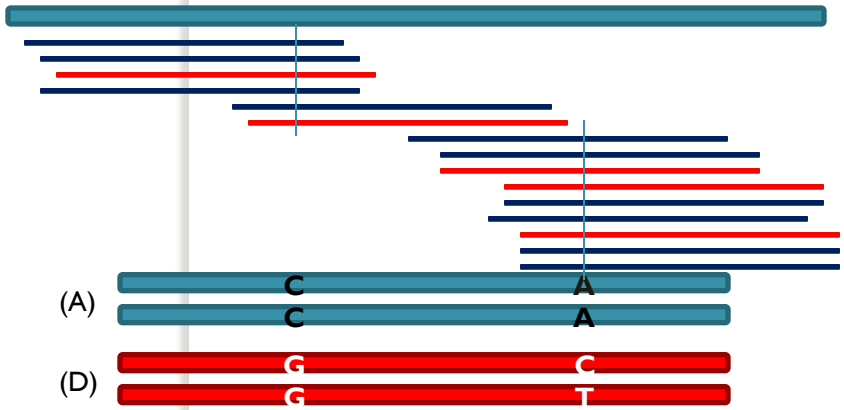


Comparison and Evaluation of Cotton SNPs Developed by Transcriptome, Genome Reduction on Restriction Site Conservation and RAD-based Sequencing

Hamid Ashrafi

Amanda M. Hulse, Kevin Hoegenauer, Fei Wang, Jeremy Schmutz, Andrew Paterson, Joshua A. Udall, David M. Stelly, and Allen Van Deynze

2012 ICGI Research Conference

Raleigh, North Carolina, USA

October 11, 2012

UC DAVIS UNIVERSITY OF CALIFORNIA

The University of Georgia

BYU

HudsonAlpha INSTITUTE FOR BIOTECHNOLOGY

NGS has facilitated discovery of a large number of SNPs in plants and animals

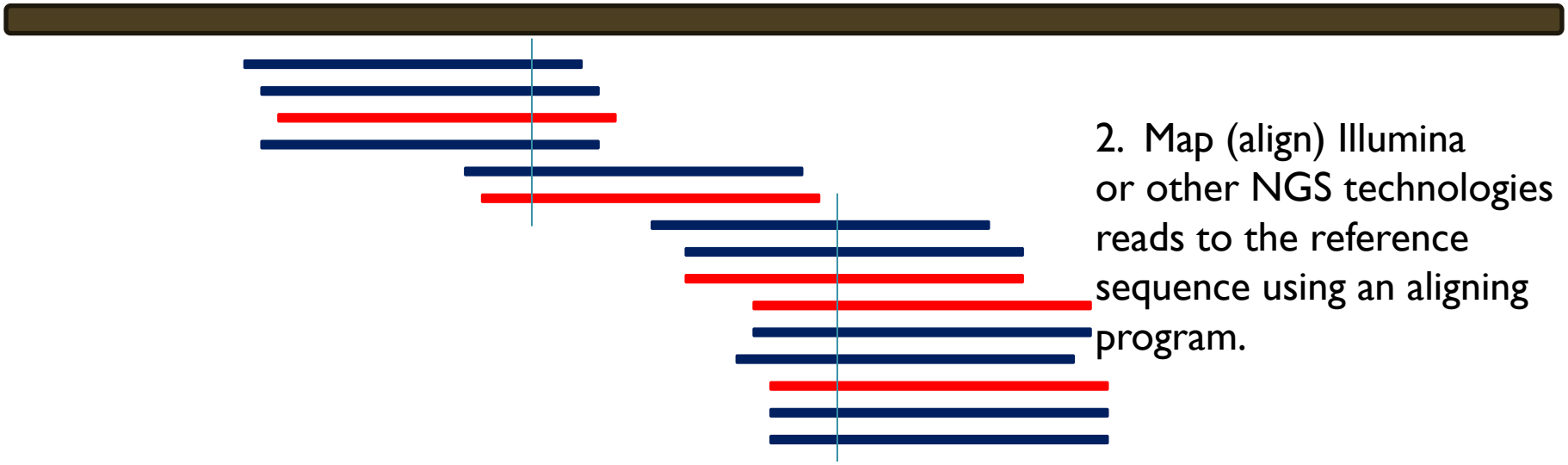
Various sequencing strategies target different regions of the genome.

- EST-based sequencing - targets transcribed regions of the genome.
- Gene-enriched sequencing uses methylation sensitive digestion.
- Genome reduction sequencing or Reduced Representation Libraries (RRL).

