003	Text Doc...
sorg_Jan2013_Filter2_c8.hmp.txt	667,513,134	Text Doc...
sorg_Jan2013_Filter2_c7.hmp.txt	616,327,520	Text Doc...
sorg_Jan2013_Filter2_c6.hmp.txt	1048,787,660	Text Doc...
sorg_Jan2013_Filter2_c5.hmp.txt	1047,007,421	Text Doc...
sorg_Jan2013_Filter2_c4.hmp.txt	1284,135,657	Text Doc...
sorg_Jan2013_Filter2_c3.hmp.txt	1505,358,845	Text Doc...
sorg_Jan2013_Filter2_c2.hmp.txt	1493,688,785	Text Doc...
sorg_Jan2013_Filter2_c1.hmp.txt	1675,855,191	Text Doc...

Chromosome	Unfilter	Filter2 (mnMAF 0.01, mnTCov 0.01, mnSCov 0.01, mnF 0.8)
1	1288506	83460
2	999490	74192
3	1077534	74976
4	868854	64199
5	528132	52401
6	706426	51633
7	569982	28153
8	462068	30221
9	618291	48143
10	659650	49595
Total	**7778933**	**556973**

Optimizing GBS in New Species

Literature

Discovery pipeline

UNEAK pipeline

GBS applications

OPEN ACCESS Freely available online

PLoS one

A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species

Robert J. Elshire[1], Jeffrey C. Glaubitz[1], Qi Sun[2], Jesse A. Poland[3], Ken Kawamoto[1], Edward S. Buckler[1,4], Sharon E. Mitchell[1]*

1 Institute for Genomic Diversity, Cornell University, Ithaca, New York, United States of America, **2** Computational Biology Service Unit, Cornell University, Ithaca, New York, United States of America, **3** Hard Winter Wheat Genetics Research Unit, United States Department of Agriculture/Agricultural Research Service, Manhattan, Kansas, United States of America, **4** Plant, Soil and Nutrition Research Unit, United States Department of Agriculture/Agricultural Research Service, Ithaca, New York, United States of America

OPEN ACCESS Freely available online

PLOS | GENETICS

Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol

Fei Lu[1], Alexander E. Lipka[1,2], Jeff Glaubitz[1], Rob Elshire[1], Jerome H. Cherney[3], Michael D. Casler[4,5], Edward S. Buckler[1,2], Denise E. Costich[1,2¤*]

1 Institute for Genomic Diversity, Cornell University, Ithaca, New York, United States of America, **2** Agricultural Research Service, United States Department of Agriculture, Ithaca, New York, United States of America, **3** Department of Crop and Soil Sciences, Cornell University, Ithaca, New York, United States of America, **4** Agricultural Research Service, United States Department of Agriculture, Madison, Wisconsin, United States of America, **5** Department of Agronomy, University of Wisconsin–Madison, Madison, Wisconsin, United States of America

Population genomic and genome-wide association studies of agroclimatic traits in sorghum

Geoffrey P. Morris[a,1,2], Punna Ramu[b,1], Santosh P. Deshpande[b], C. Thomas Hash[c], Trushar Shah[b], Hari D. Upadhyaya[b], Oscar Riera-Lizarazu[b], Patrick J. Brown[d], Charlotte B. Acharya[e], Sharon E. Mitchell[e], James Harriman[e], Jeffrey C. Glaubitz[e], Edward S. Buckler[e,f,g], and Stephen Kresovich[a]

[a]Department of Biological Sciences, University of South Carolina, Columbia, SC 29208; [b]International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad 502 324, Andhra Pradesh, India; [c]ICRISAT-Sadoré, BP 12404 Niamey, Niger; [d]Department of Crop Sciences, University of Illinois, Urbana, IL 61801; [e]Institute for Genomic Diversity and [f]Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853; and [g]Agricultural Research Service, Department of Agriculture, Ithaca, NY 14853

Edited by Ronald L. Phillips, University of Minnesota, St. Paul, MN, and approved November 21, 2012 (received for review September 14, 2012)

Genome Wide Association Study (GWAS)

Association Mapping (AM)

(or)

Linkage Disequilibrium mapping (LD mapping)

(or)

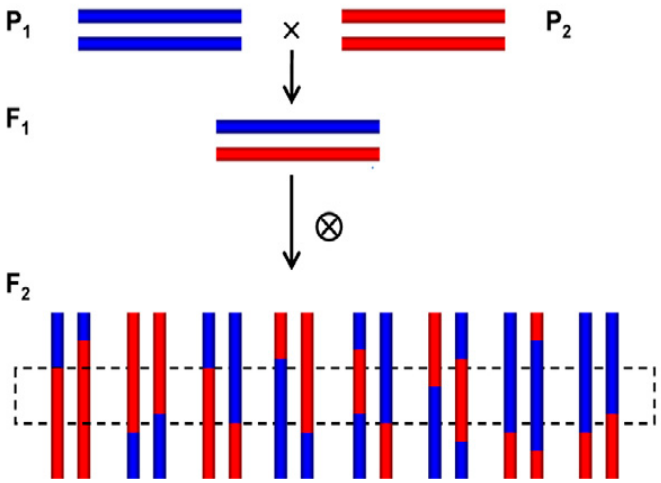
Population mapping

Linkage mapping (or) Family mapping

Generation of mapping population (RILs, NILs, DH, BC, F2)

Genotyping – polymorphic markers

Phenotyping – trait of interest



Limitations

Resolution power is low (10–30 cM)

Small population size

Modest degree of recombination within the population

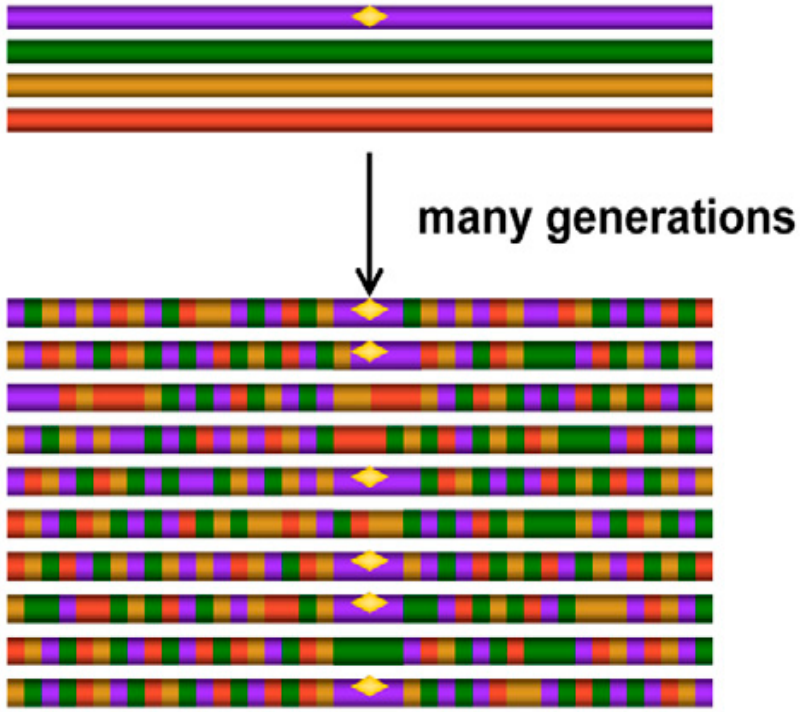
Linkage mapping – limited to sampling only two alleles at a given locus in any given bi-parental population

ASSOCIATION MAPPING

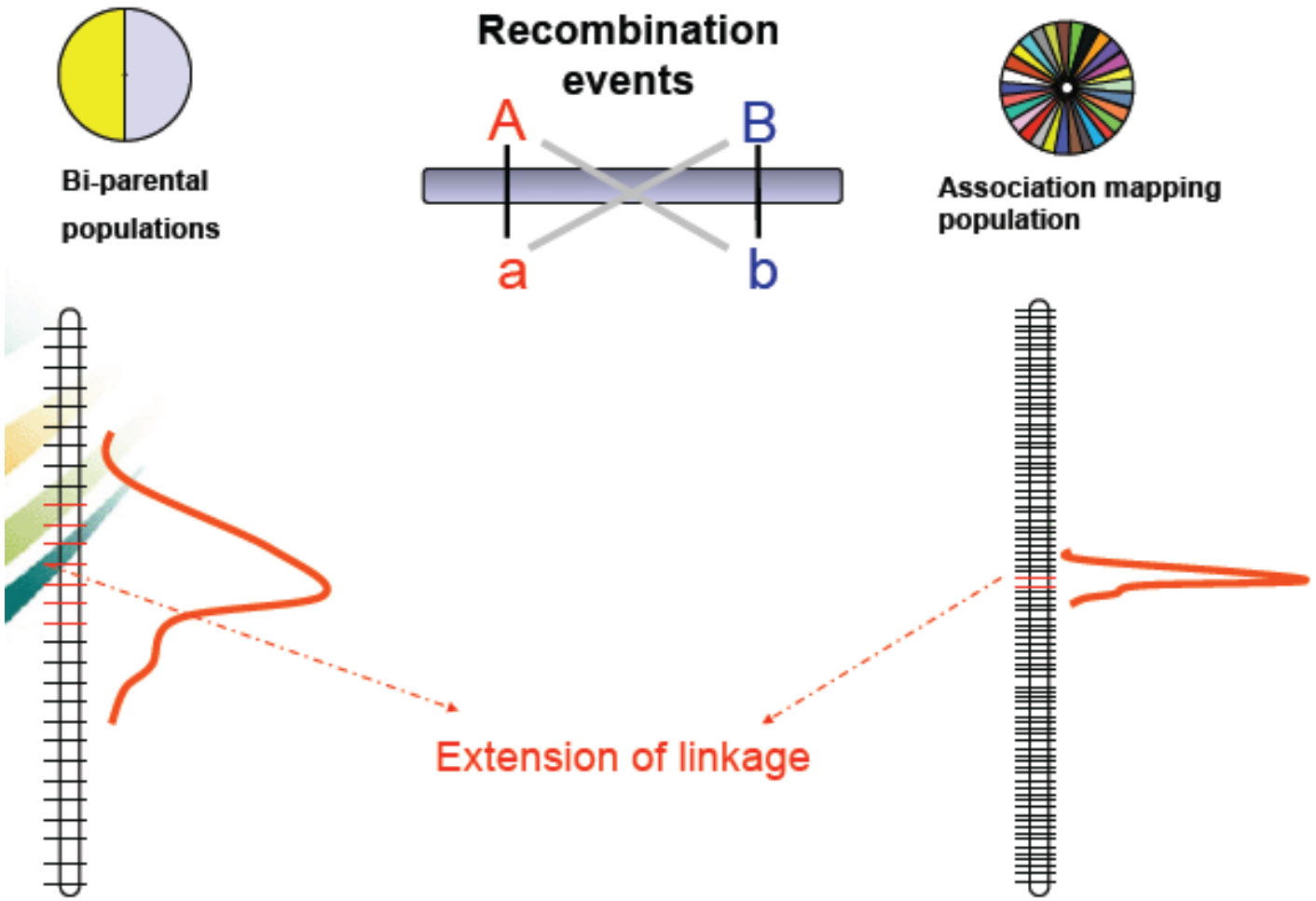
- Currently existing natural populations are used **Vs** generating a population via a biparental cross
- No need to develop mapping population
- A potentially large number of alleles per locus – as opposed to only two – can be surveyed simultaneously
- Resolution can be dramatically increased (e.g. 2000 bp in diverse maize inbred lines) - - - Fine mapping.
- Reduces time
- Considering recombination of history/evolution

