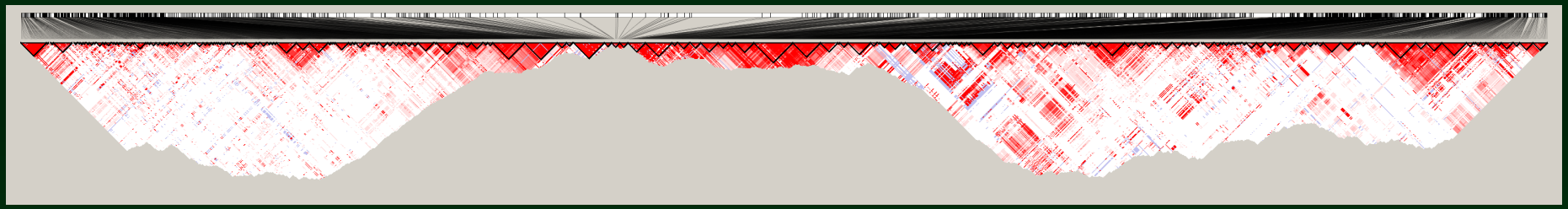
Update on SNP Genotyping of the USDA Soybean Germplasm Collection

David L. Hyten, Qijian Song, Gaofeng Jia,
Randall L. Nelson, Vincent R. Pantalone,
James E. Specht, and Perry B. Cregan



United States Department of Agriculture
Agricultural Research Service

Beltsville Area / Soybean Genomics and Improvement Laboratory

The completed soybean genome draft sequence is only a first step

Soybean Genome is 1.1 billion DNA base-pairs

20 chromosomes	**Chr. Ave. 55 million bases long**

Soybean Genome Project, DoE Joint Genome Institute (Schmutz et al., 2010, Nature, 463:178-183)

Assembled and anchored 950 million bases	**Predicted ~46,430 protein-coding genes**

Next steps

Improve soybean reference sequence	**Define genetic variation in germplasm and correlate with agronomically important traits**

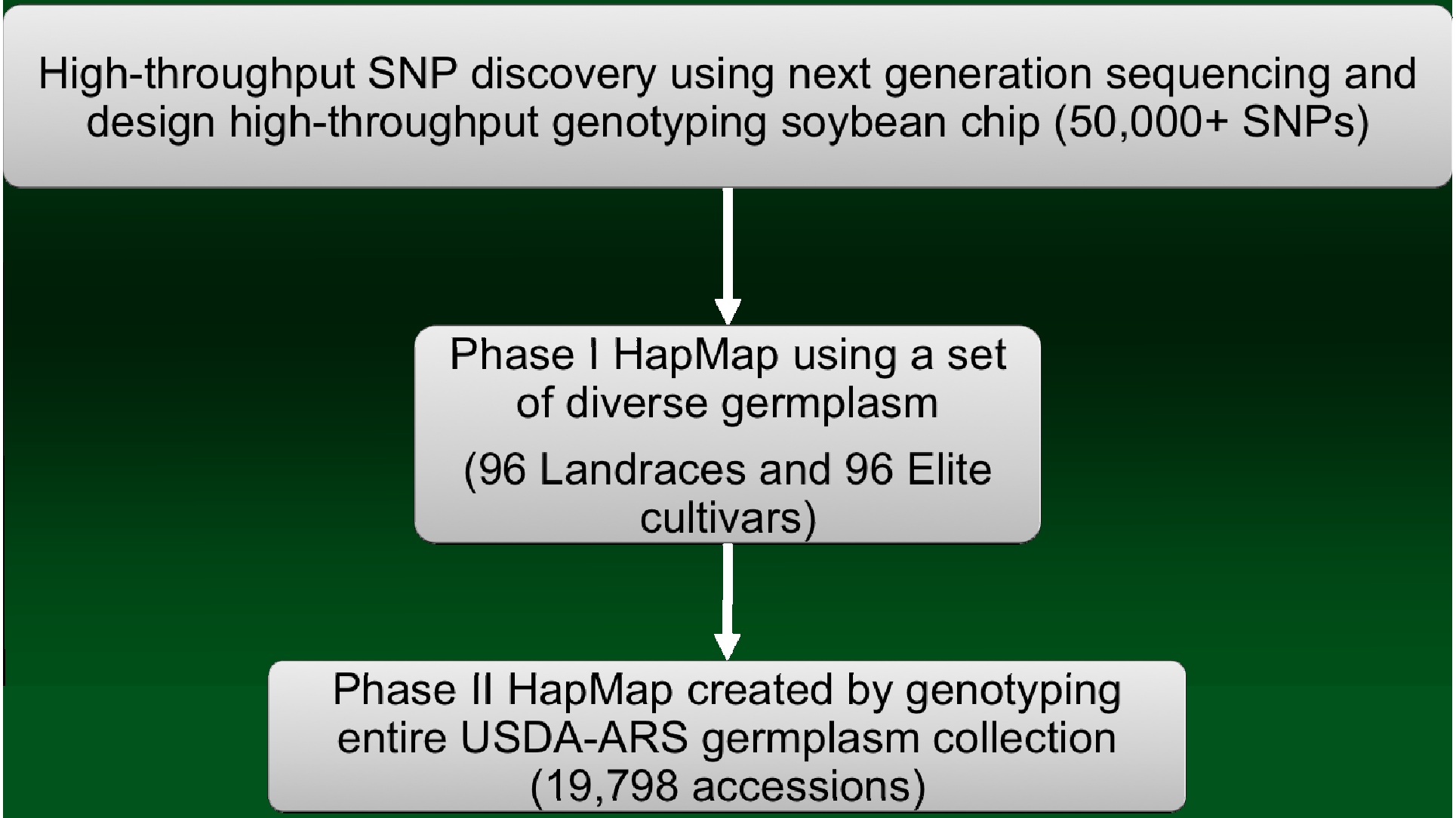
Creating a next generation map that charts out soybean diversity (Soy HapMap)