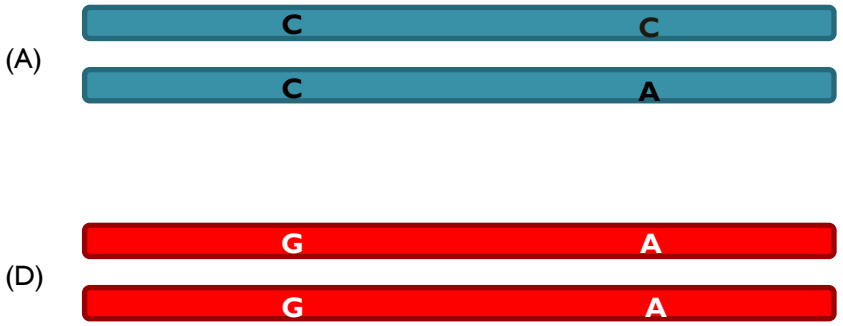
Byers et al, TAG 124, 1201-1214, 2012

Bioinformatically there are three steps involved in calling putative SNPs

1. Construct a reference sequence (*de novo*) or use an existing one



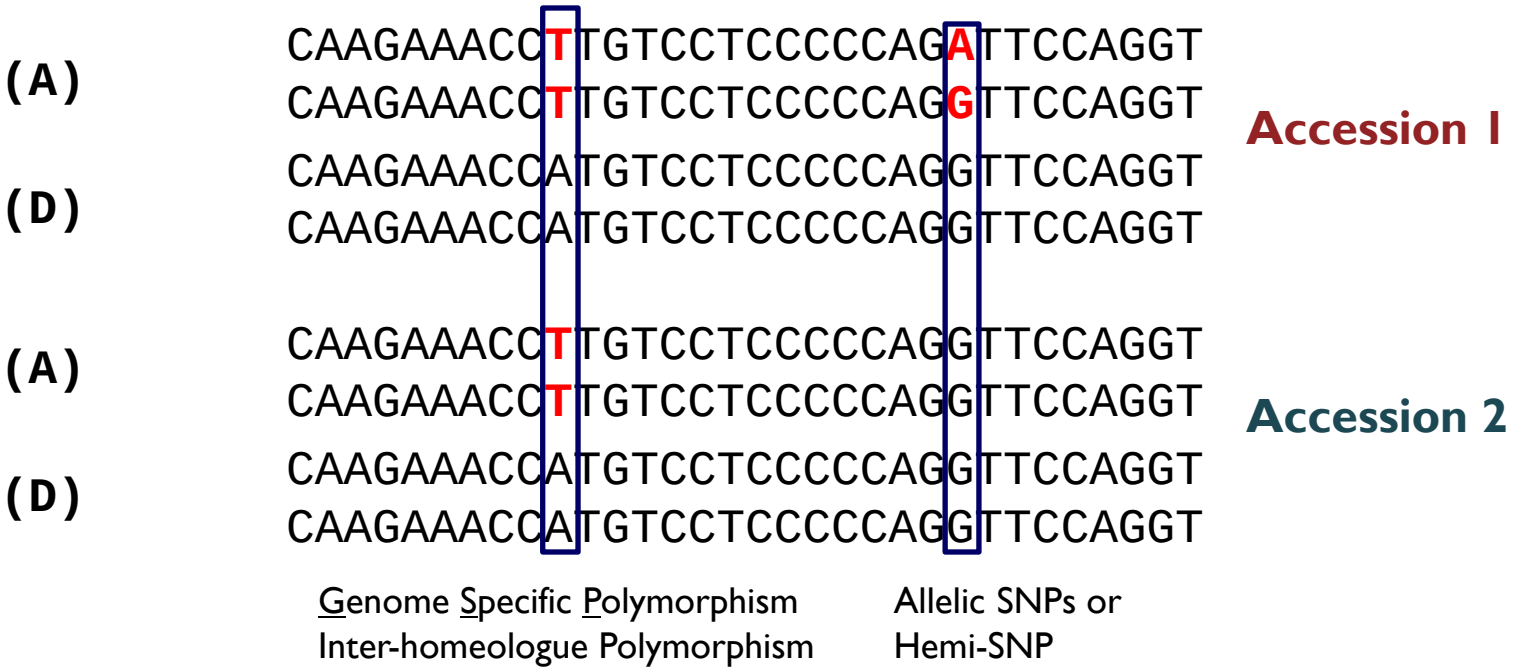
2. Map (align) Illumina or other NGS technologies reads to the reference sequence using an aligning program.



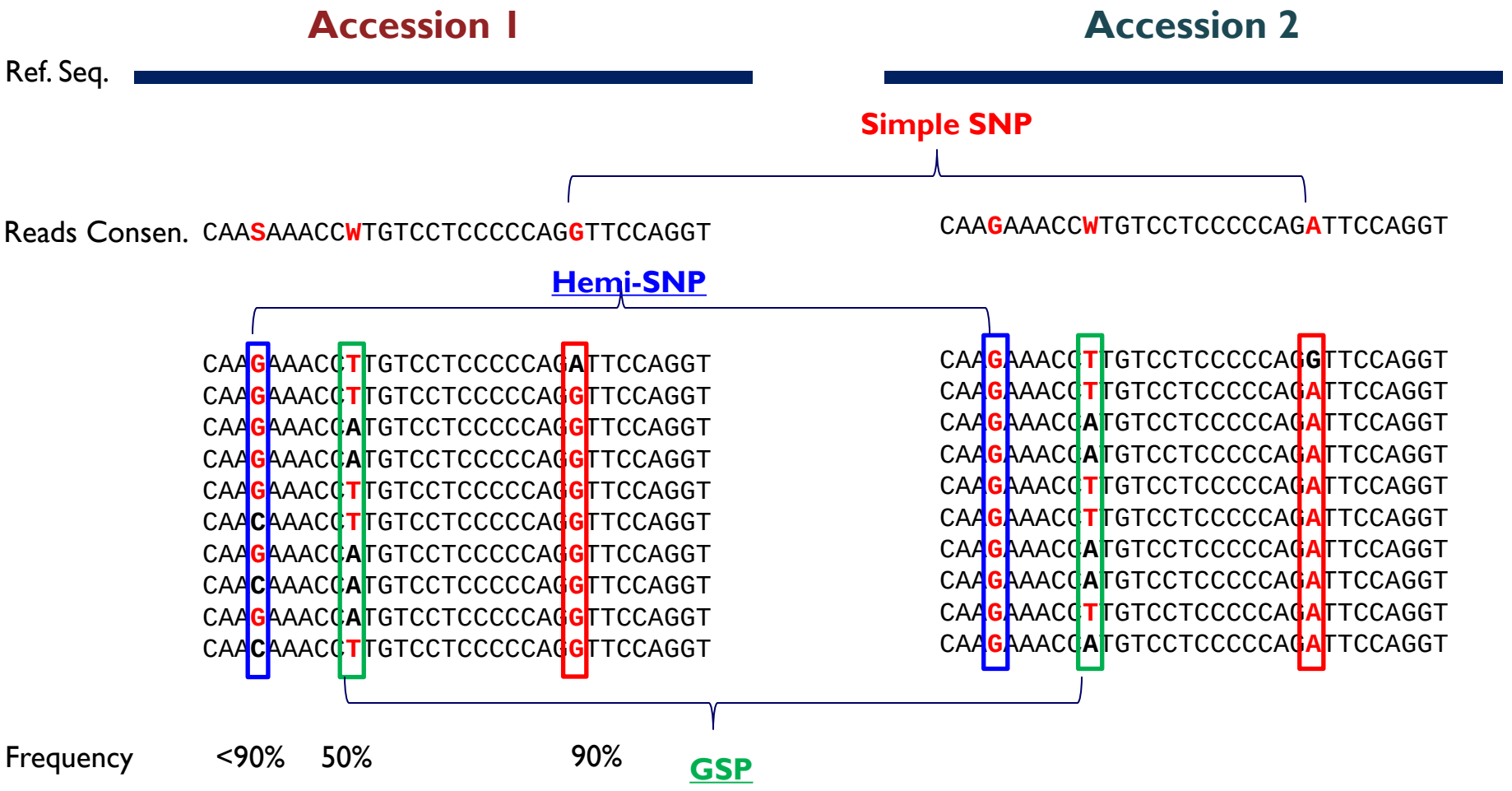
3. Call the putative SNPs based on quality of the reads (statistics), depth, allele frequency,