AM is a multi-disciplinary field

- **Genomics**
- **Genetics**
- **Molecular Biology**
- **Statistical Genetics**
- **Bioinformatics**



Recombination: key of Genetic variation or Success of Breeding



GWAS Types

Success of either methods depends on population size and degree of LD

1. **Genome wide scanning or AM**
   - Markers spanned across the genome
   - Moderate to extensive LD
2. **Candidate gene scanning or AM**
   - Sequencing only candidate gene
   - Low LD

AM in human and plants

HUMANS

AM is responsible for identification and cloning of

- **Cystic fibrosis gene**
- **Diastrophic displasia gene**
- **Alzheimer's disease**

PLANTS

First AM – Candidate gene analysis

Flowering time and **dwarf8 (d8) gene** in **Maize**

First AM – Genome scan is in

Sea beet (*Beta vulgaris* ssp. *maritine*)

GENOME-WIDE ASSOCIATION MAPPING (GWA)

Sps – **Self-fertile** **: Arabidopsis, rice**

Clonally propagated **: Switch grass, grape**

If LD is high, GWA is useful with low resolution mapping

Number of markers to screen determined by

- sample size,
- Extent of LD

E.g.:

Human – 70,000 markers

Arabidopsis – 2,000 markers

Diverse Maize Landraces – 750,000 markers

Elite Maize lines – 50,000 markers

Sorghum – 556,000 markers

CANDIDATE GENE APPROACH

Multi-disciplinary approach

- Mutagenesis
- Biochemical analysis
- Expression profiling
- Comparative genome mapping
- Bioinformatics
- Linkage mapping

Positional candidates

or

Candidate genes

How to start?

Basic resources for AM

- Germplasm choice
- Trait evaluation
- Identification of candidate polymorphism
- Estimation of population structure
- Statistical analysis

GERMPLASM CHOICE

- Encompass as much **phenotypic variation** as possible, and perhaps represent the breeding pool of a crop species
- **Genetic or phenotypic surveys** can be used to identify genotypically diverse subsets of the available germplasm in order to maximize the range of alleles sampled in the population
- In some species, **core sets of germplasm** have already been defined and characterized, and can be used to initiate preliminary association studies
- Germplasm should have maximum diversity of gene pool with more extensive recombination events

Community resources

Maize 300 maize inbred lines (Flint-Garcia et al., 2005)

NAM panel consists of 5000 RILs (Yu et al. 2008)

Sorghum 377 Association mapping panel (Casa et al. 2008)

107 Sorghum Diversity Research Set (SDRS) (Shehzad et al. 2009)

384 Reference set of sorghum (Billot and Ramu et al. 2013,)

Barley 3840 Diverse germplasm set

✓ Model based study found that more power is achieved by increasing the no of individuals rather than increasing the no of markers

Linkage Disequilibrium (LD)

- **Non-random association between alleles at different loci**
- **LD extends to a much longer distance in self-pollinated crops than in cross-pollinated species**
- **Genome-wide LD determines the mapping resolution and marker density for a genome scan**

Plant populations amenable for association studies can be classifiable into one of five groups

(i) ideal sample with subtle population structure and familial relatedness
(ii) multi-family sample
(iii) sample with population structure
(iv) **sample with both population structure and familial relationships, and**
(v) sample with severe population structure and familial relationships.

Due to local adaptation, selection, and breeding history in many plant species, many populations for association mapping would fall into category **four**.

Factors affecting LD

Mutation – Base for producing polymorphism

Recombination – weakens the intra-chromosomal LD

inter-chromosomal LD broken by **independent assortment**

Population size – Small pop – effect of genetic drift result in constant loss of rare allele combinations, thus LD increases

Population mating system – Self vs Out cross

LD decay more – out cross sps – because recombination leads to heterozygosity

Admixture – introduction of chromosomes of different ancestry and allele frequency

Gene flow between genetically distant populations

Selection – Bottle neck effect very few allele

- combinations are passed to next generations
- Locus specific alleles

LD Decay

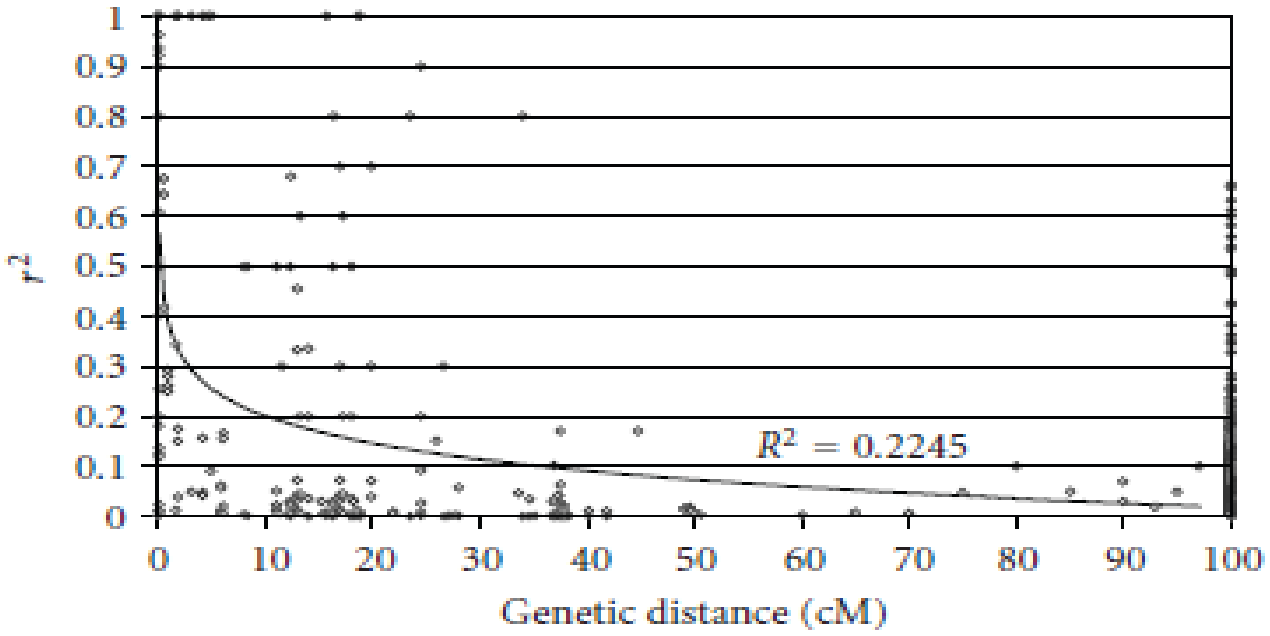


FIGURE 3: Linkage disequilibrium (LD) decay plot depicted from the LD values of a hypothetical marker data to demonstrate a measure of an average genome-wide LD block sizes. A pairwise LD values (r^2) are plotted against a genetic distance. Inner fitted trend line is a nonlinear logarithmic regression curve of r^2 on genetic distance. LD decay is considered below $r^2 = 0.1$ threshold and based on trend line it is around 38–40 cM in above plot. A pairwise LD between unlinked marker loci is assigned to 100 cM distance point. Note: this is for demonstration purposes only and does not have any real impact or correspond to any genomic fragment of an organism.

2. Genotyping

CANDIDATE GENE MARKERS

- May be from related or unrelated
- Comparative genomics and Bioinformatics
- SNPs are preferred markers system in candidate gene approach
- Promoters, Introns, Exons and 5' / 3' UTRs – targets for identification of candidate gene SNPs
- Rate of LD decay – denotes the no of SNPs per unit length
- It is not essential to screen all SNPs. SNPs that cause phenotypic variation due to alter in protien function (Coding SNPs) or gene expression (regulatory SNPs) should be a top priority
- If there are block of several SNPs in significant LD, there is no need to do all SNPs. An alternatively, select and score a fraction of SNPs (Tag SNPs) that capture most of the haplotype blocks in a candidate gene region – more cost effective

WHOLE GENOME SCANNING

The extent of LD - dictates the number of markers

SNP marker – large number…..resolution is more

SSR markers – small number …..resolution is low

3. PHENOTYPING DATA

AM –

a relatively large number of diverse accessions phenotypic data

collection with adequate replications across multiple years

multiple locations is challenging

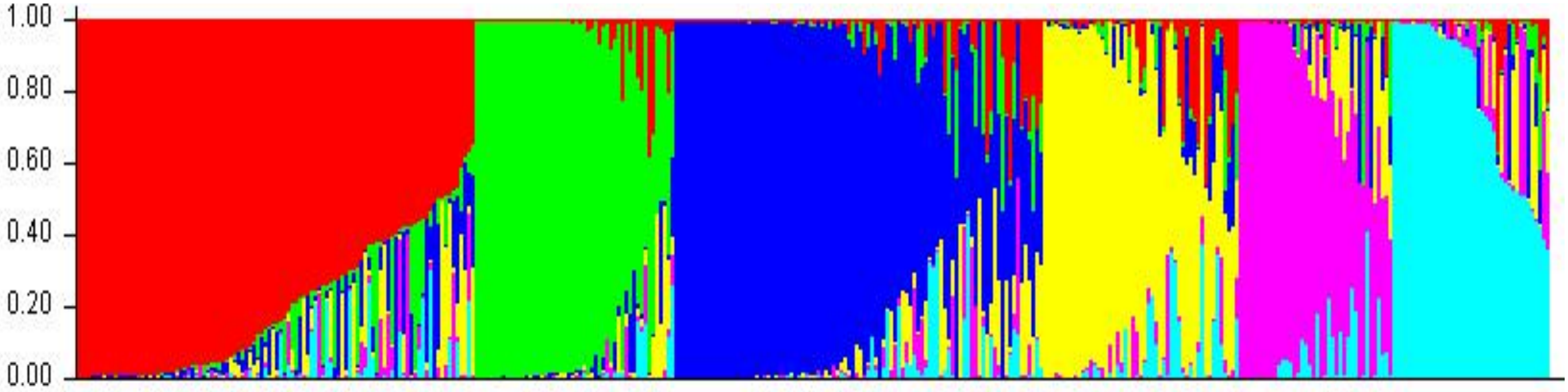
- Efficient field design with incomplete block design (e.g., α-lattice)
- Appropriate statistical methods (e.g., nearest neighbor analysis and spatial models)
- Consideration of QTL × environmental interaction should be explored to increase the mapping power, particularly if the field conditions are not homogenous (Eskridge, 2003).

