De novo SNP discovery and mapping of tetraploid cotton genome via a simplified genotype-by-sequencing (GBS) approach

Surender Verma, Carla Jo Logan-Young, Ryan McCormick, Richard Percy, Alan Pepper and John Yu

Interspecific Linkage Map

What are the two parents, and why do we use them?

G. hirsutum TM-1 and *G. barbadense* 3-79, the genetic standards for their respective species.

An interspecific RIL mapping population was developed using these two parents.

Genetic maps were constructed using this RIL population.

Reference:

Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. Young-Hoon Park, Magdy S. Alabady, Mauricio Ulloa, Brad Sickler, Thea A. Wilkins, John Yu, David M. Stelly, Russell J. Kohel, Osama M. El-Shihy, Roy G. Cantrell. **Molecular Genetics and Genomics (2005) 274: 428-441.**

Key metrics of map:
193 loci (EST, CSR), 1277 cM,

Reference:

Cotton genome mapping with new microsatellites from Acala 'Maxxa' BAC-ends. James E. Frelichowski, Michael B. Palmer, Dorrie Main, Jeffrey P. Tomkins, Roy G. Cantrell, David M. Stelly, John Yu, Russell J. Kohel, Mauricio Ulloa. **Molecular Genetics and Genomics (2006) 275: 479–491.**

Key metrics of map:
433 loci (SSR, BAC ends), 2126.3 cM, 4.9cM

Reference:

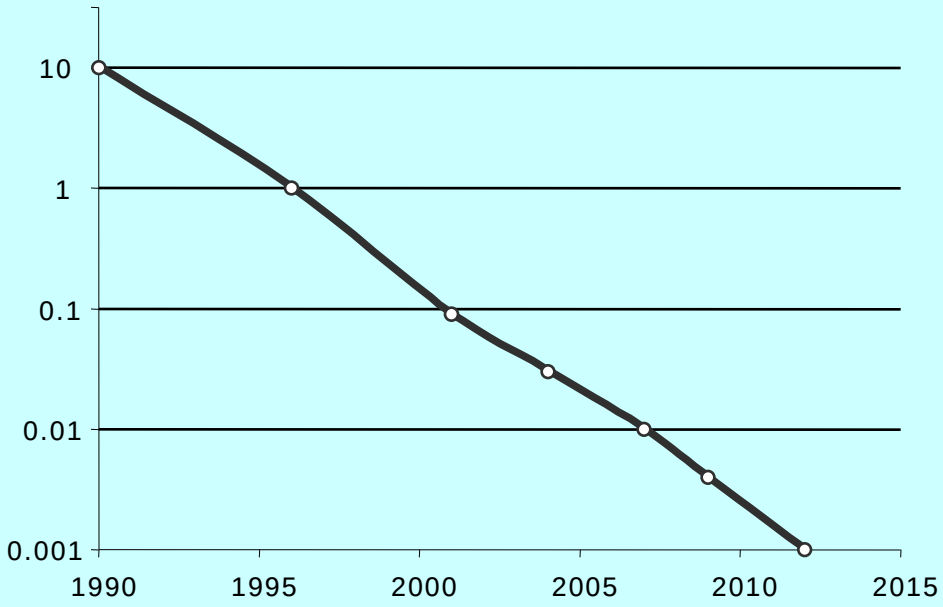
A High-Density Simple Sequence Repeat and Single Nucleotide Polymorphism Genetic Map of the Tetraploid Cotton Genome

John Z. Yu, Russell J. Kohel, David D. Fang, Jaemin Cho, Allen Van Deynze, Mauricio Ulloa, Steven M. Hoffman, Alan E. Pepper, David M. Stelly, Johnie N. Jenkins, Sukumar Saha, Siva P. Kumpatla, Manali R. Shah, William V. Hugie, Richard G. Percy. (2012) **G3: Genes, Genomes, Genetics 2:43-58.**

Key metrics of map:

2072 loci (SSR and SNP/llumina Golden Gate markers), 3380 cM, 1.63 cM density

Cost of DNA sequencing per finished base-pair (US$)



Reducing to half every 22-month

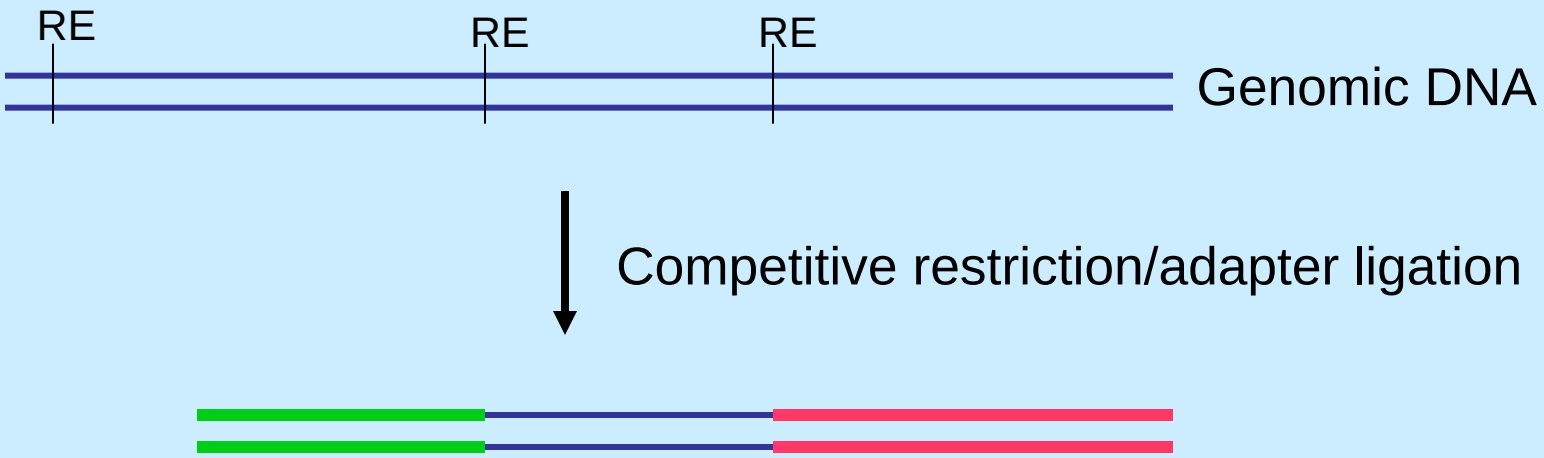
- Genotype-by-sequencing (GBS)
- Restriction site associated DNA (RAD-seq)
- Reduced representation libraries (RRLs)
- Digital genotyping (DG)

technically simple, highly multiplexed, suitable for population studies, germplasm characterization, trait mapping. Low per-sample cost – based on NGS of genome targeted by restriction enzymes.