Determine population structure of germplasm collection

Find signatures of selection

Associate haplotypes with traits

An overview of the soybean HapMap project at the USDA



Discovered 177,347 SNPs using 21 gigabases of next generation sequence from multiple reduced representation libraries with a validation rate of 86%

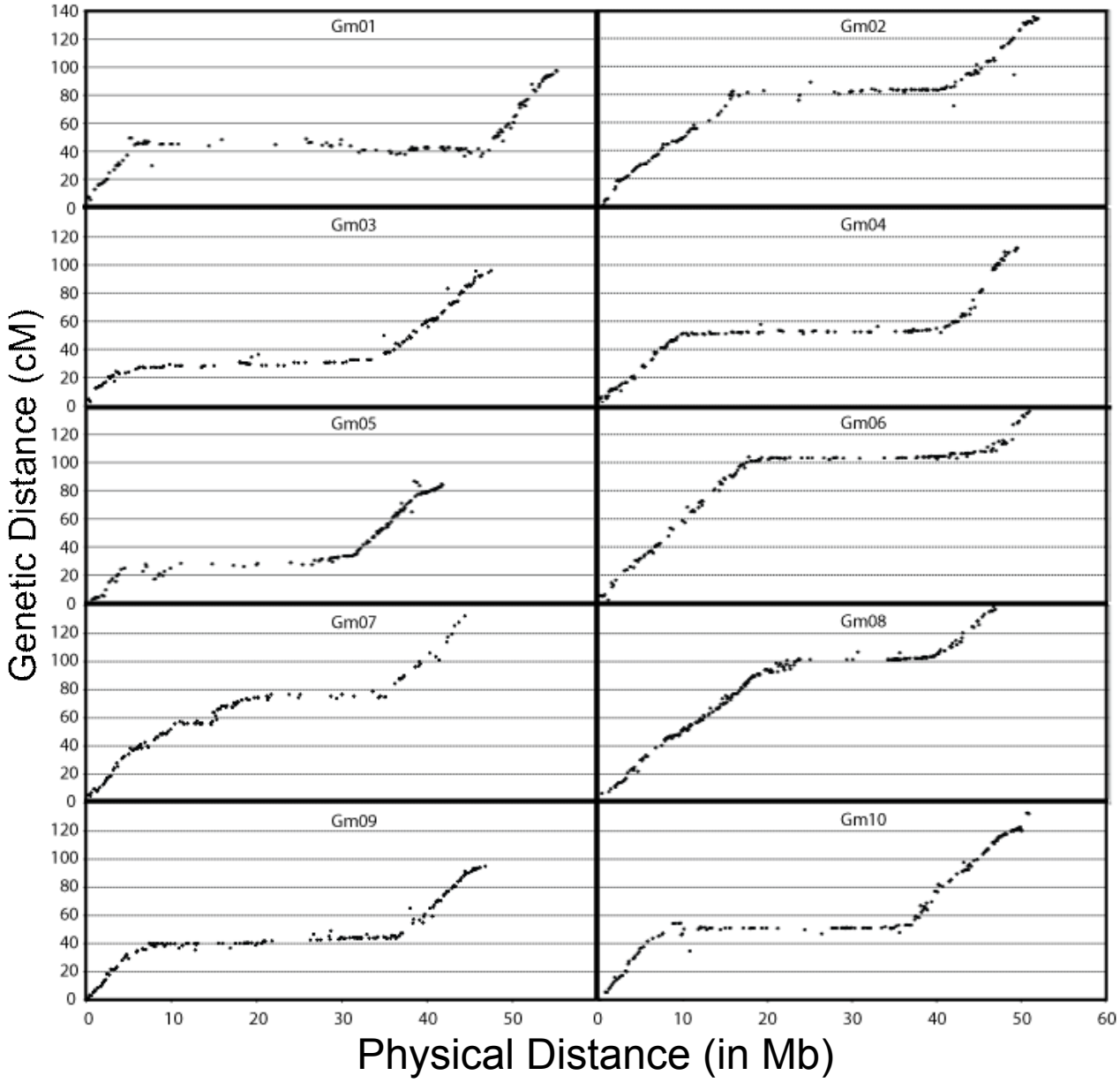
Williams 82 whole genome reference sequence