Data collection – repository (Database)

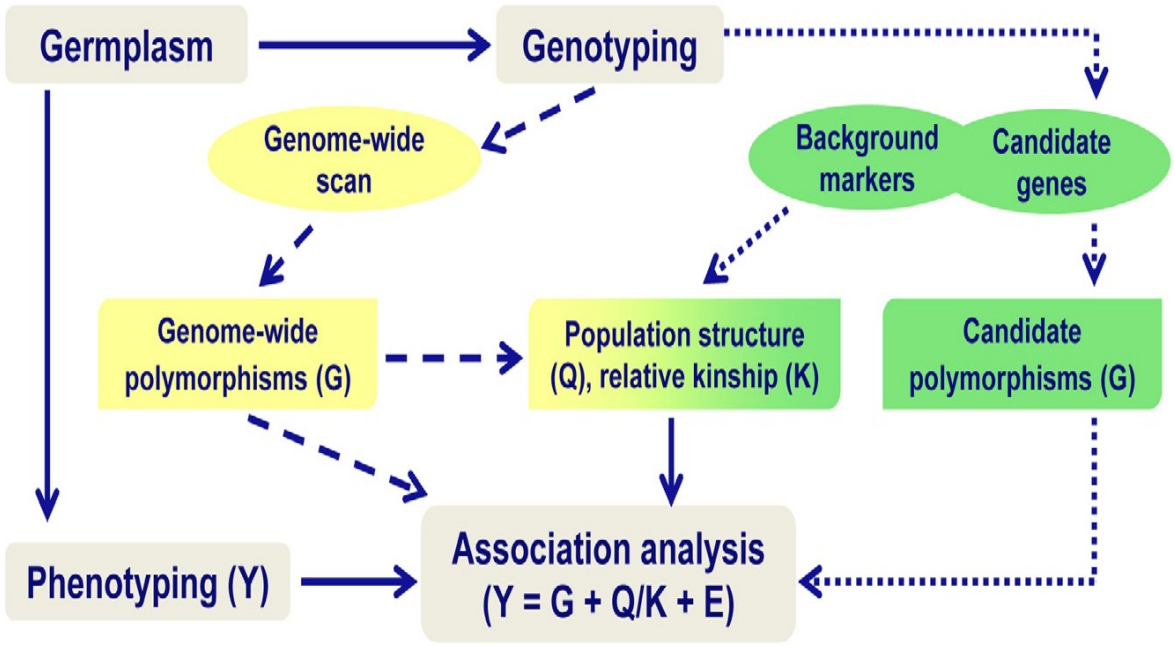
4. POPULATION STRUCTURE

Statistical methods for calculating population structure

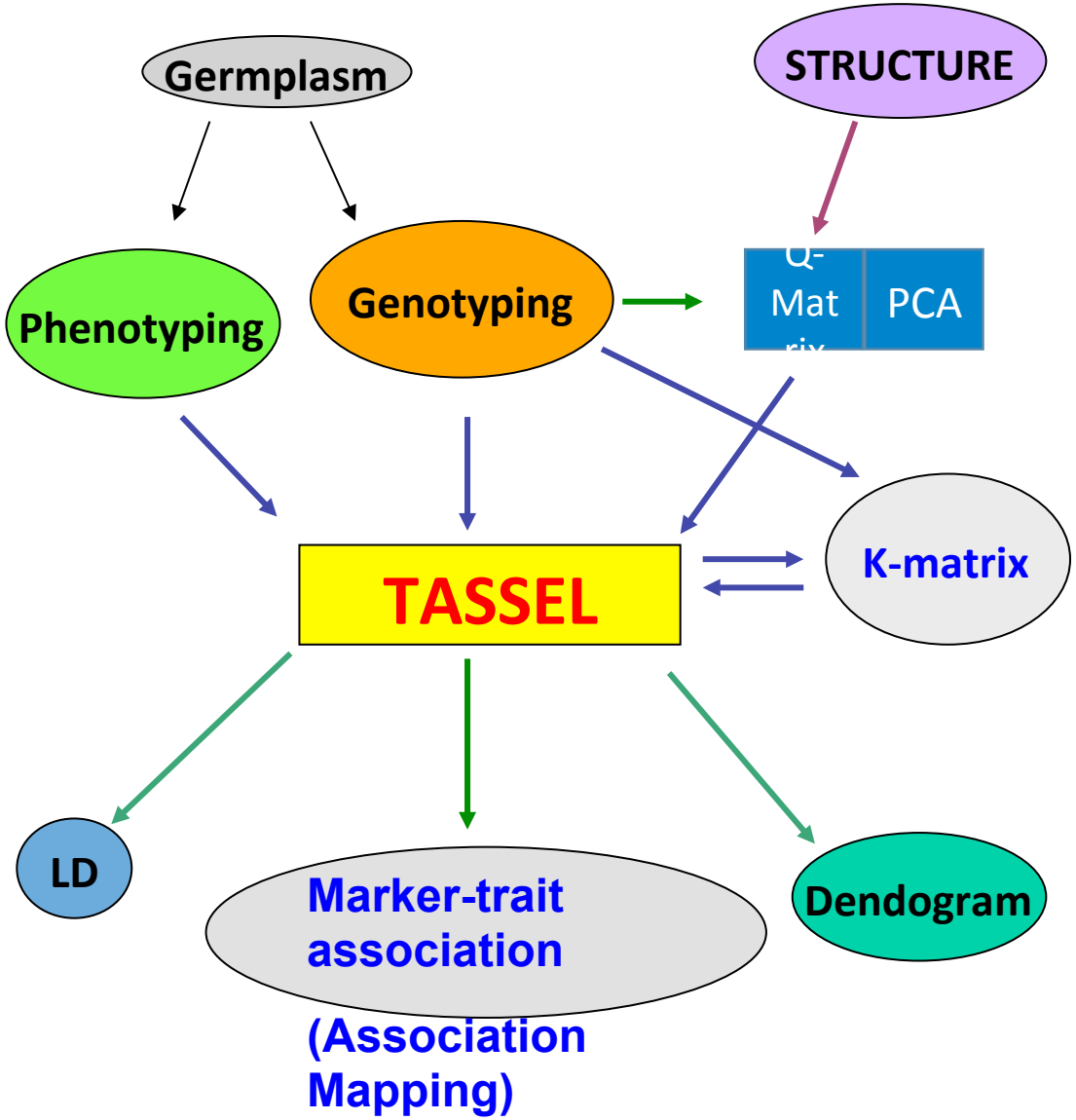
- **Structured associations (SA)** - uses a set of random markers to estimate population structure (Q) and then incorporates this estimate into further statistical analysis
- **Mixed model approach** - random markers are used to estimate Q and a relative kinship matrix (K), which are then fit into a mixed-model framework to test for marker-trait associations
- **Principal component analysis (PCA)** - summarizes variation observed across all markers into a smaller number of underlying component variables



5. Statistical Analysis



Genome-wide association mapping	Candidate-gene association mapping
It is a comprehensive approach to systematically search the genome for causal genetic variation. A large number of markers are tested for association with various complex traits, and prior information regarding candidate genes is not required. It works best for a research consortium with complementary expertise and adequate funding.	Candidate genes are selected based on prior knowledge from mutational analysis, biochemical pathway, or linkage analysis of the trait of interest. An independent set of random markers needs to be scored to infer genetic relationships. It is a low cost, hypothesis-driven, and trait-specific approach but will miss other unknown loci.



TASSEL = **T**rait **A**nalysis by a**SS**ociation, **E**volution, **L**inkage

Points for Analysis

GLM – Genotypic data + Phenotypic data + Q-matrix/PCA

MLM – Genotypic data + Phenotypic data + Q-matrix/PCA + K-matrix

- Association mapping without consideration of population structure would result in a high rate of false positive **Type I** errors
- Mixed-model approach to account for multiple levels of relatedness simultaneously, as detected by use of genetic markers, has improved control of both **type I** and **type II** error rates

- **False negative** : the declaration of an outcome as statistically non-significant, when the effect is actually genuine.

- **False positive** : the declaration of an outcome as statistically significant, when there is no true effect.

AM results

Trait	Marker	marker_p	markerR2
FT	S6_51258751	0.002526	0.118313
FT	S6_49476889	0.006728	0.104095
FT	S6_49476896	0.006728	0.104095
FT	S6_51074660	0.007553	0.06921
FT	S6_51074666	0.007553	0.06921
FT	S6_57125381	0.008782	0.073529
FT	S6_57125382	0.008782	0.073529
FT	S6_50090655	0.010302	0.09183
FT	S6_48617413	0.011146	0.092174
FT	S6_50083373	0.011984	0.060825
FT	S6_49946596	0.01502	0.06651
FT	S6_48622968	0.019185	0.085937
FT	S6_48825076	0.021259	0.063186
FT	S6_51074670	0.021644	0.074513

Association Mapping vs Linkage Mapping

WHICH ONE IS BEST?