How to detect putative SNPs in allotetraploid cotton?

- There are factors that complicate SNP detection in cotton genome
  - Polyploidy (homeologous mutations)
  - High frequency of duplicated genes (paralogous mutations)



How can we define, GSP, simple, and hemi-SNPs using NGS reads of allotetraploid cotton?



Cotton SNP Discovery Based on Transcriptome Assembly of *Gossypium hirsutum cv. TM-1 (UCD assembly)*

G. hirsutum (TM-1) Sequence Sources

- cDNA library of TM-1 and Illumina Sequencing
- cDNA library of TM-1 and 454 Sequencing
- GenBank EST sequences of TM-1

A hybrid assembly of 454 and Sanger EST sequences was constructed (MIRA Software)

Reads assembled:	1,435,805
GenBank EST Sequences:	32,576
Avg. total coverage:	7.83
Number of contigs:	64,113
Total consensus nt:	62,8Mb
Largest contig:	13,697
N50 contig size:	1,168
Max coverage:	593

SNP detection

- BWA and CLC software packages were used to map Illumina reads of genotypes to the assembly.
- SAMtools was used to call the putative SNPs.
- In house Perl scripts were used to filter putative SNPs.

TM-1 transcriptome assembly was used as a reference to identify SNP in the following genotypes (UCD data)

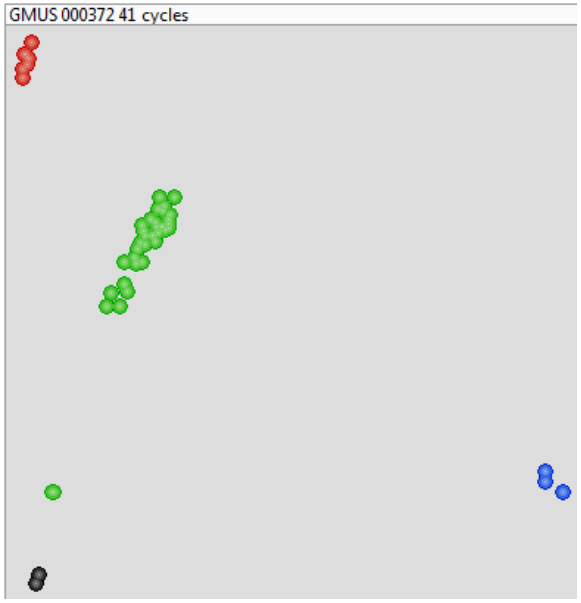
Genotype/Species	Genome	Class I SNPs	Class II SNPs	Class III SNPs
G. hirsutum TM-1 Acala PD-1 Sealand FiberMax	$(AD)_1$	764	428	4
TM-1 and Acala			Simple SNPs 511, Hemi=96,392	
G. barbadense 3-79	$(AD)_2$	3,257	6,386	1,246
G. tomentosum	$(AD)_3$	1,520	6,526	1,474
G. mustelinum	$(AD)_4$	1,678	7,584	1,726
G. armourianum	(D_{2-1})	7,331	14,523	5,120
G. longicalyx	(F_1)	14,577	18,960	4,711

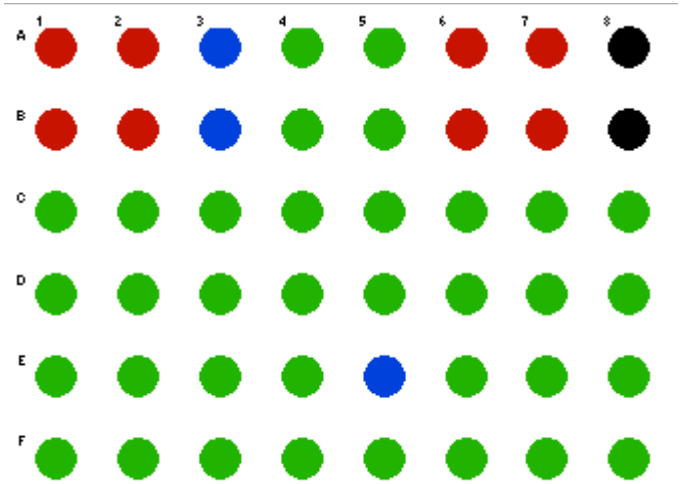
Coverage 10 and minor allele frequency 90%