Our Methods

Barcoding scheme for *Bsr*GI and *Hin*p1I

RE = *Bsr*GI or *HIn*P1I



Restriction enzymes

*Bsr*GI – methylation not sensitive

5´ - GTAC

*Hinp*1I – methylation blocked

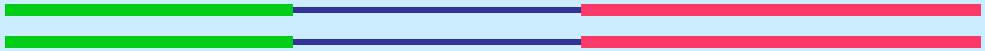
5´ - CG

Barcoding scheme for *Bsr*GI and *Hin*p1I

RE = *Bsr*GI or *HIn*P1I

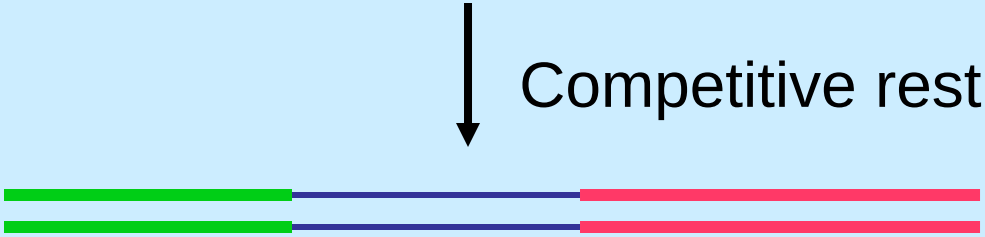
Genomic DNA

Competitive restriction/adapter ligation

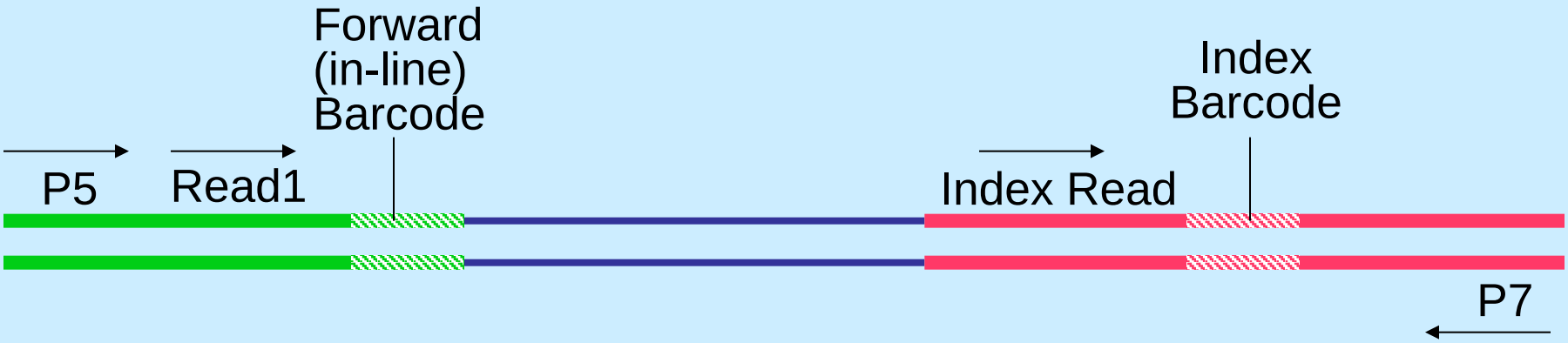


RE = *Bsr*GI or *Hln*P1I

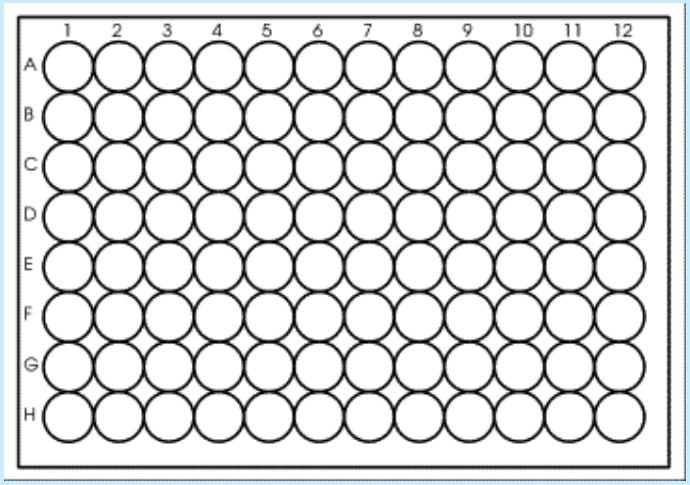
Genomic DNA

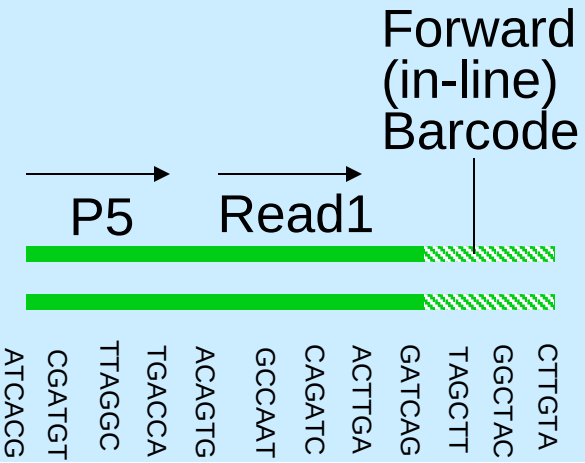


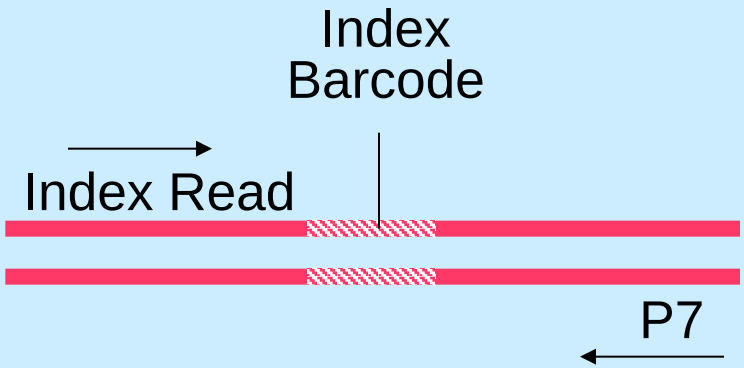
Competitive restriction/adapter ligation