Obtained a total of 542 million Illumina GA reads

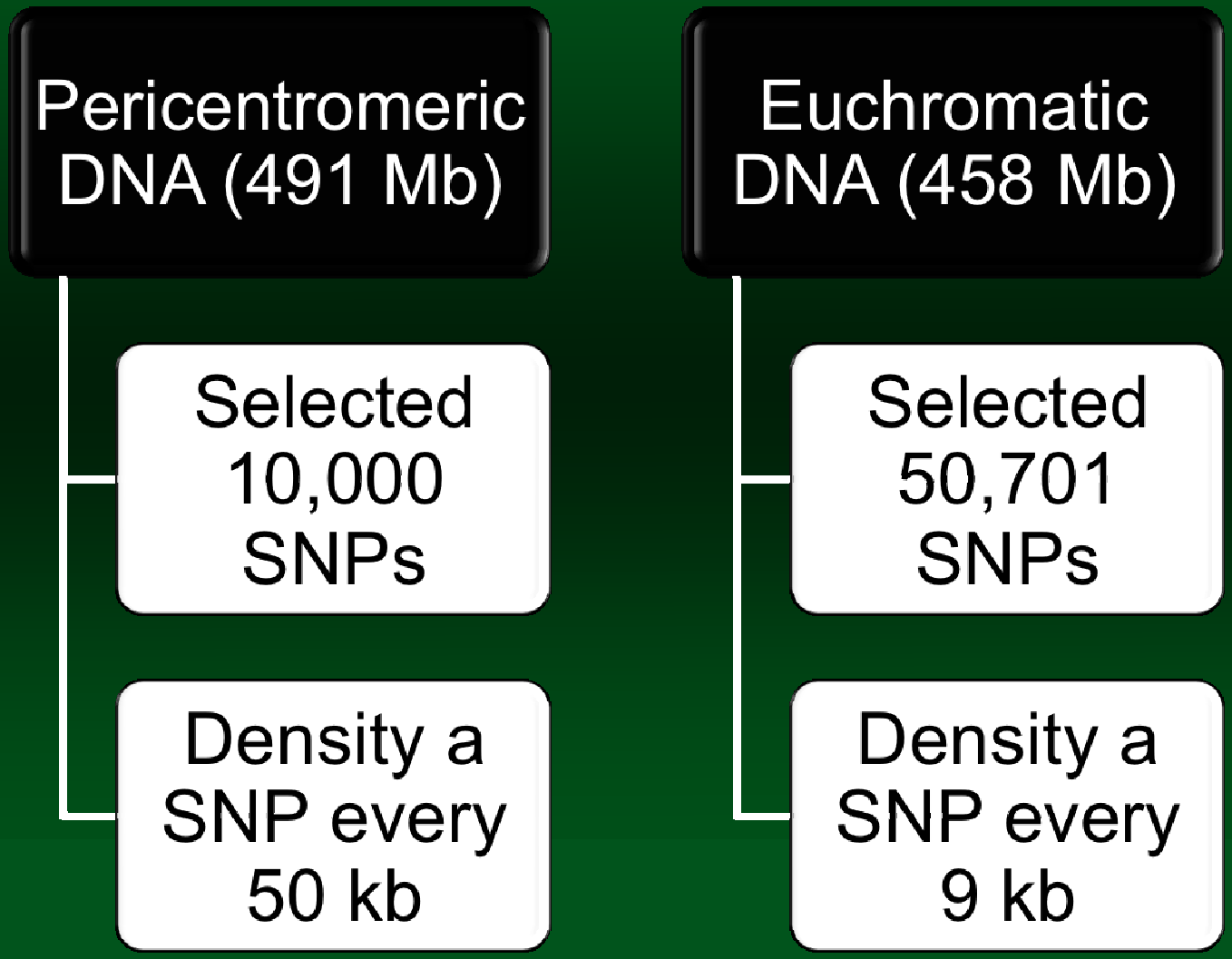
Illumina GA reads ranged from 35 to 50 bp in length