Class I = SNPs in contigs with no GSP
Class II = SNPs in contigs with GSP but not in the vicinity of 50 bases of the SNPs
Class III = SNPs in contigs with GSP and one GSP in the vicinity of 50 bases of the SNP

In a validation experiment 100 putative *G. mustelinum* SNPs used in KASPar assay

SNP Panel 6 for G. mus	1	2	3	4	5	6	7	8
A	TM-1 (USDA)	TM-1(WAR)	G.mus	F1 (TM1 X mus)	F1 (mus X TM1)	3--79	G.tom	H2O
B	TM-1 (USDA)	TM-1(WAR)	G.mus	F1 (TM1 X mus)	F1 (mus X TM1)	3--79	G.tom	H2O
C	F1 (H01 X mus)	F1 (Te 15 LO X mus)	F1 (H06 X mus)	F1 (H25 X mus)	F1 (H7 X mus)	F1 (H16 X mus)	F1 (H3 X mus)	F1 (H17 X mus)
D	F1 (H01 X mus)	F1 (Te 15 LO X mus)	F1 (H06 X mus)	F1 (H25 X mus)	F1 (H7 X mus)	F1 (H16 X mus)	F1 (H3 X mus)	F1 (H17 X mus)
E	F1 (H2 X mus)	F1 (H4 X mus)	F1 (Te 5 LO X mus)	F1 (Te 8 LO X mus)	F1 (Te 11 LO X mus)	F1 (Te 12 LO X mus)	F1 (Te 12 SH X mus)	F1 (H12 X mus)
F	F1 (Te 14 LO X mus)	F1 (H18 X mus)	F1 (Te 20 LO X mus)	F1 (Te 20 SH X mus)	F1 (Te 11 SH X mus)	F1 (Te 22 LO X mus)	F1 (Te 22 SH X mus)	F1 (Te 26 SH X mus)

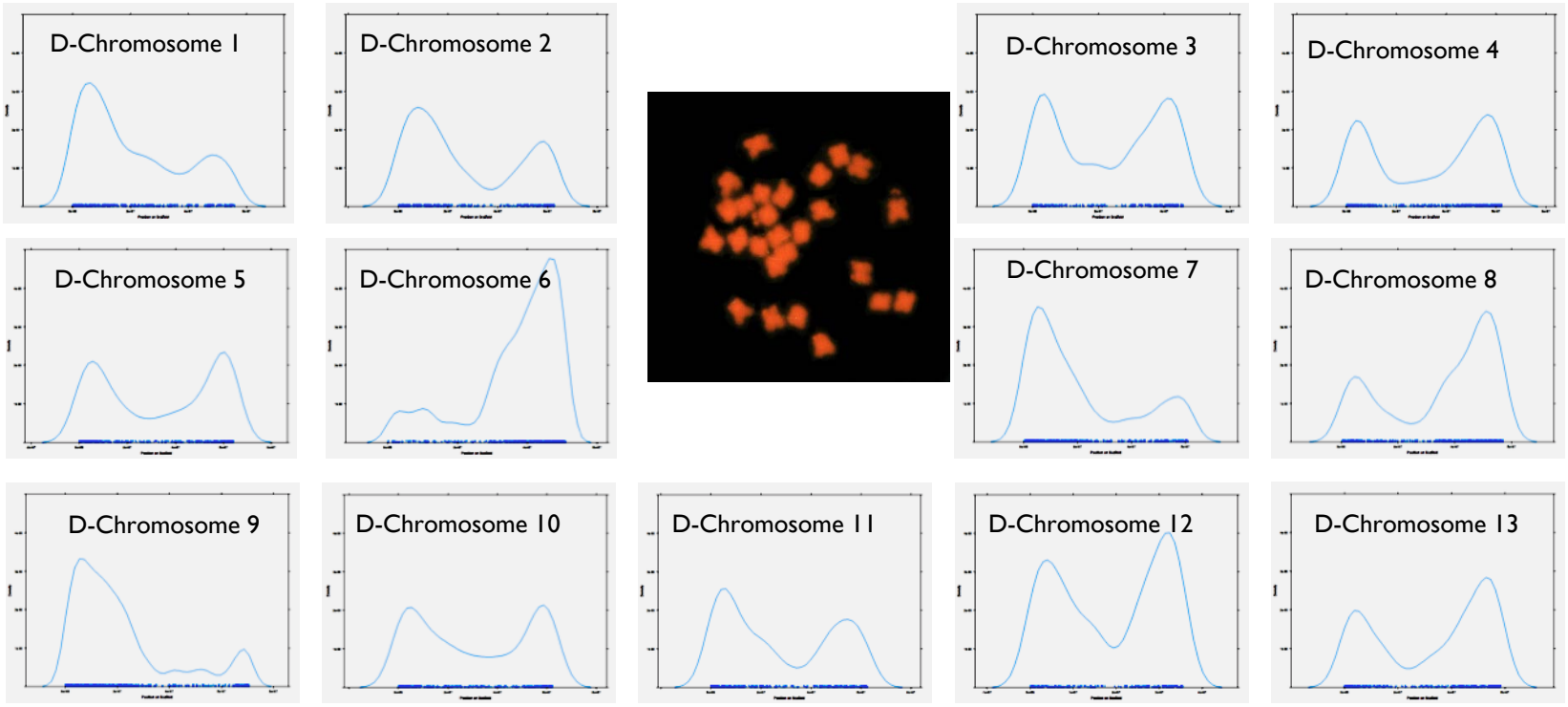




Two parents polymorphism= 62/100

The Plate Polymorphism= 71/100

Density plots depict the abundance of mapped SNPs across the D5 genome scaffolds. Distributions ~match expectations for (sub)metacentrics