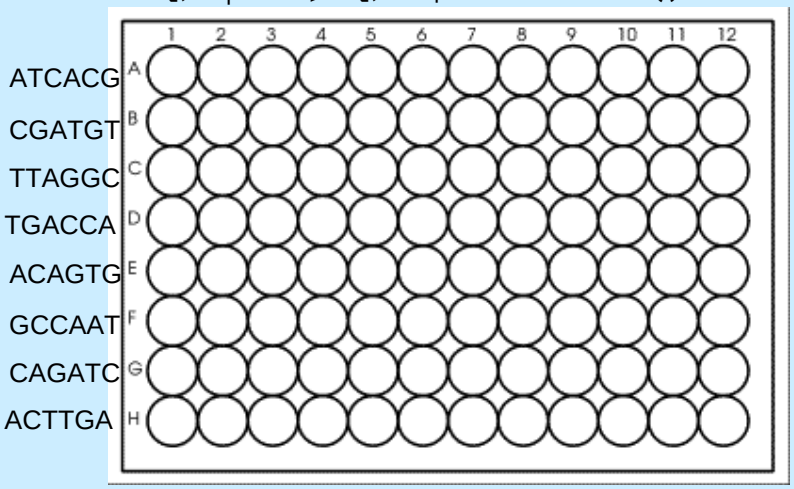


Genomic DNA isolation from 186 individuals in population (+parents and RILS)