AM is not a replacement of LM rather is a **complementary** to LM

- AM – species specific and population specific
- Sps with low genetic diversity LM is superior to AM

Ideal method

Combination of AM and LM

- where strength of each method are used to conduct
- high-resolution power and
- high-resolution tests

JOINT FAMILY POPULATION MAPPING

Low detection power – low allele frq and QTLs with small effects

More false positives – genotype – phenotype covariance

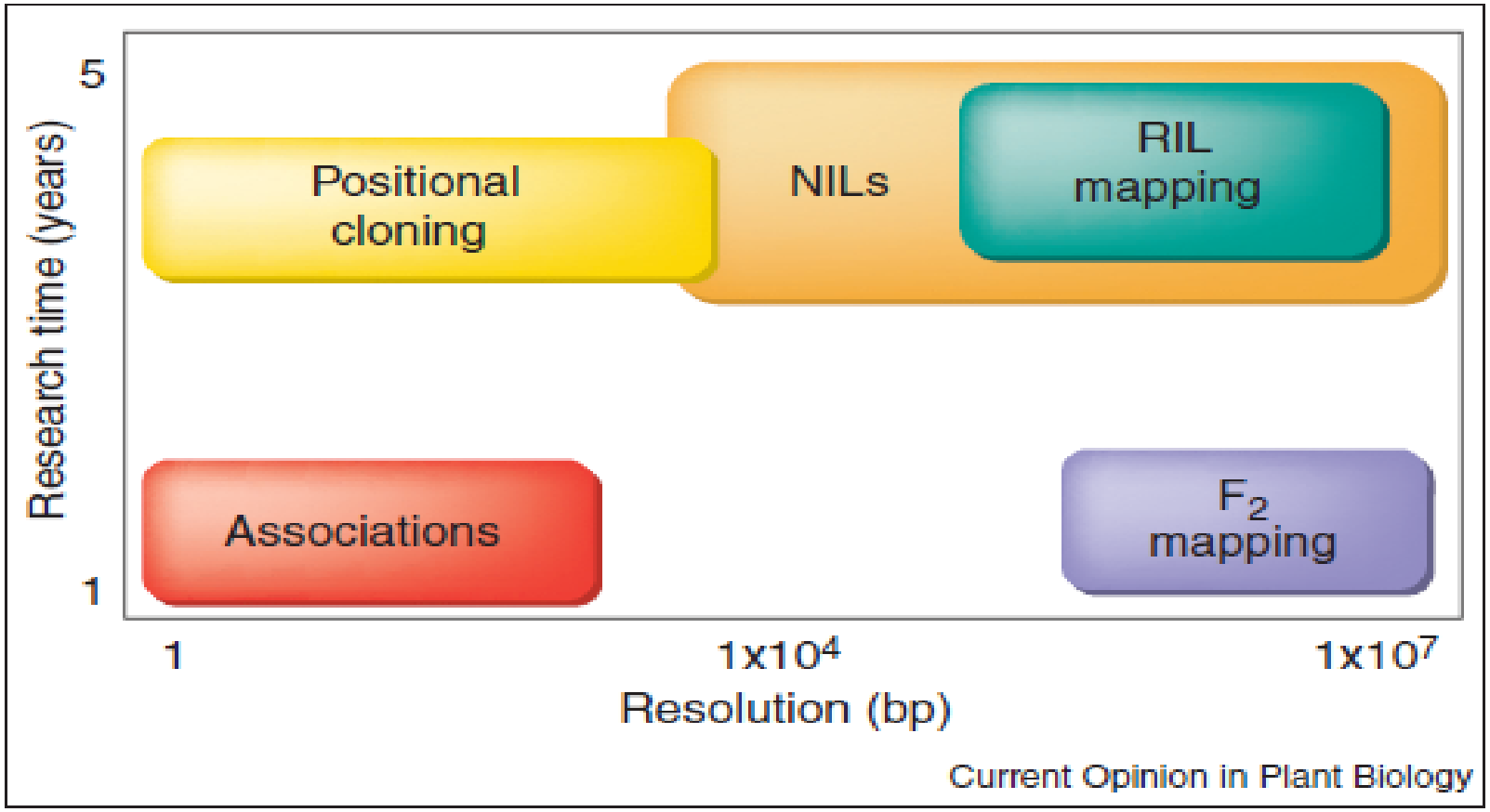
Manipulate allele free and population structure by controlled crosses and using family mapping to enhance the power

Nested Association Mapping (NAM)

(Joint linkage association mapping)

Relevant question is not what the nature of the genotyping data will be, but rather how to select germplasm to maximize the allelic diversity and the power to detect complex traits

Resolution vs Research Time



Comparison of resolution and research time for various approaches to dissect quantitative variation. The research times assume the target species has only two generations per year. NIL, near-isogenic line; RIL, recombinant inbred line.

GBS-SNP APPLICATIONS IN CROP PLANTS

SORGHUM & MAIZE

- ❑ Marker Discovery
- ❑ Phylogeny/Kinship
- ❑ Fine-Mapping
- ❑ Genomic Selection
- ❑ NAM-GWAS
- ❑ GWAS
- ❑ Improving reference genome assembly

Marker Discovery

MAIZE

Total Accessions : > 30000

Total SNPs : > 1.5M

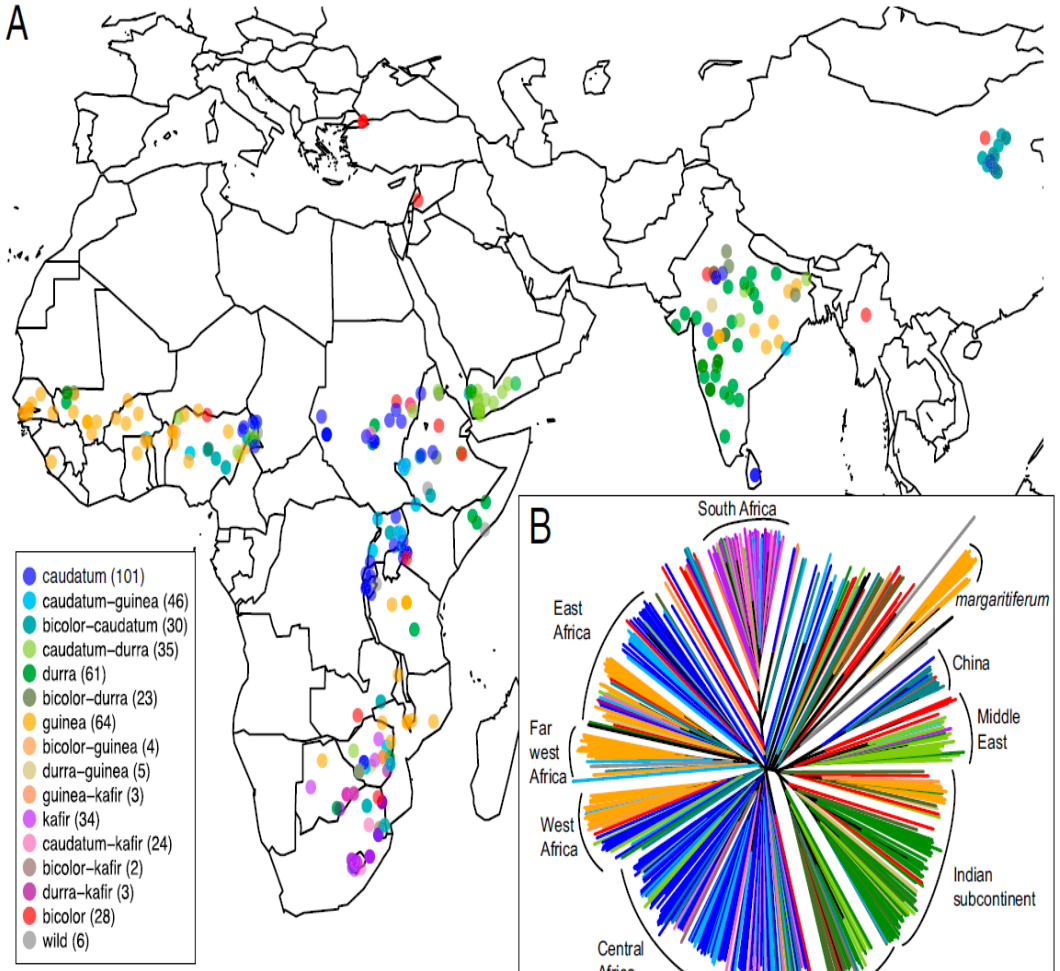
SORGHUM

Total Accessions : 7173

Total SNPs : 556,969

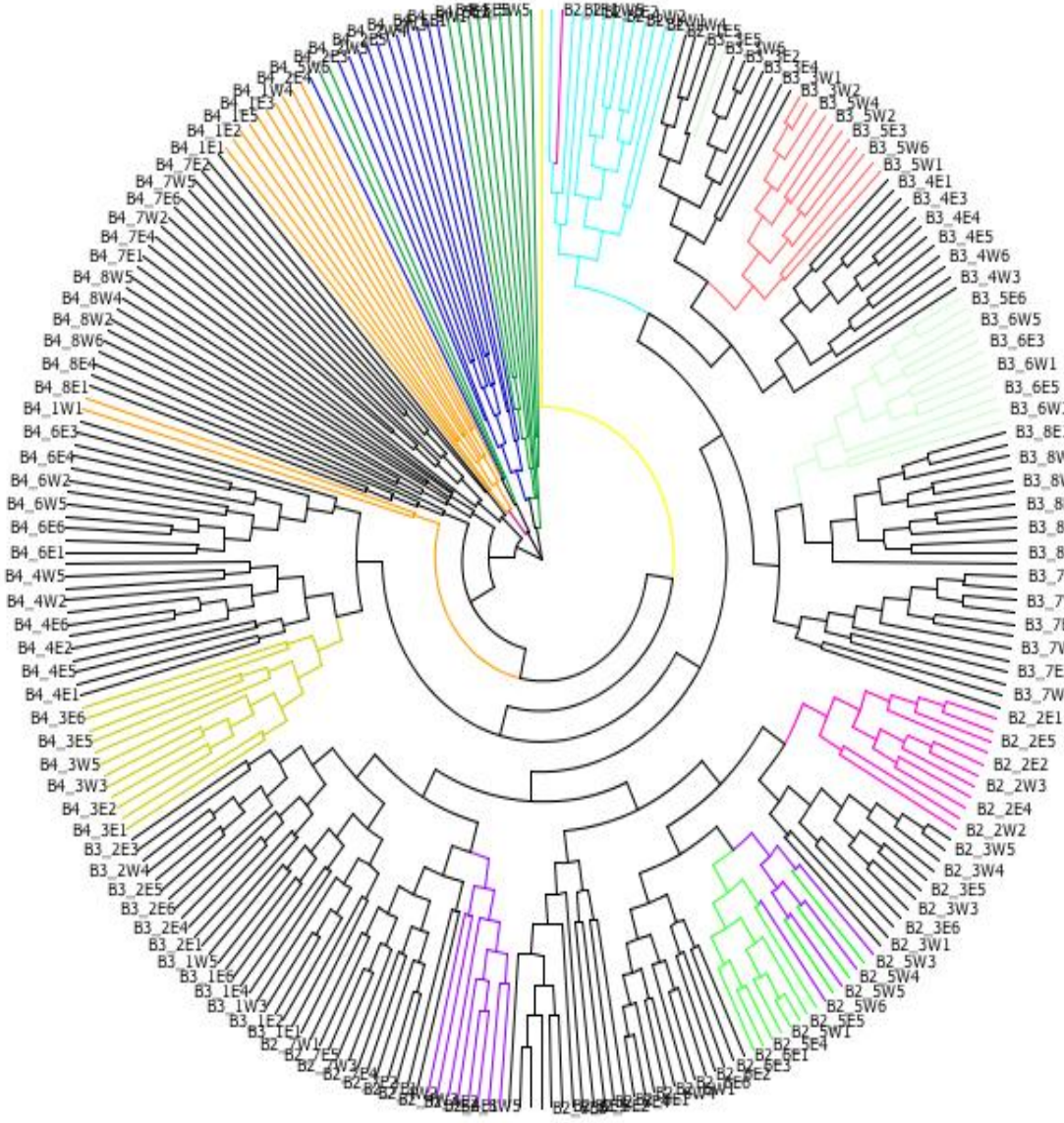
Total Data points : 3,995,138,637

Genetic Diversity/phylogeny



Cladogram of Sorghum

Source: Dr. Ram Sharma



Cladogram of Maize Hybrid and Landraces

Biasness in coordinate analysis in SSR, Chip-SNPs and GBS-SNPs in maize

Non stiff stalk,nss (106)