Cotton SNPs Based on Genome Reduction on Restriction Site Conservation (BYU data)

Methodology:

- Double digest the DNA with rare and frequent restriction enzymes (2 enzyme system, *Eco* RI & *Bfa* I)
- Add ligation adaptors and end label the end of 6-base recognition site with 5'-biotin molecule. Leave the 4-base recognition site unlabeled.
- Multiplexing can be achieved by adding bar code sequences using PCR with different primers complementary to the adapter sequence.
- If several individuals are supposed to be sequenced, equimolar amount of each can be pooled together and further size selection can be done by electrophoresis and gel isolation.

Maughan et al. 2009 Plant Genome

Summary of GR-RSC SNP Discovery

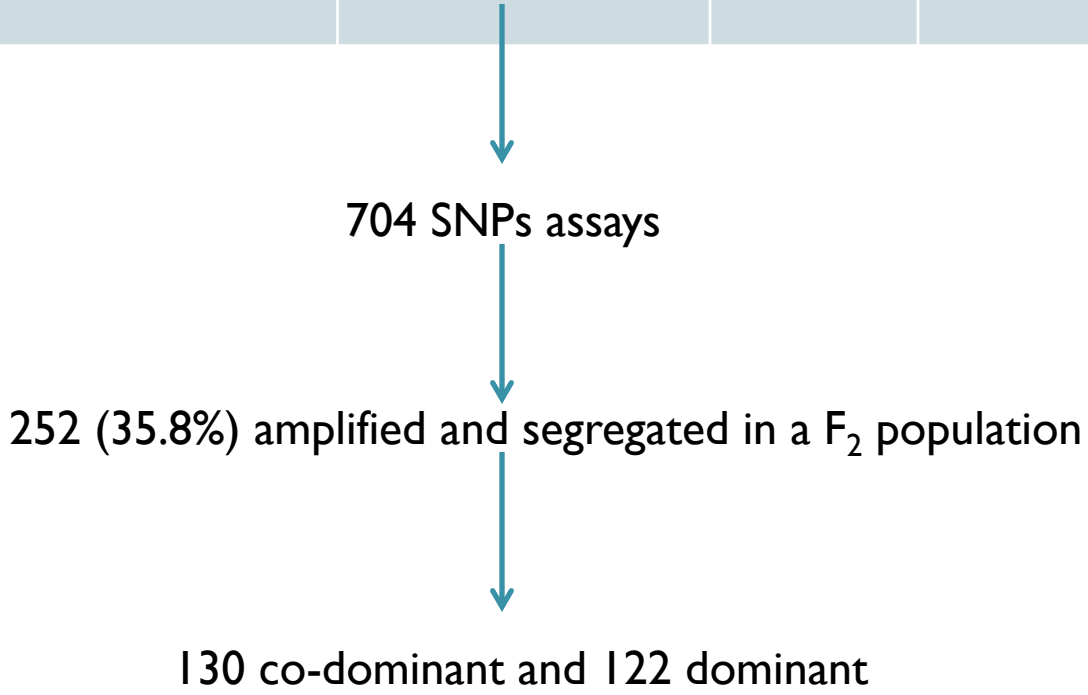
Category	Assembly	Accessions	SNPs	Contigs with SNPs	SNPs / contig
By individual	*G. hirsutum*	Acala	45,590	8660	5.3
	G. hirsutum	TX2094	42,166	8201	5.1
	G. barbadense	Pima-S6	26,662	4934	5.4
	G. barbadense	K101	29,420	5455	5.4
Between Accessions	*G. hirsutum*	Acala and TX2094	11,834	6469	1.8
	Reduced G. hirsutum	Acala and TX2094	4,045	2176	1.9
	G. barbadense	Pima-S6 and K101	1,679	965	1.7

Coverage 8 and minor allele frequency 80%

Byers et al, TAG 124, 1201-1214, 2012
With permission

Validation of GR-RSC SNPs

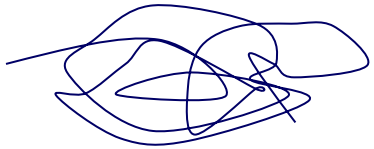
Category	Assembly	Accessions	SNPs	Contigs with SNPs	SNPs / contig
Between Accessions	*G. hirsutum*	Acala and TX2094	11,834	6469	1.8



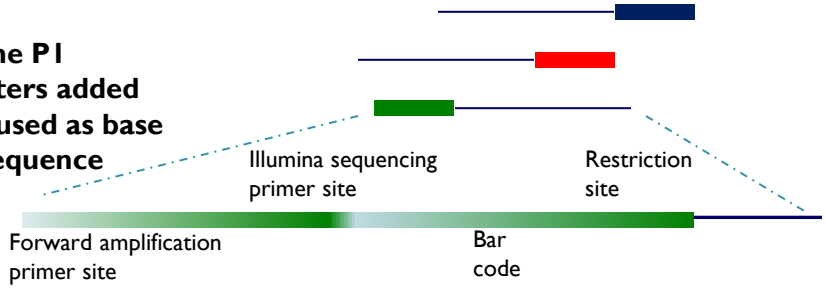
Byers et al, TAG 124, 1201-1214, 2012
With permission

Cotton SNPs Based on Restriction Site Associated DNA Sequencing (RADseq SNP discovery)

A) Genomic DNA is digested with one RE (ie. Pst I).