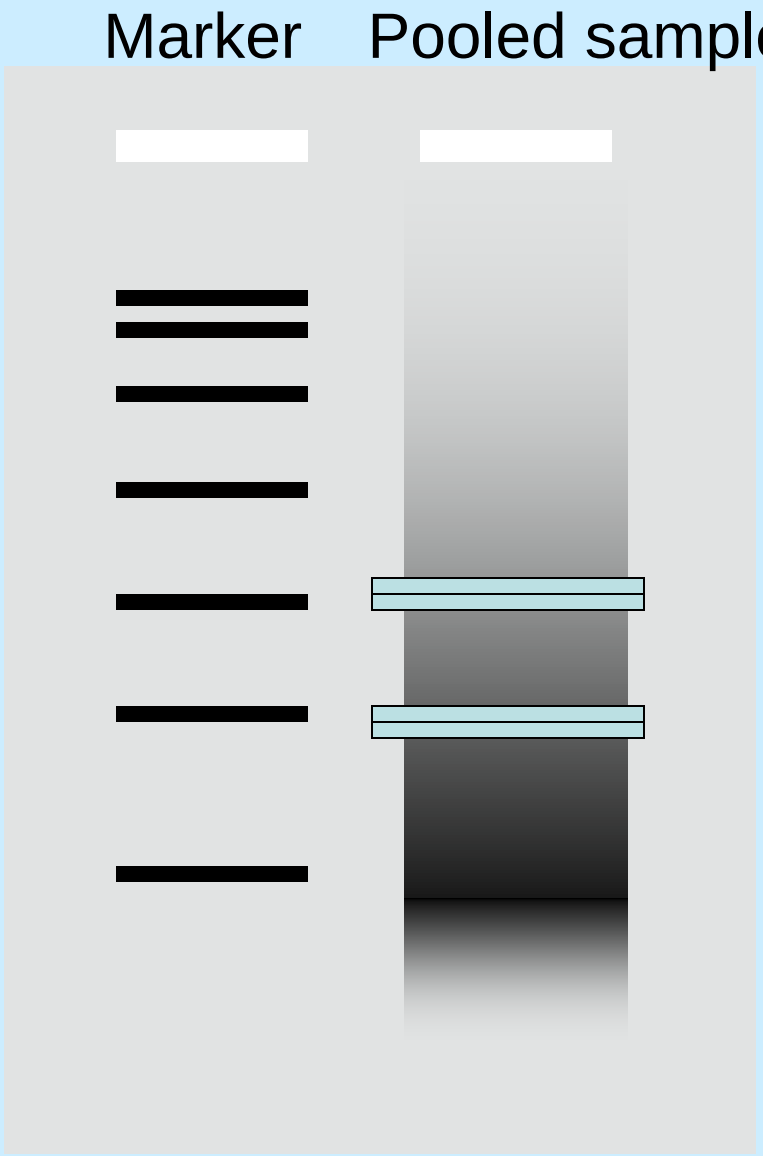
Adaptor scheme

RE = *Bsr*GI or *Hln*P1I

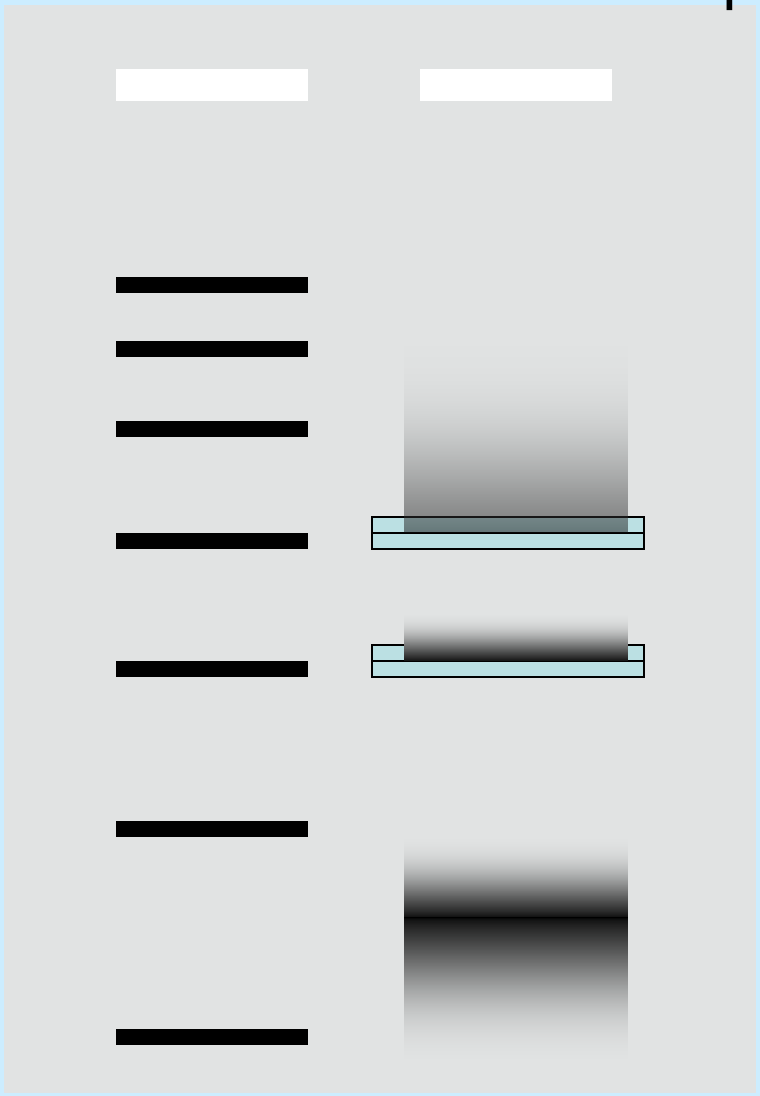
Genomic DNA

Competitive restriction/adapter ligation

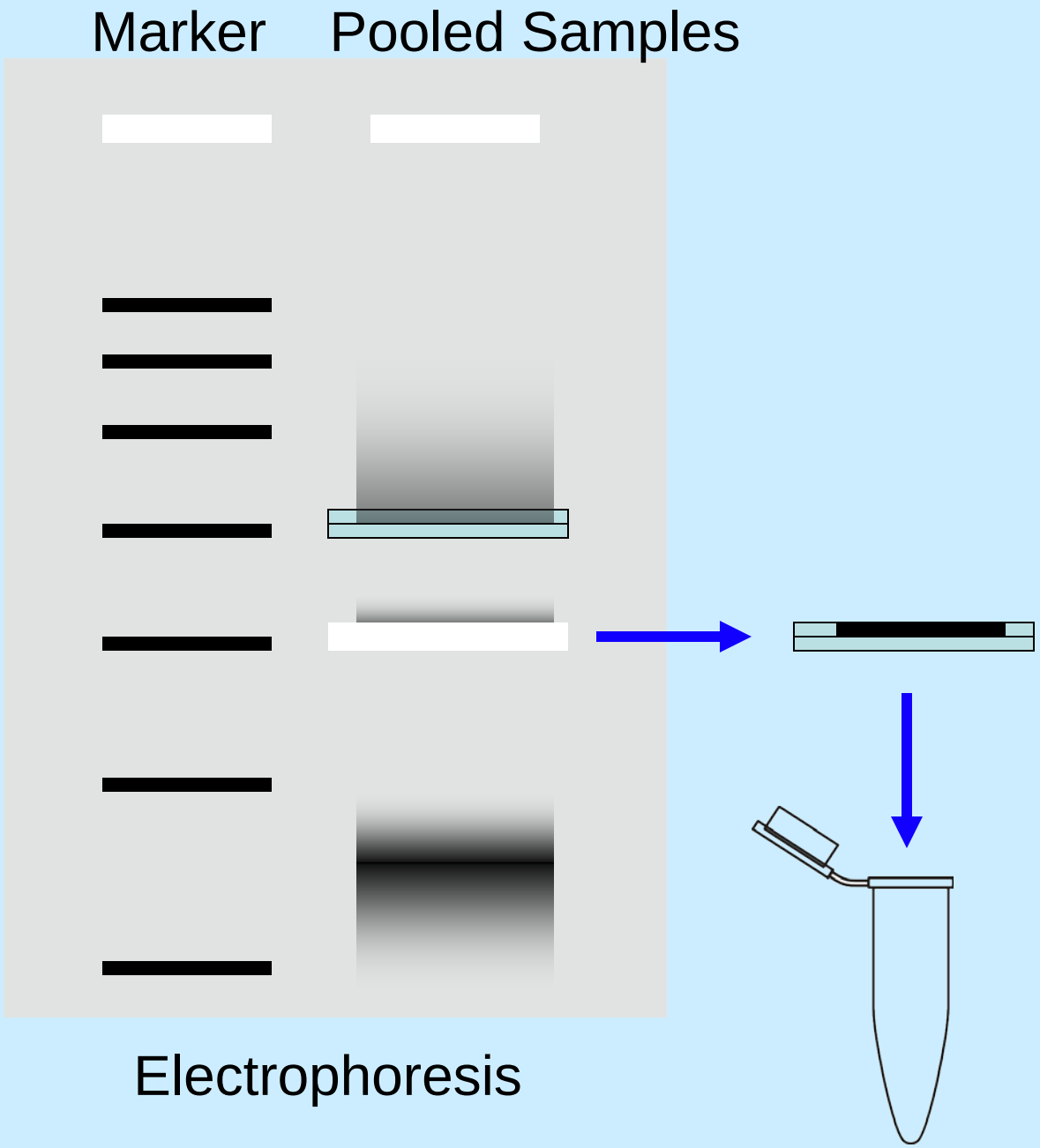
Pool <u>all</u> samples

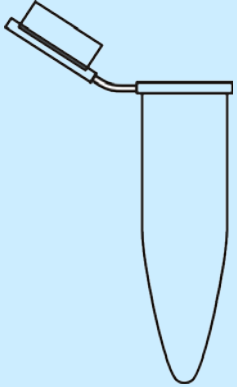

Marker
Pooled samples
Insert Reco chips
Electrophoresis (Gel Green)