Stiff stalk, ss (28)

Tropical/sub tropical (66)

Popcorn (9)

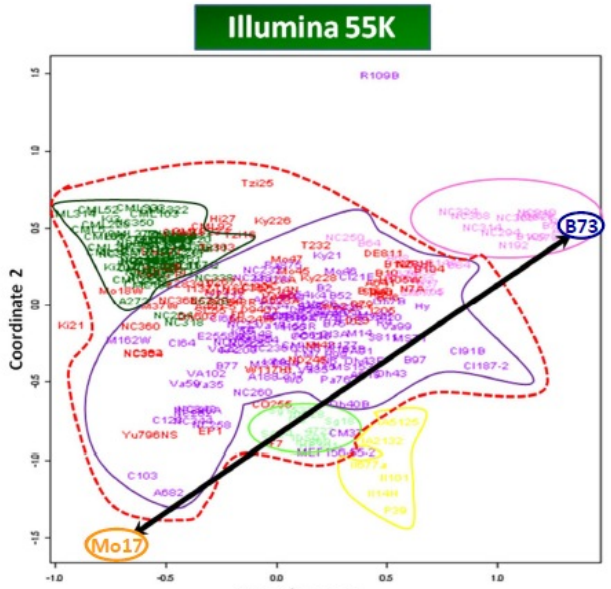
Sweet corn (6)

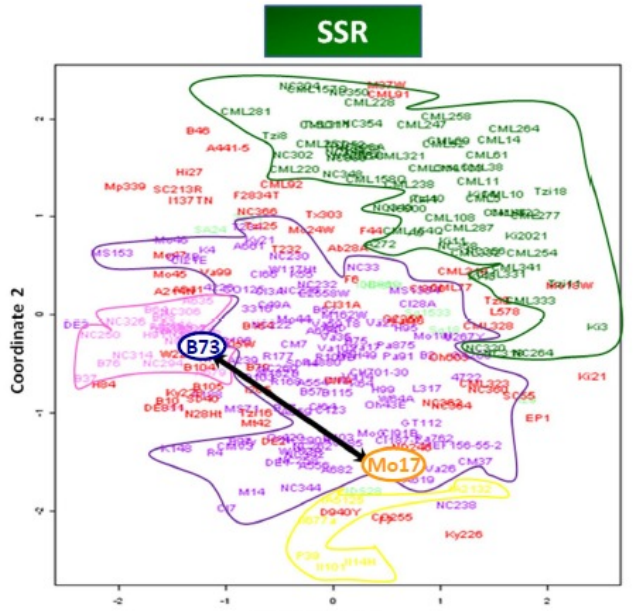
Unclassified(67)

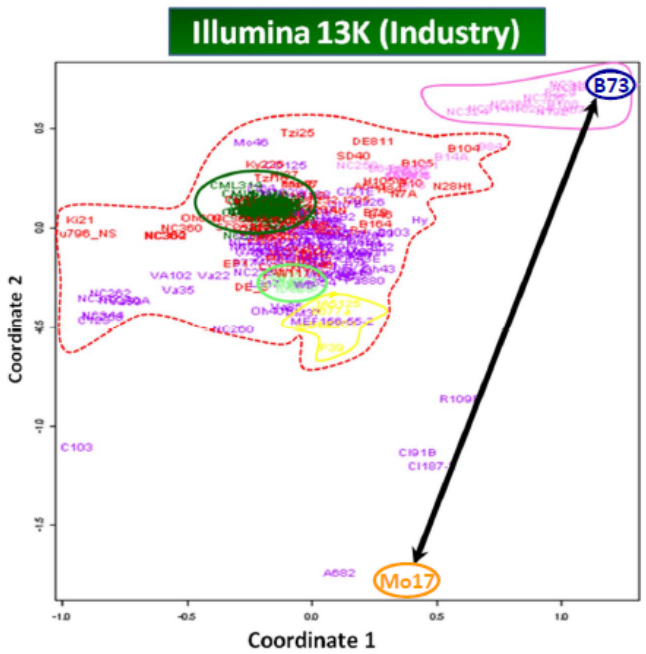
B73 (ss)

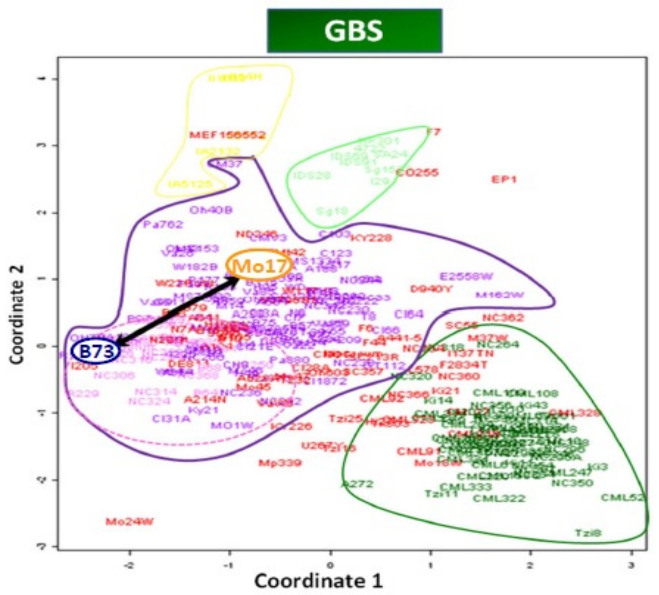
Mo17 (nss)

Source: Dr. Ram Sharma

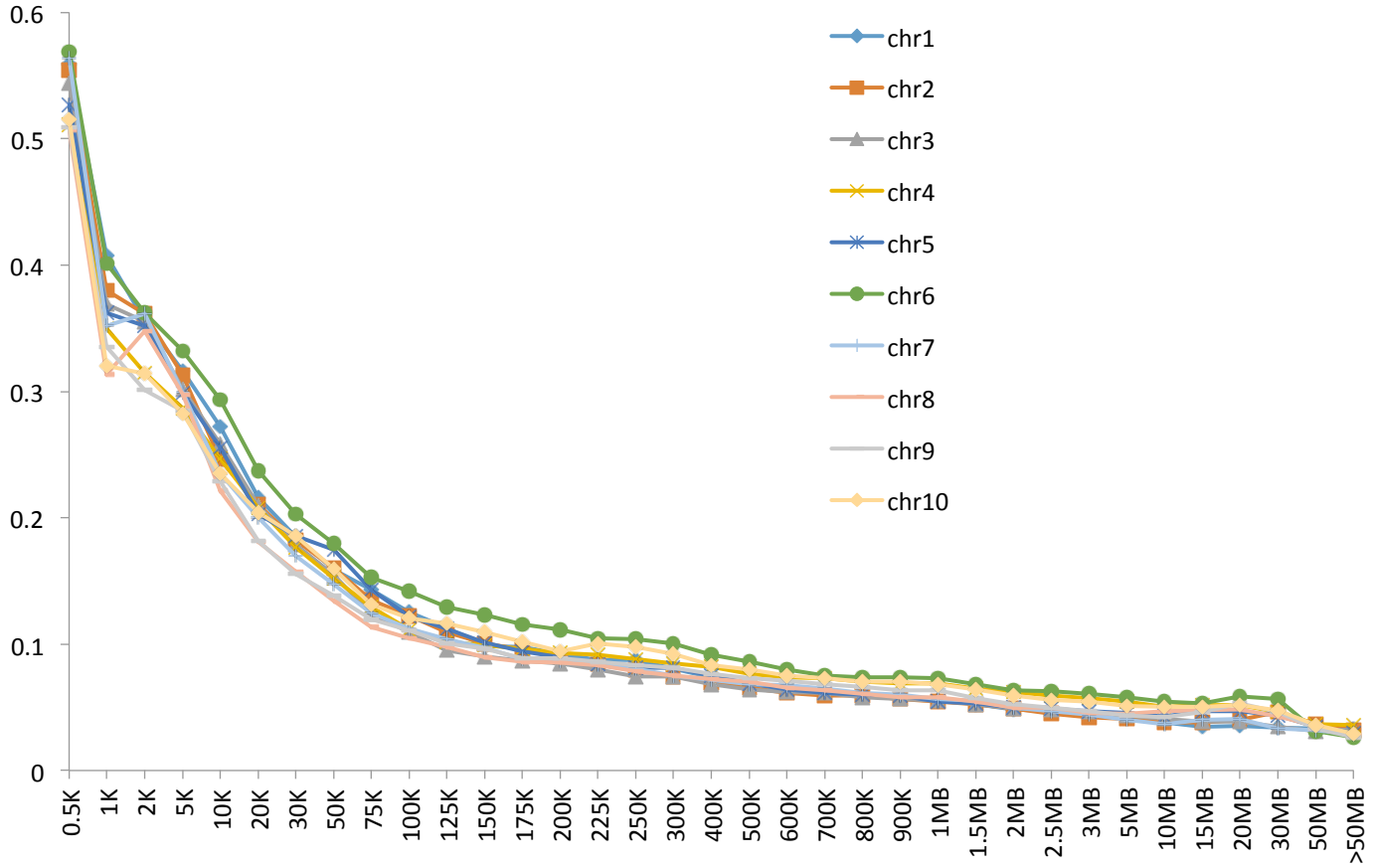








LD decay in sorghum



Chr	LD decay (kb)
1	125-150
2	150-175
3	100-125
4	125-150
5	150-175
6	300-400
7	125-150
8	100-125
9	125-150
10	225-250