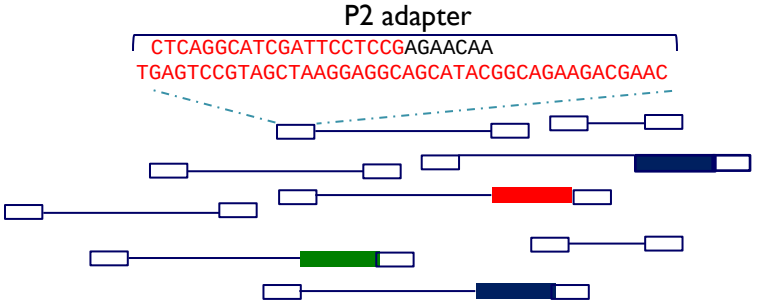


B) The P1 adapters added and used as base for sequence



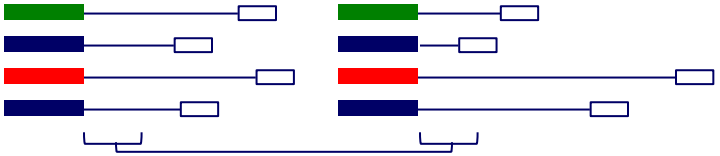
C) Pool bar coded samples and shear

D) Ligate P2 adapter to sheared fragments



The fragments undergo PCR amplification using Illumina primers and gel purified again to the size of 300-700 bp followed by Illumina sequencing

E) Selectively amplify RAD tag



Adapted from Baird et al. PLOS one 2008

Cotton SNPs Based on Restriction Site Associated DNA Sequencing (RADseq SNP discovery)

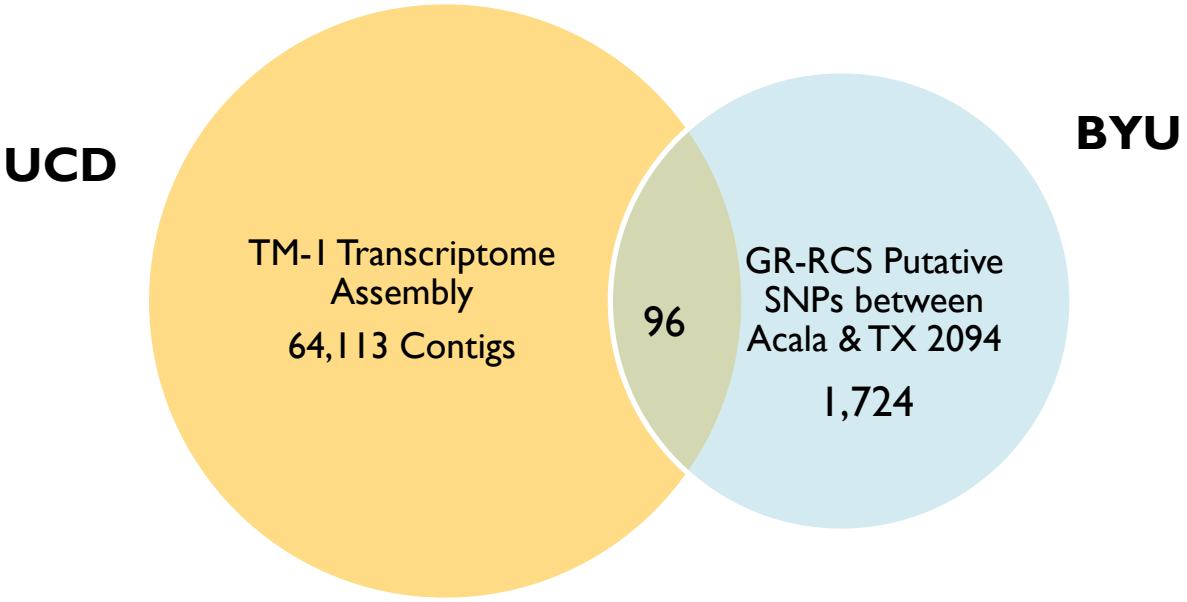
De novo Assembly Results for TM1:	81,261 contigs
N50 contig length:	463 bp
Assembled nt:	34.2 Mb

SNP Discovery between Tm-1 and Acala

52,190	Putative variant alleles (with minimum 14x seq coverage in both samples).
1,460	Simple SNPs (Alleles fixed / homozygous in each cultivar).
1,827	Hemi-SNPs (Allele fixed in one cultivar, heterozygous in other).

Comparison of SNPs identified by GR-RCS vs. transcriptomes assembly of TM-1

Approx. what percentage of GR-RCS putative SNPs are genic or non-genic?