Marker Pooled Samples



Electrophoresis


Marker
Pooled Samples
Electrophoresis

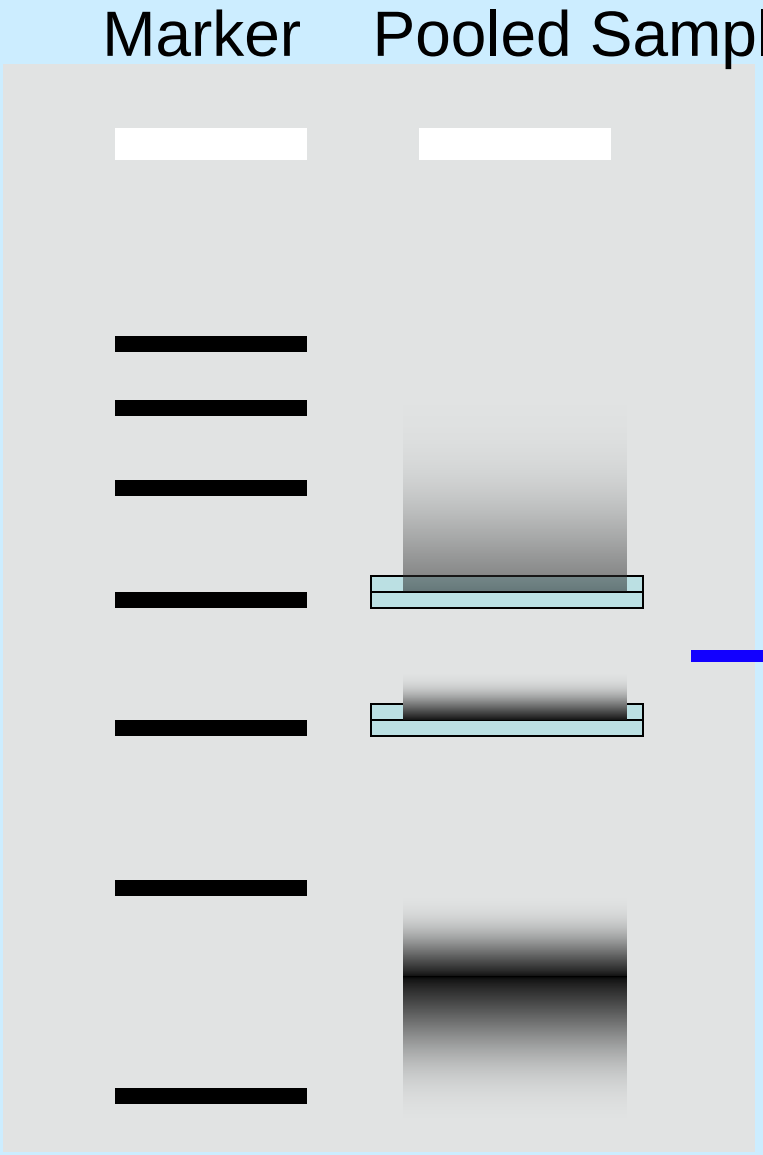


NEB ssDNA isolation kit

PCR (P5 and P7 primers)

Purify (Ampure beads)

100 bp single-end sequence on Illumina Hi-seq 2500
(96 individuals per lane)



GBS system is fully ‘tunable’

Wider size range extracted
- More fragments sequenced
- Lower depth of coverage

Narrow size range extracted
- Fewer fragments sequenced
- Higher depth of coverage

Bioinformatics pipeline

1. Fastq output from Illumina Hi-seq 2500

2. Trim for quality (http://www.geneious.com/)

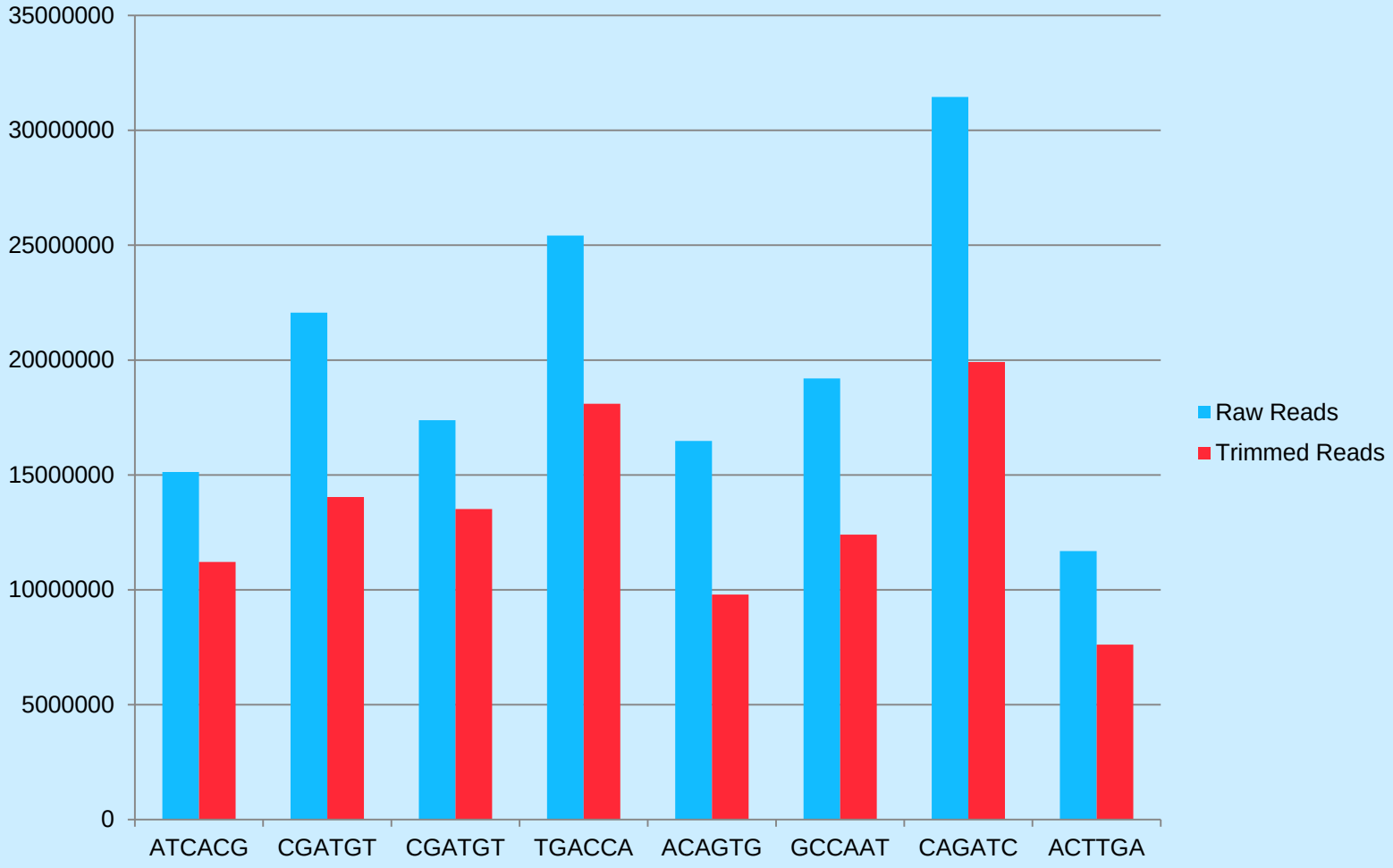
3. Filter for quality and minimum read length (94bp) (http://www.geneious.com/)