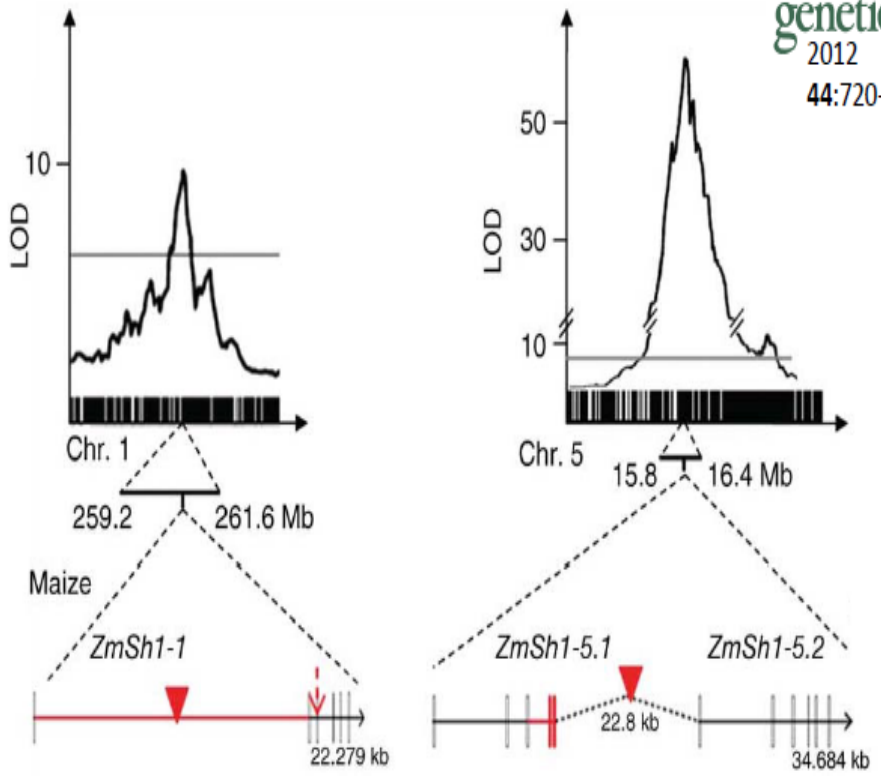
Source: Dr. Punna Ramu

Bi-Parental Linkage mapping/Fine Mapping

Parallel domestication of the *Shattering1* genes in cereals

Zhongwei Lin[1], Xianran Li[1], Laura M Shannon[2], Cheng-Ting Yeh[3,4], Ming L Wang[5], Guihua Bai[1,6], Zhao Peng[7], Jiarui Li[7], Harold N Trick[7], Thomas E Clemente[8], John Doebley[2], Patrick S Schnable[3,4], Mitchell R Tuinstra[9], Tesfaye T Tesso[1], Frank White[7] & Jianming Yu[1]

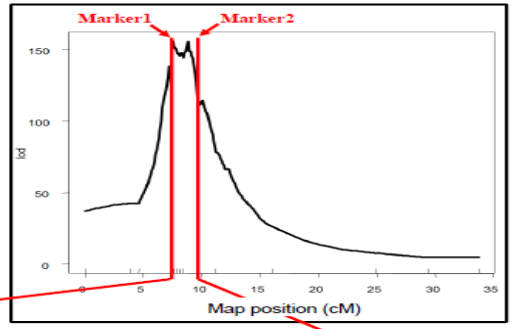
nature genetics 2012 44:720-725

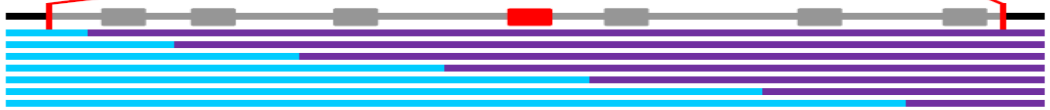


Maize *Sh1* orthologs are located at seed shattering QTLs

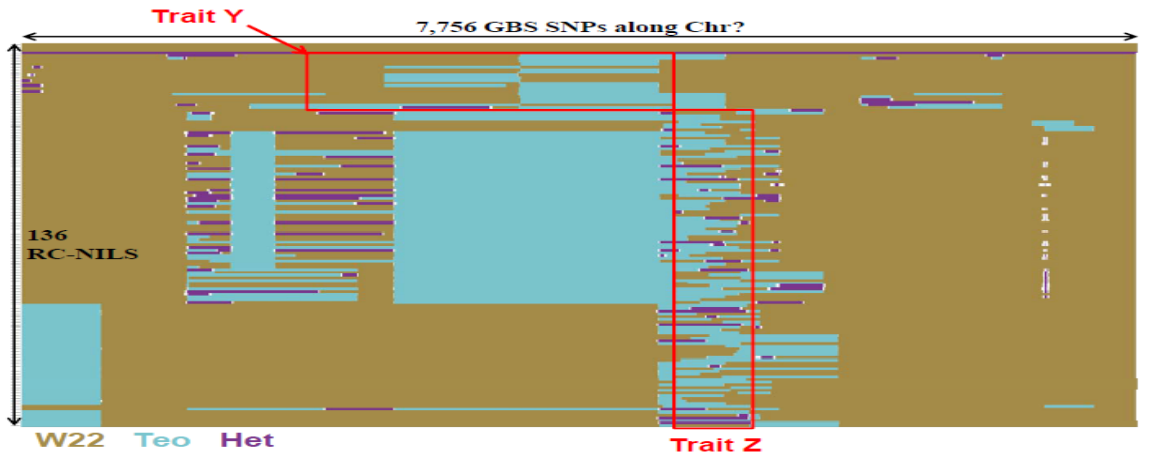
Fine mapping QTL

- 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

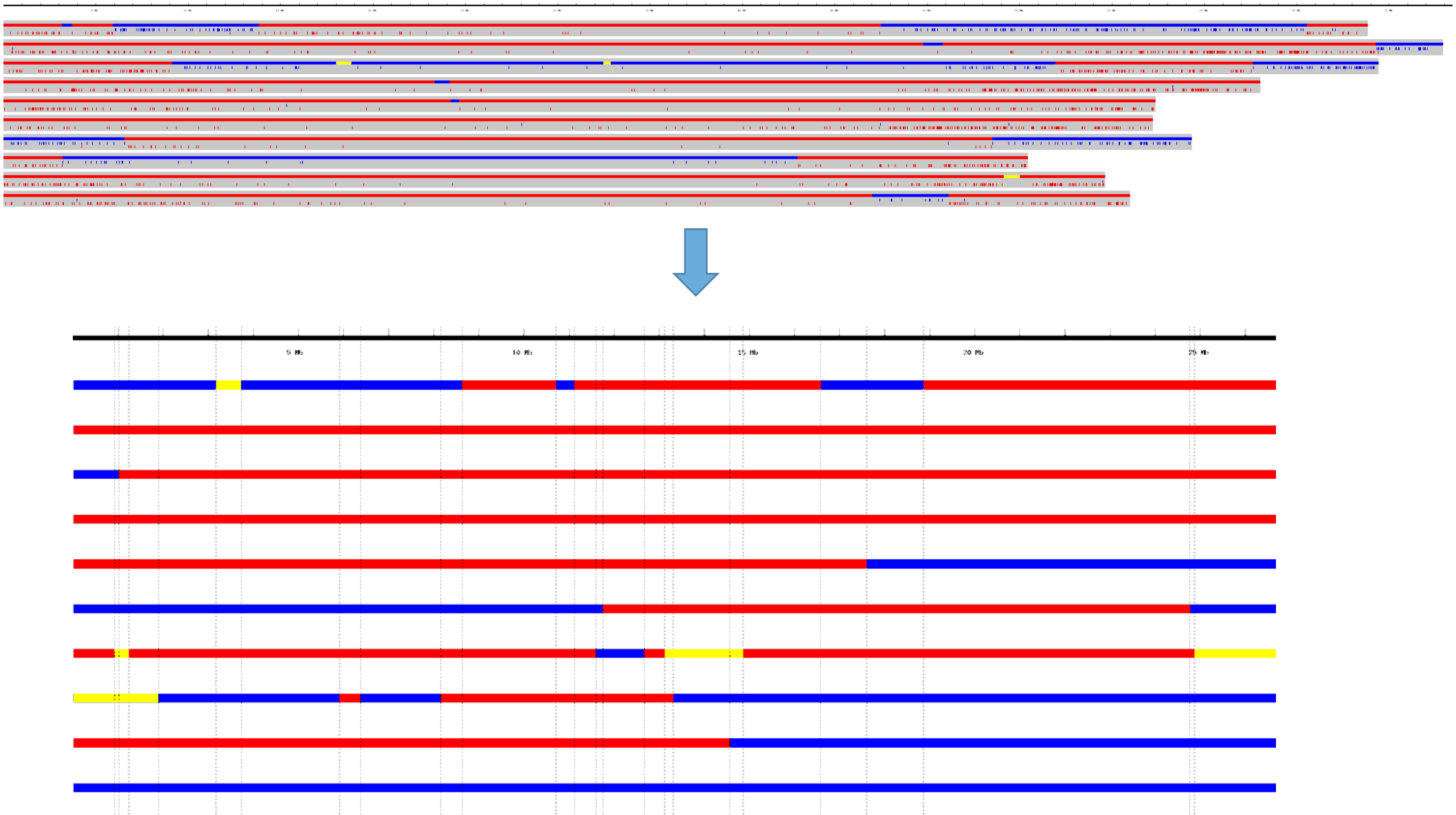




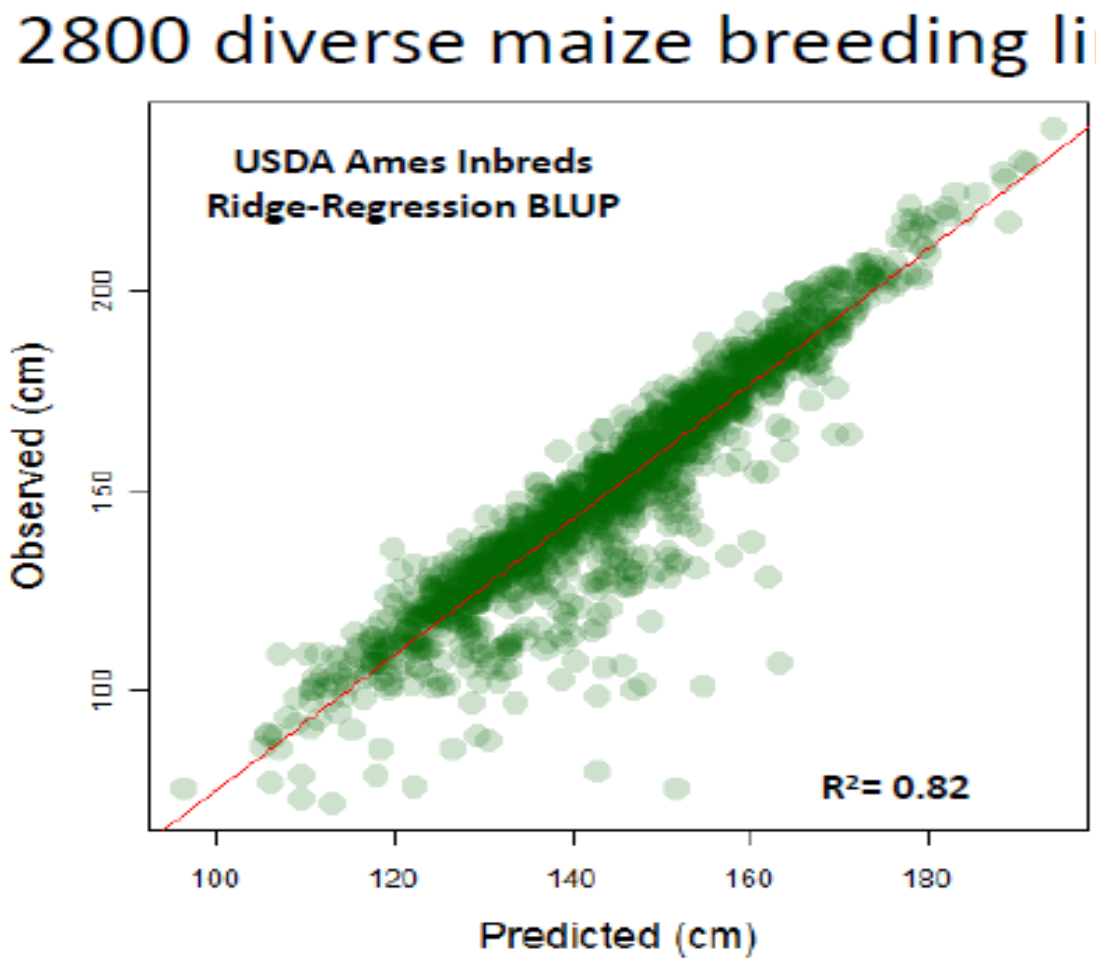
Fine Mapping of Domestication QTL in Maize