Most recombination in soybean occurs in the euchromatic DNA

- 93% recombination occurs in the euchromatic DNA (43% of the genome)
- Genetic-to-physical ratios
  - 197 kb/cM for euchromatic DNA
  - 3.5 Mb/cM for pericentromeric regions
- Data and figure from Schmutz et al. 2010, *Nature,* 463:178-183

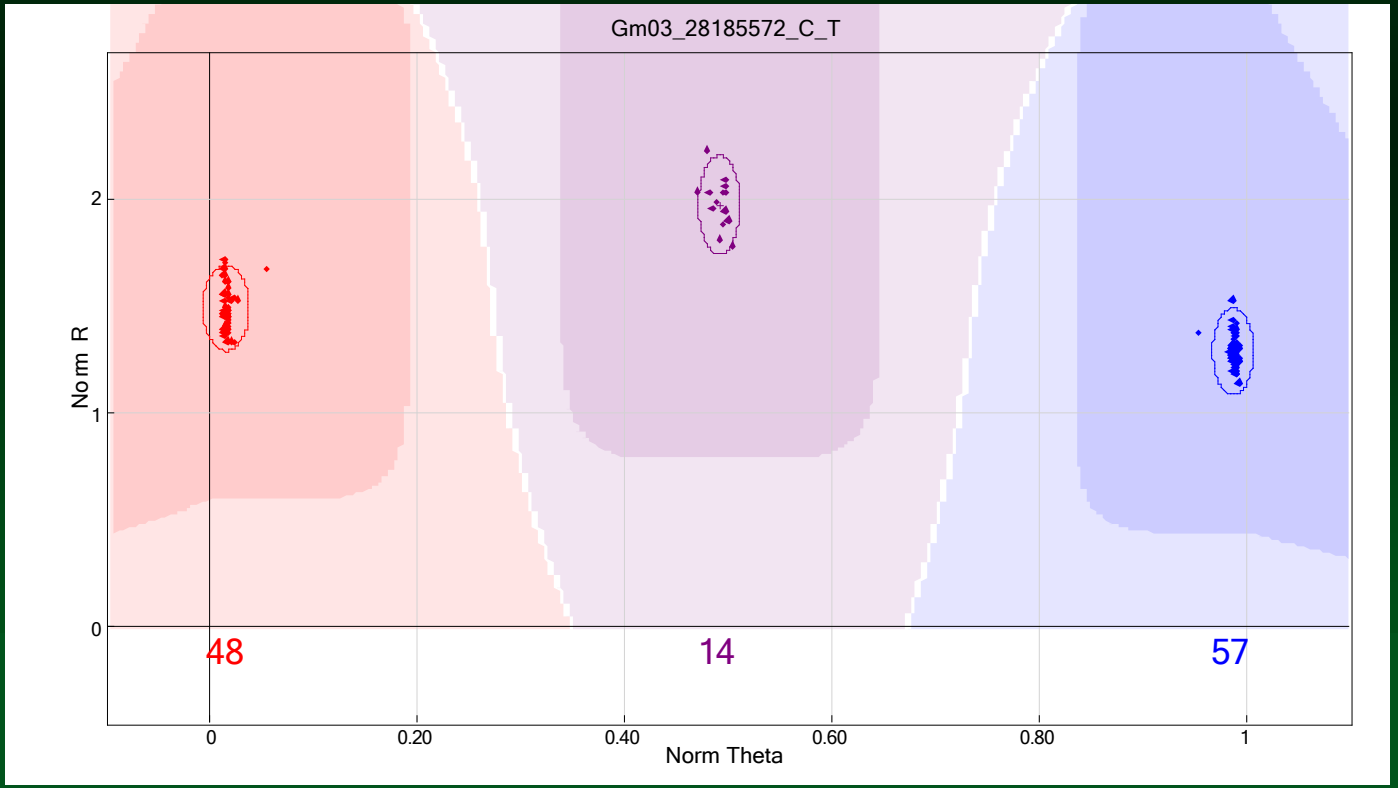


The 60,800 SNPs selected for the Infinium chip were mostly located within euchromatic DNA