3. Trim to uniform read length 94 bp (http://www.clcbio.com/)

4. Stacks
(Catchen *et al*. (2011), **G3: Genes, Genomes, Genetics** 1:171-182)
(http://creskolab.uoregon.edu/stacks/)

process_radtags
denovo_map.pl

Index (P7) Reads *Bsr*GI



Total Reads= 158,812,061 (15.8 Gb)

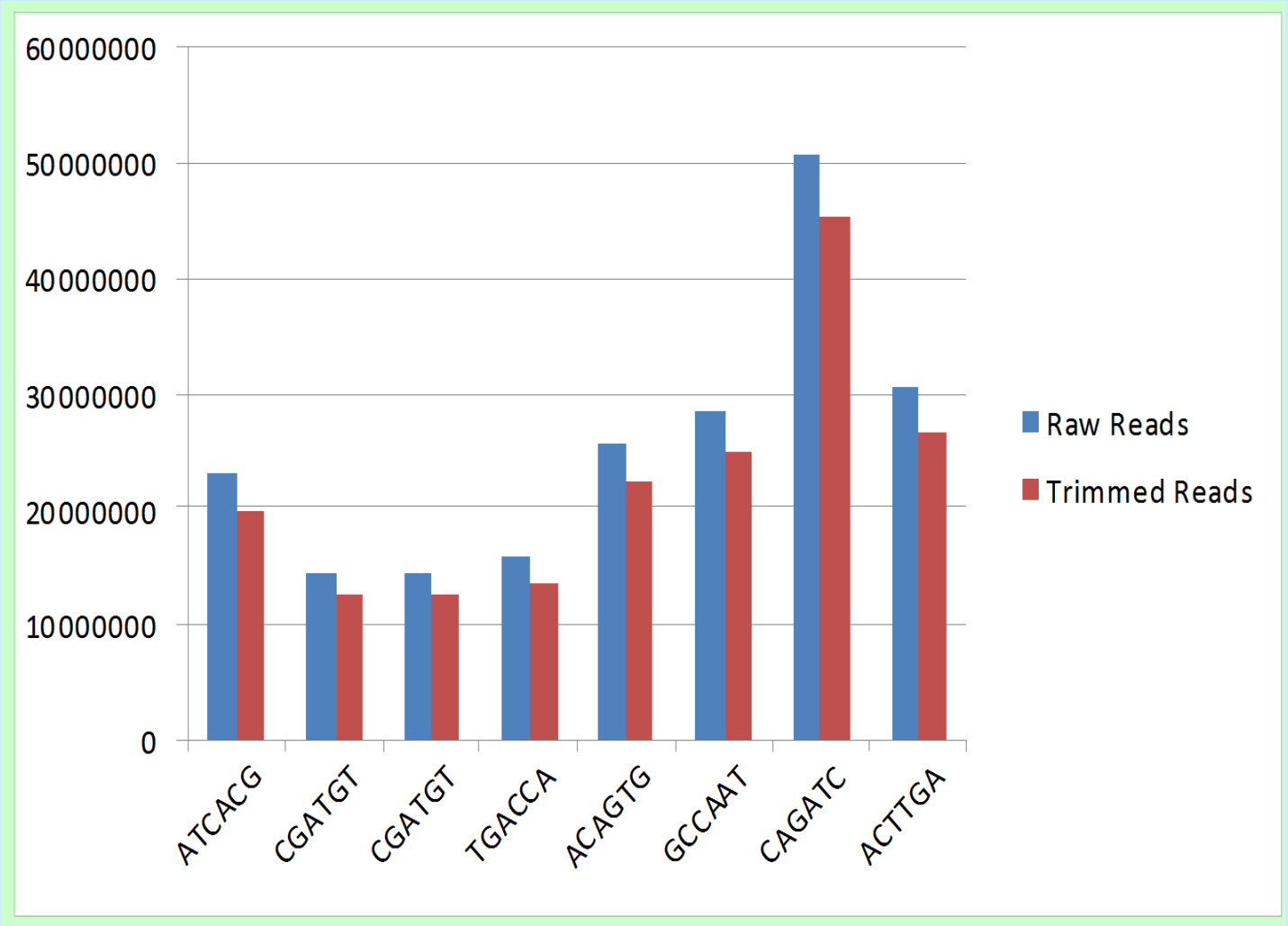
Sequence metrics

*Bsr*GI, methylation insensitive

Barcodes	Reads	Stacks	Polymorphic stacks
ATCACG	11,214,539	47,921	2,308
CGATGT	14,037,139	47,921	2,308
CGATGT	13,514,319	47,921	2,308
TGACCA	18,101,188	47,921	2,308
ACAGTG	9,791,644	47,921	2,308
GCCAAT	12,404,218	47,921	2,308
CAGATC	19,915,098	47,921	2,308
ACTTGA	7,614,837	47,921	2,308