Recombination break points



Genomic prediction in maize

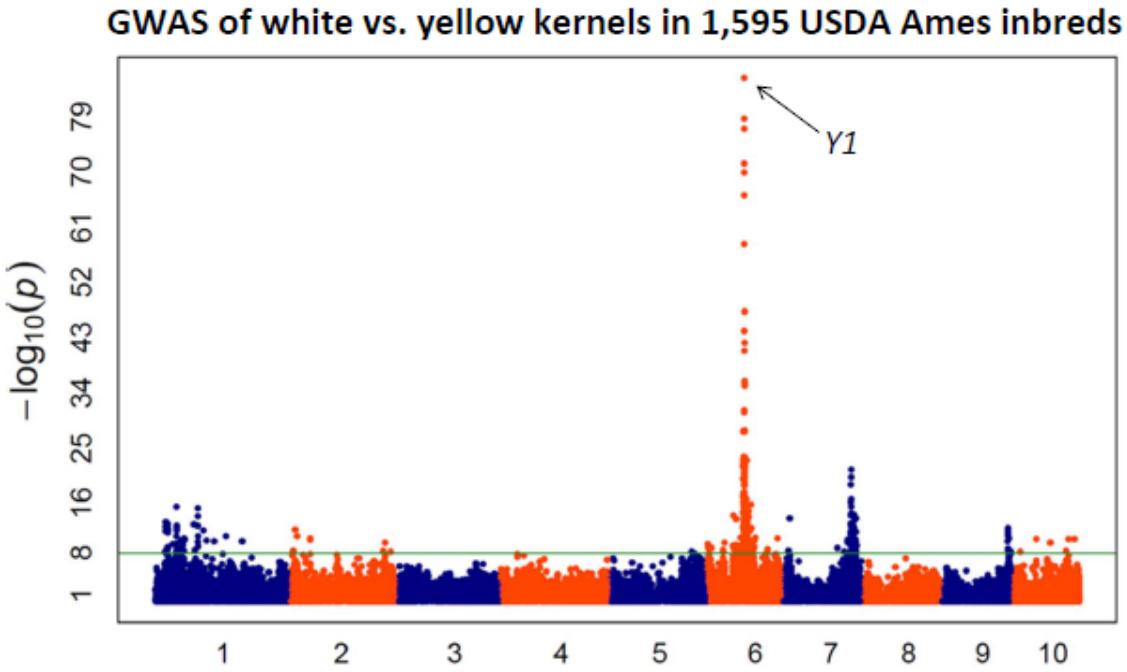


Source : Jason Peiffer

Accurate genomic prediction of height based on GBS data in maize

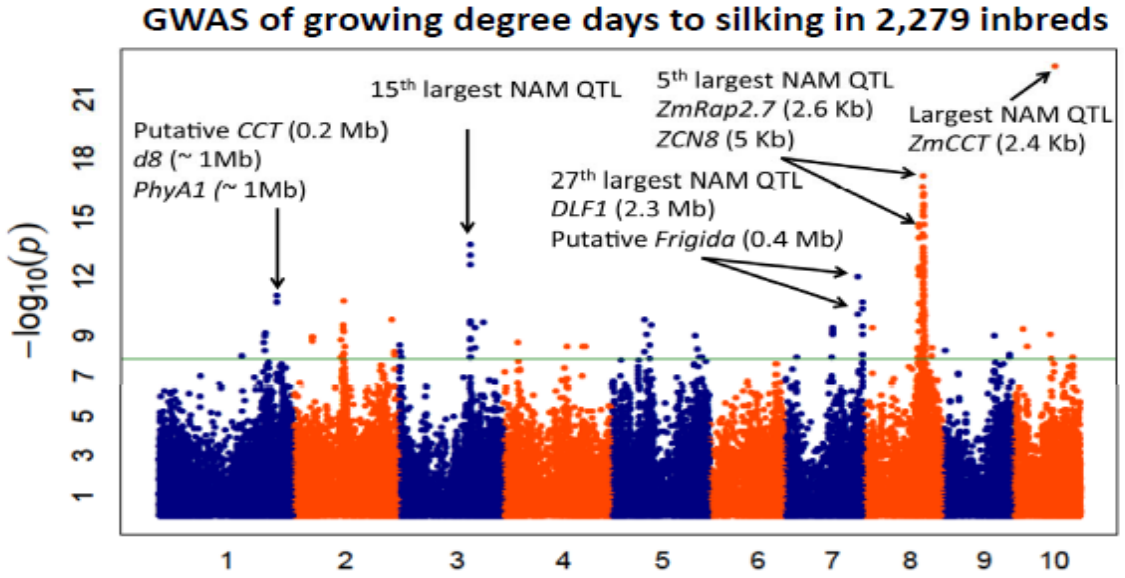
GWAS examples in plants

GWAS directly hits known Mendelian traits



The best hit for kernel color lies within *Y1*

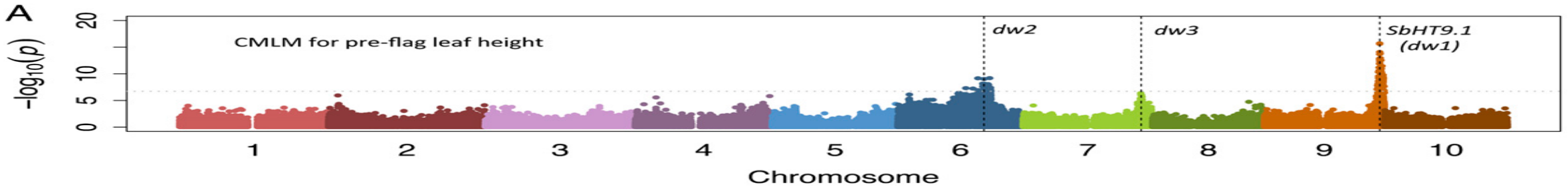
GWAS of a more comples traits directly hits known flowering time

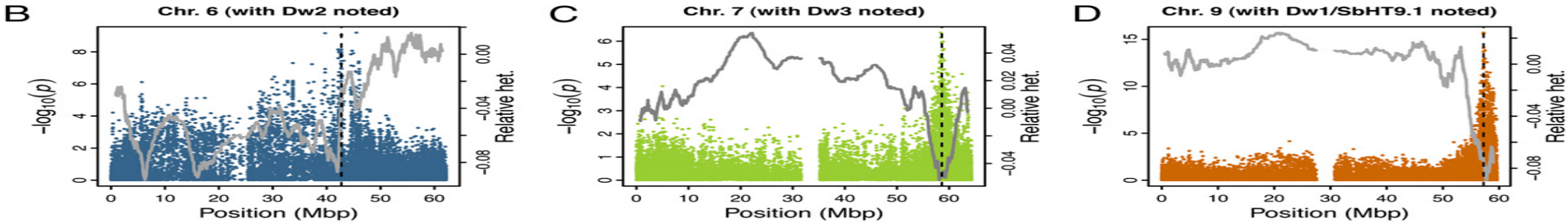


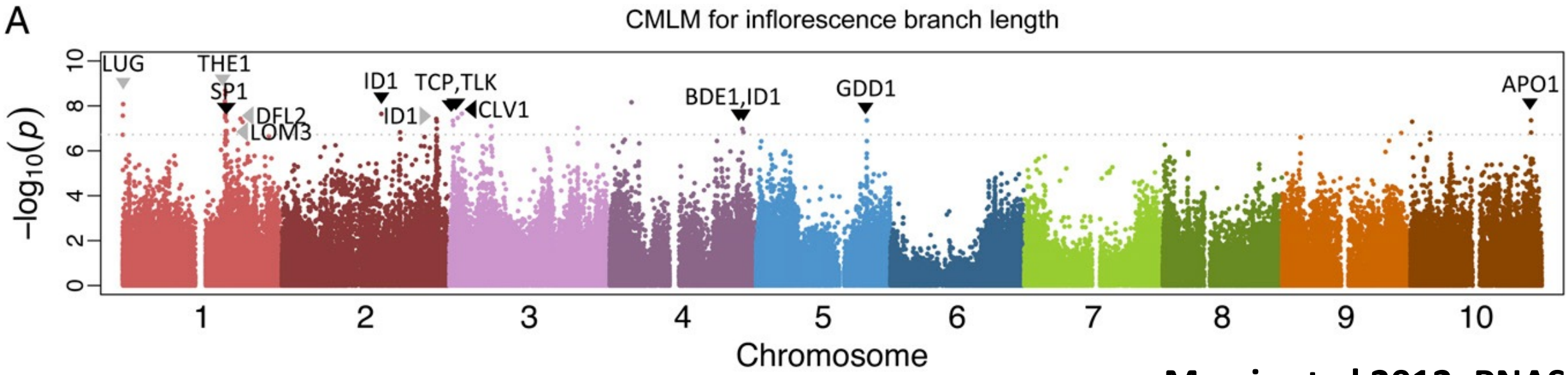
Even with ~660K SNPs we almost missed ZmCCT (only one significant SNP)

Roamy et al. 2013 ; Genome Biology

Cont....