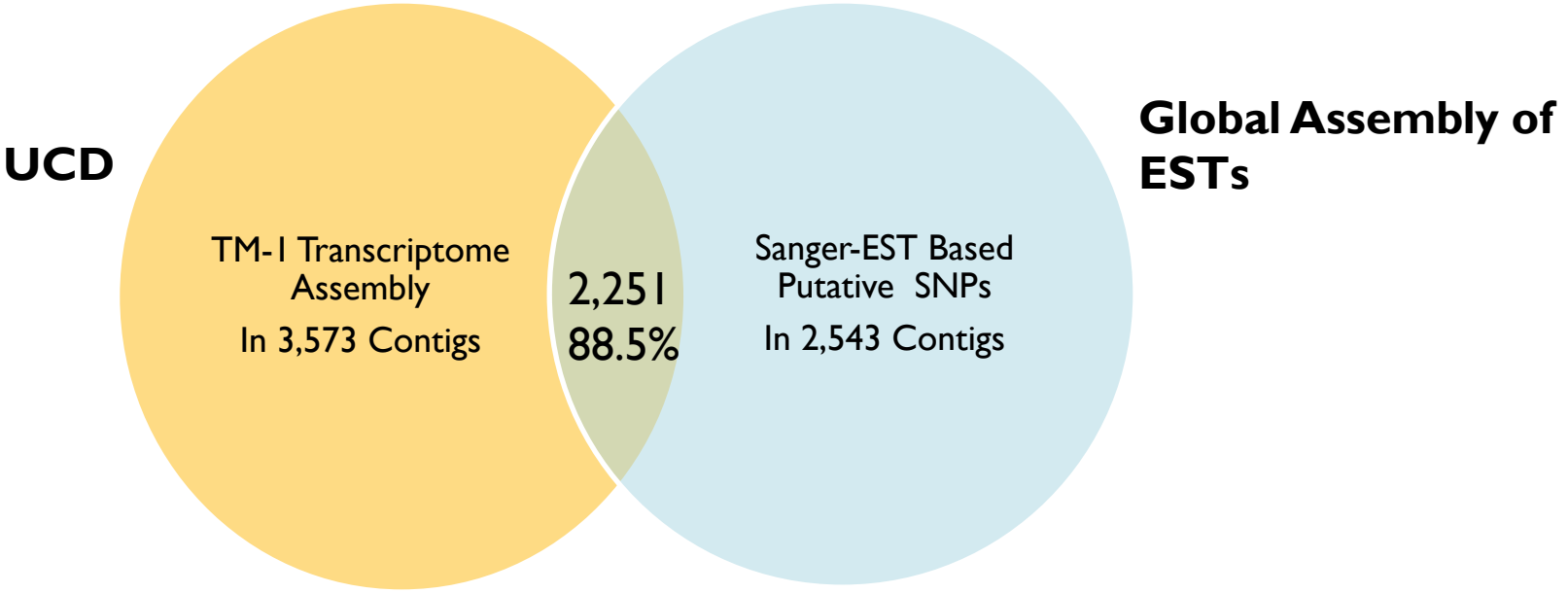
BLAST a sample of GR-RCS SNPs (101 nt) against TM-1 Assembly.

5.5% genic, 94.5% non-genic

Comparison of TM-1 transcriptome SNPs vs. SNPs identified in a global assembly of cotton ESTs

Udall et al. Genome Research 16(3) 2006

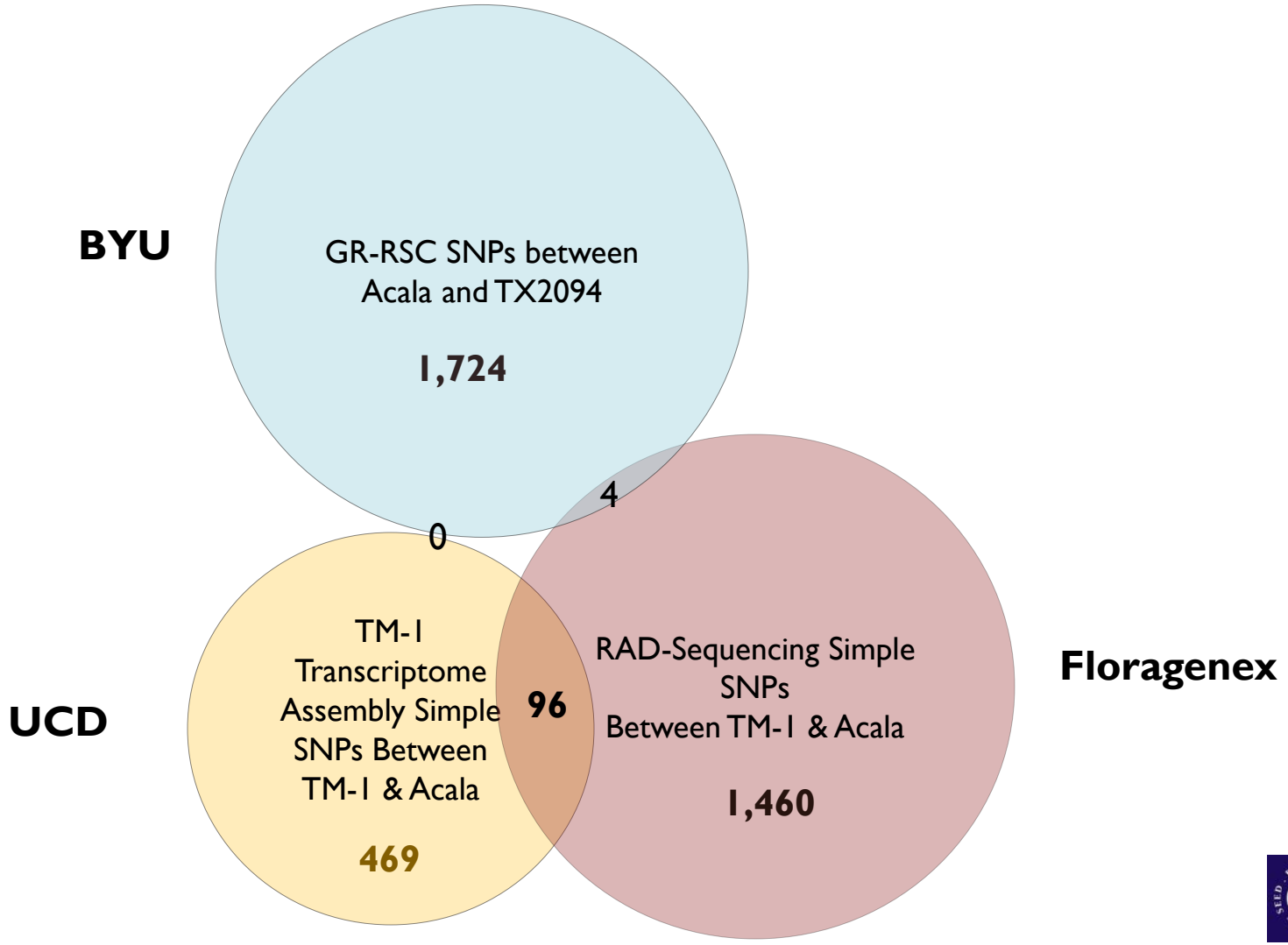
Approx. what is the overlap between SNPs identified by UCD and global assembly of cotton ESTs?