4.8% of fragments have interspecific SNPs

Index Reads *Hin*p1I



Total Reads = 203,281,003 (20.3 Gb)

Sequence metrics

*Hin*p1I, methylation sensitive

Barcodes	Reads	Stacks	Polymorphic stacks
ATCACG	19,907,200	15,785	1,721
CGATGT	12,605,099	15,785	1,721
CGATGT	12,527,273	15,785	1,721
TGACCA	13,622,588	15,785	1,721
ACAGTG	22,265,602	15,785	1,721
GCCAAT	24,955,208	15,785	1,721
CAGATC	45,522,752	15,785	1,721
ACTTGA	26,526,217	15,785	1,721

10.9% of fragments have interspecific SNPs

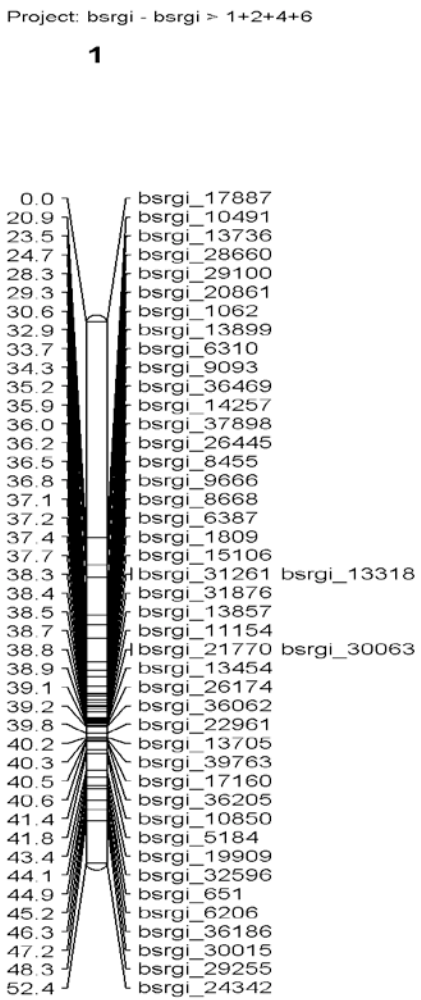
Mapping method

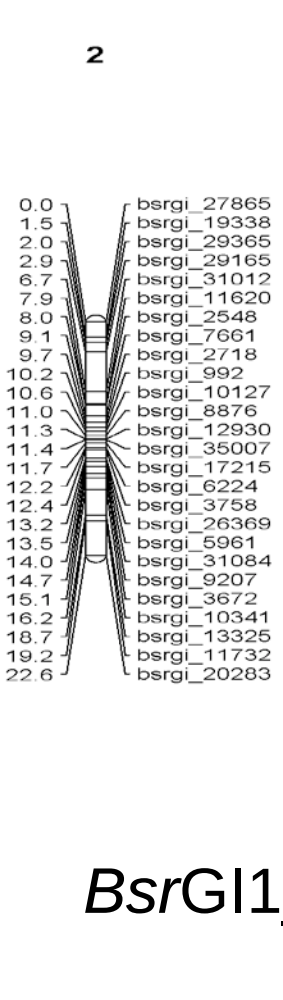
Joinmap 4.0
(http://www.kyazma.nl/index.php/mc.JoinMap)
Van Ooijen *et al*. **Genetics Research** (2011) 93: 343-349

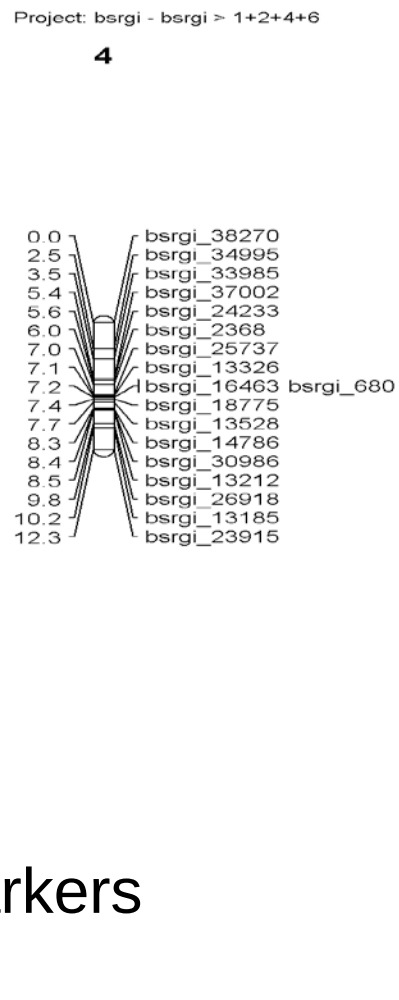
TM-1 x 3-79 RIL mapping validation (ongoing)