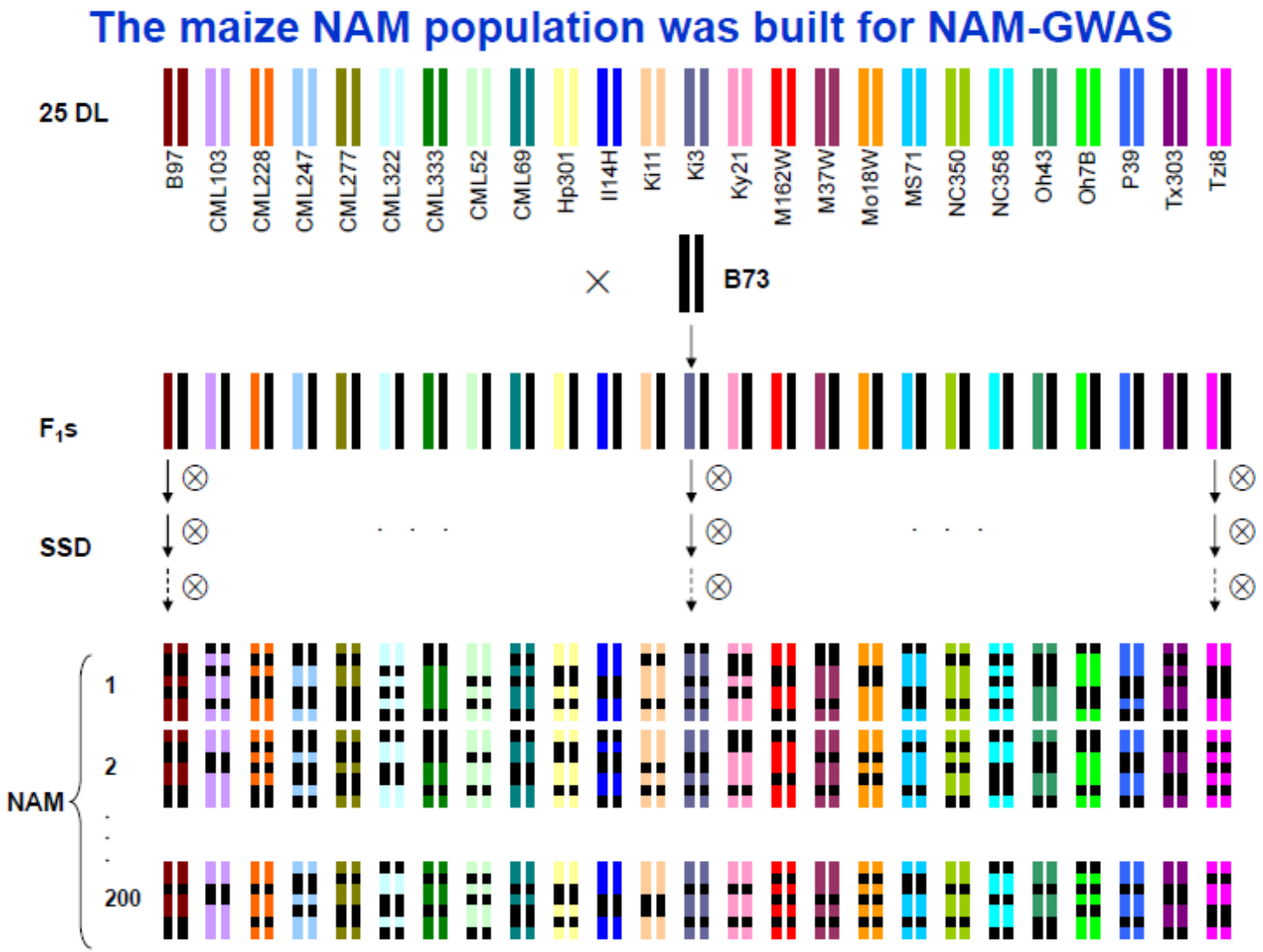




Morris et al 2012; PNAS

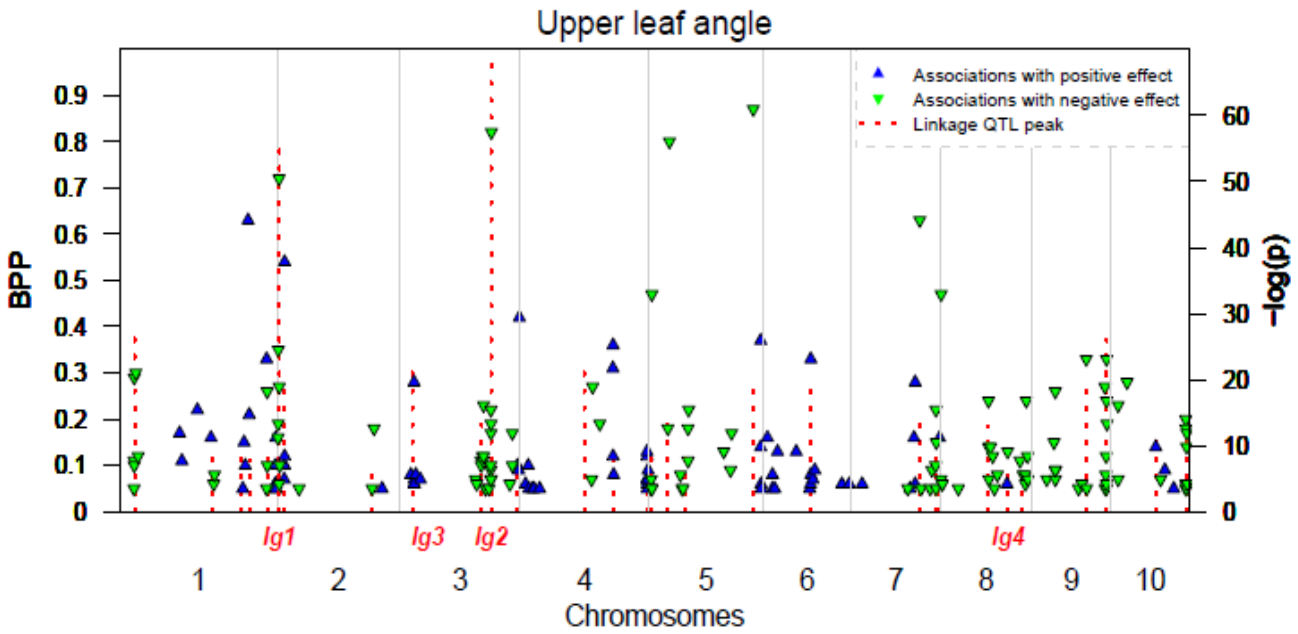
NAM (nested association mapping)-GWAS in maize

Development of NAM mapping population in maize



Cont….

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



Tian, Bradbury, et al 2011 Nature Genetics

GBS data improves the resolution of joint linkage analysis in the NAM population

Trait: Days to silk (flowering time)

Markers	Median QTL support interval
1,106 array SNPs	6.2 cM
171,479 GBS SNPs	2.6 cM

The Maize B73 reference genome: room for improvement?

- ❑ The B73 reference genome accurate for B73 but less so far other maize lines line M017

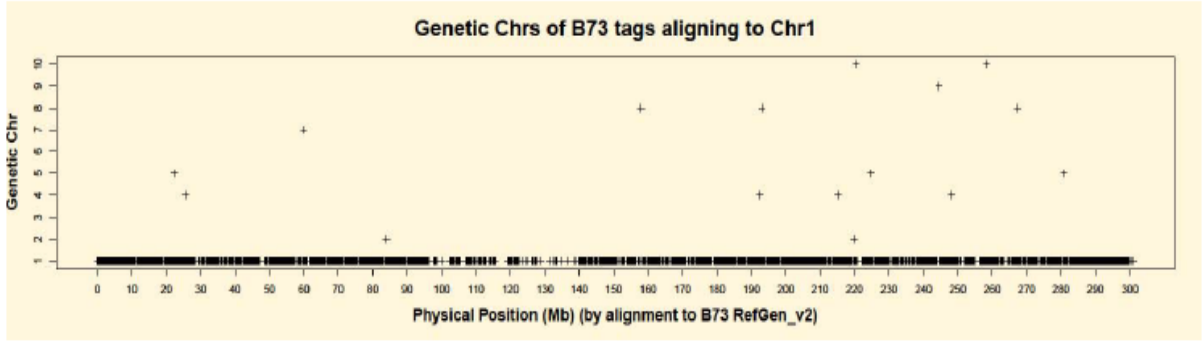
- ❑ Even for 73, some regions of the genome are in the wrong location

- ❑ Some large (multiple BAC) contigs could not be anchored
  - assigned to ‘chromosome 0’
  - 30 chr0 contigs in B73 RefGenV1
  - 30 chr0 contigs in B73 RefGenV1

- ❑ Some regions of the genome are missing
  - ~ 5% of the 73 sequence is not in the B73 reference genome

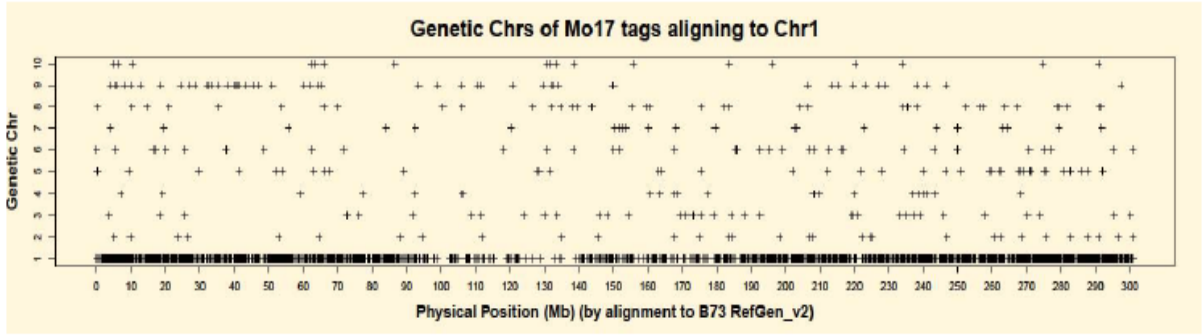
Cont....

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

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

Acknowledgements

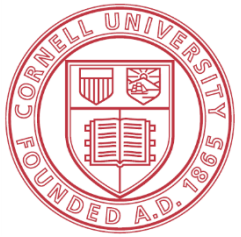
Odeny Damaris

Stefania Grando

Santosh Deshpande

Trushar Shah

Punna Ramu



Ed Buckler

Jeff Glaubitz

James Harriman

Terry C

Rob Elshire

Sharon E. Mitchell

Alex Lipka

Fei Lu

Thank You

for your kind attention