Comparison of SNPs identified across all three methods

Direct comparison of simple SNPs identified by each strategy.

Flanking sequence around the putative SNPs were aligned against each other.



Summary

- Using UCD assembly of TM-1 we identified:
  - -intraspecific SNPs among 5 *G. hirsutum* cultivars.
  - -fewer intraspecific SNPs between Acala and TM-1 than RAD-seq.
  - -more Hemi-SNPs than RAD-seq.
  - -more interspecific putative SNPs in other AD, D and F genomes.
- Validation of UCD SNPs showed the utility of them in breeding programs for cotton.
- Comparison of RAD-Seq and GR-RSC with UCD transcriptome assembly showed that less than 10% of SNPs identified by those technologies are genic and the remaining are non-genic.
- In terms of percentage overlap with UCD transcriptome assembly, GR-RSC and RAD-seq were comparable. Though BYU SNPs were slightly more non-genic than genic compared to RAD-seq SNPs.
- Validation of RAD-Seq SNPs is now underway and once completed, it allows us to determine the efficiency and accuracy of algorithms used to filter the SNPs by Floragenex.
- Comparison of RAD-Seq and GR-RSC showed they have almost no overlap.

Acknowledgements

UC DAVIS

Allen Van Deynze

Kevin Stoffel

Alex Kozik

Texas A&M University

David Stelly

Amanda Hulse

Kevin Hoegenauer

Fei Wang

Brigham Young University

Joshua Udall

Robert Byers

Floragenex (http://www.floragenex.com)

Rick Nipper

The University of Georgia

Andrew Paterson

The Center for Genomics and Bioinformatics – Indiana University

Keithanne Mockaitis

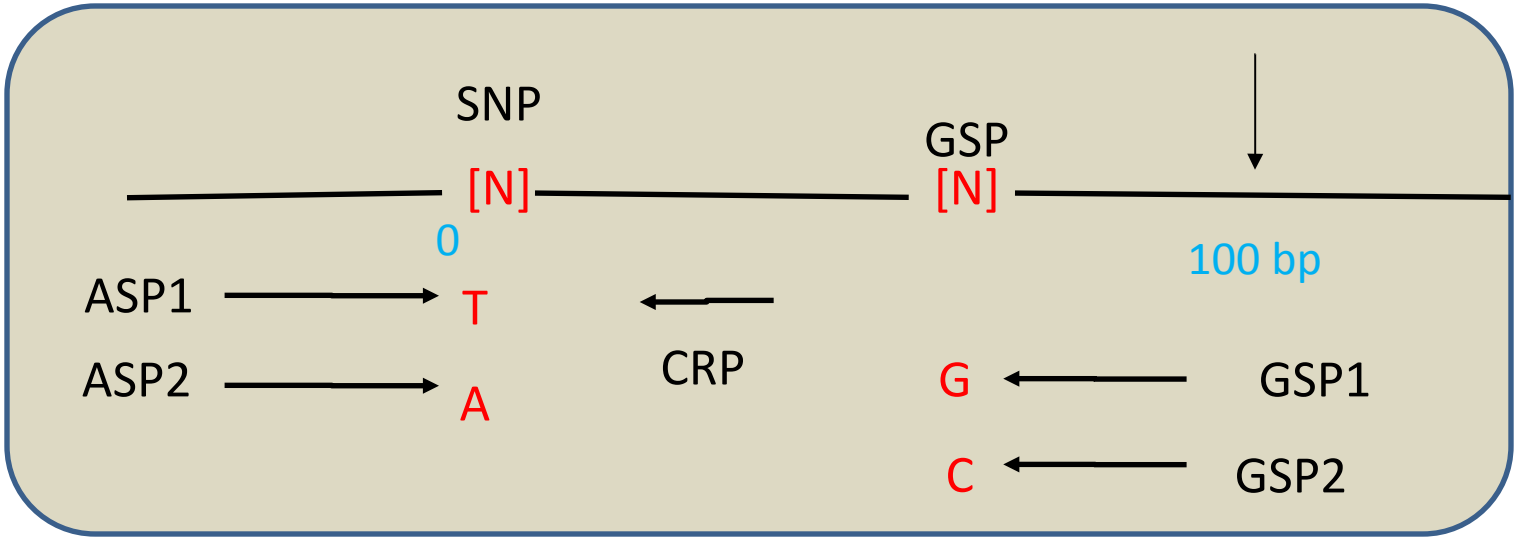
UC Davis Genome Center

Funding Agencies: Cotton Incorporated, Texas A&M Agrilife Research and extension, NSF Plant Genome Program

UCDAVIS
UNIVERSITY OF CALIFORNIA

UCDAVIS
UNIVERSITY OF CALIFORNIA

BIOTECHNOLOGY CENTER


SNP
[N]
0
GSP
[N]
100 bp
ASP1
T
ASP2
A
CRP
G
GSP1
C
GSP2