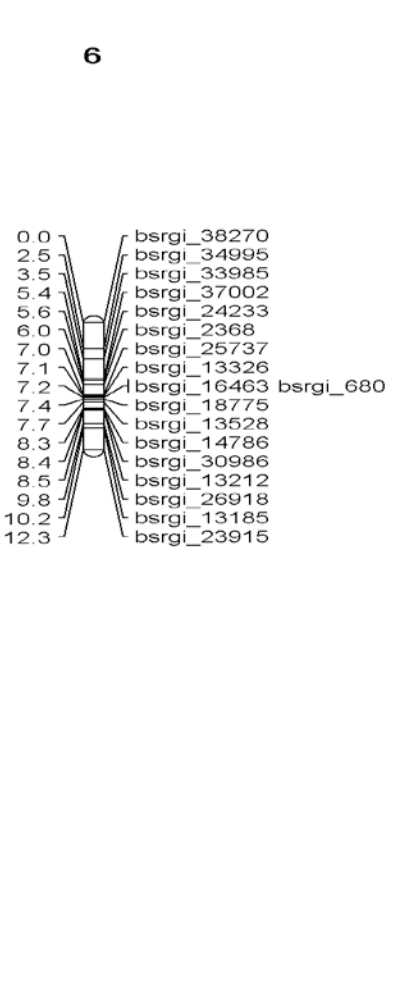
3014 (2038+976) de novo SNPs

Map examples



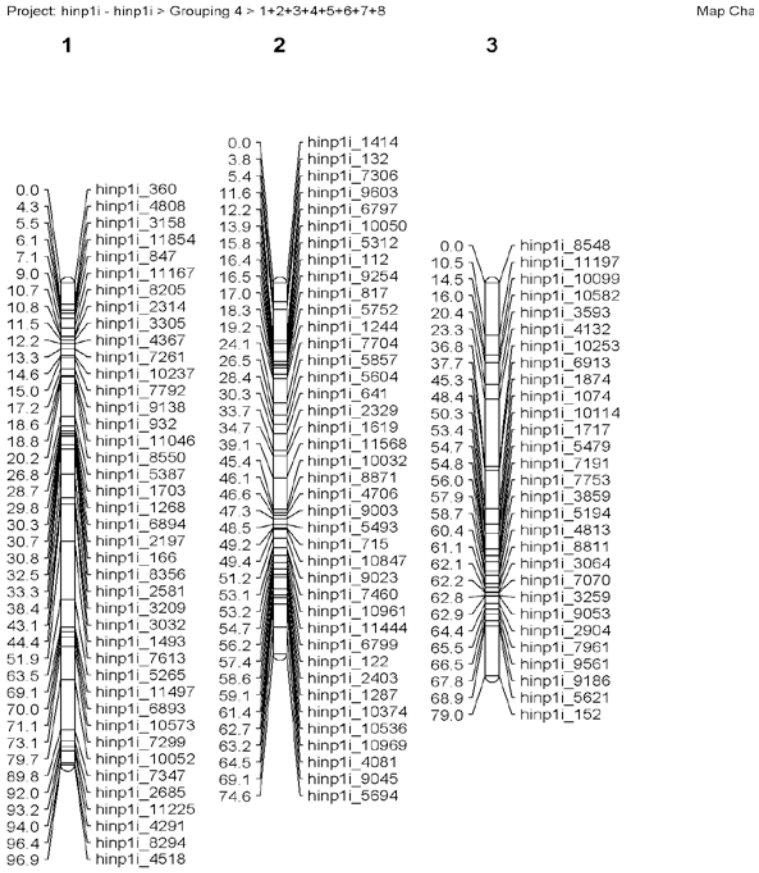


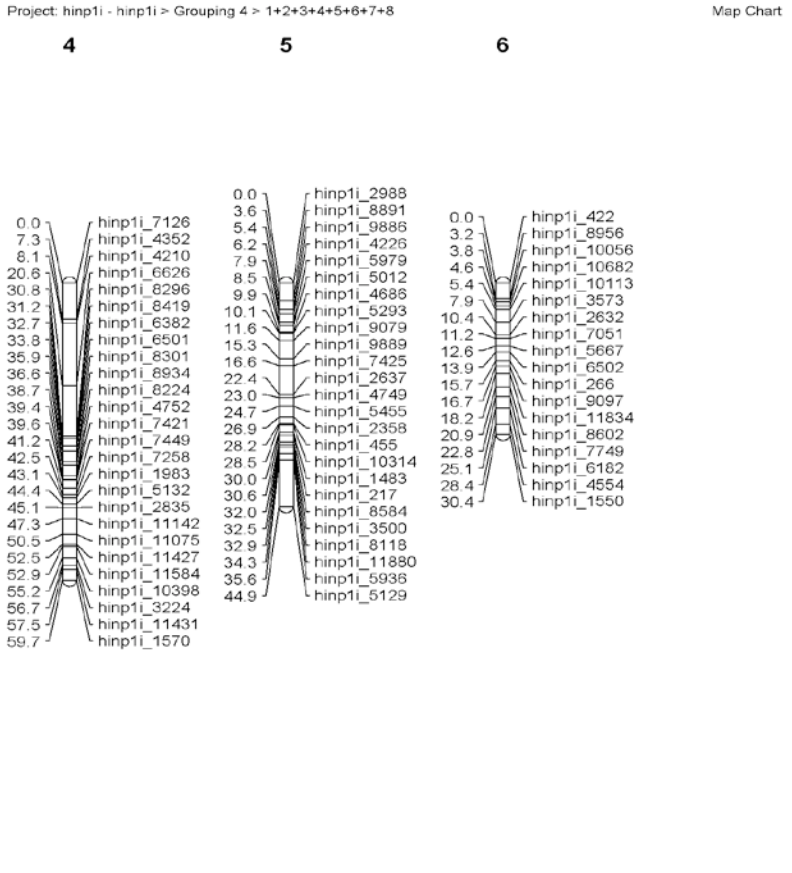




*Bsr*GI1_markers

Map examples





*Hin*P1I_markers

Summary and Conclusions

- Simplified Genotyping-by Sequencing (GBS) approach was used to create sequencing library using methylation not sensitive and methylation blocked enzymes separately.
- Stacks pipeline was used as a key bioinformatics tool.
- 2038 *de novo Bsr*GI SNPs are being mapped in the TM-1 x 3-79 RIL population
- Another 976 *de novo Hin*p1I SNPs are being mapped in the same population.
- *Hin*p1I markers were more dispersed, while *Bsr*GI were more centromeric.
- However, *Hin*p1I markers were more polymorphic, (contrary to our expectations).

Summary and Conclusions

- Project cost = ~$6,000 for sequencing
- ~$1.50 per polymorphic marker (discovered and mapped)
- Can be applied to any mapping population or association study with no prior sequencing necessary.
- Markers are sequence-based and can be mapped *in silico* to reference genomes.

Acknowledgements

- The CREST Award 2010 awarded to Surender Verma by Department of Biotechnology, Ministry of Science & Technology, Govt. of India is fully acknowledged.
- The technical support and experimental material rendered during the period of work by USDA-ARS and Texas A&M University are acknowledged.
- Special thanks to Dr. John Z. Yu, Dr. Richard G. Percy and Dr. Alan E. Pepper for hosting my fellowship study.

Thanks for

.....

kind attention