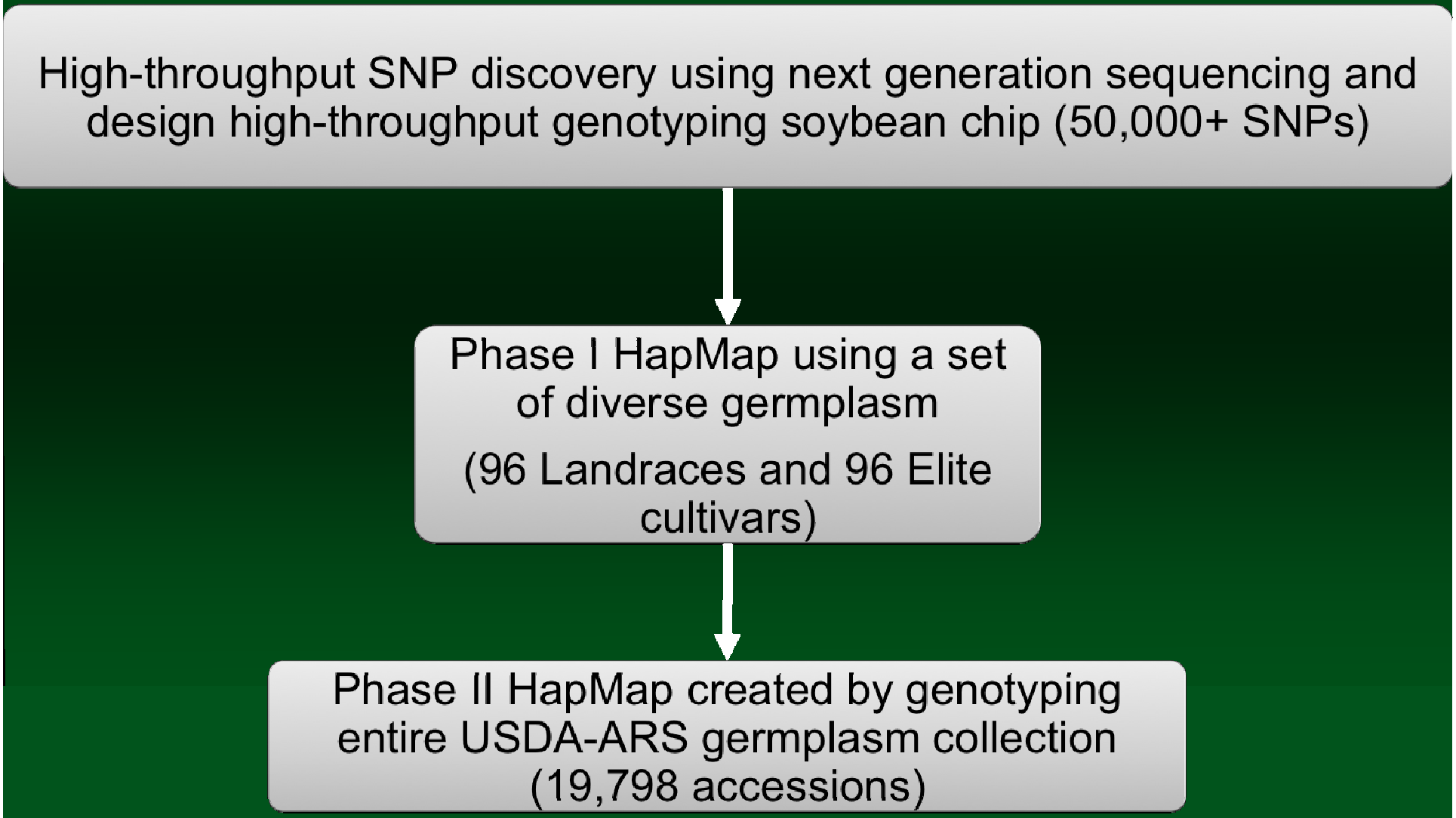


SoySNP50

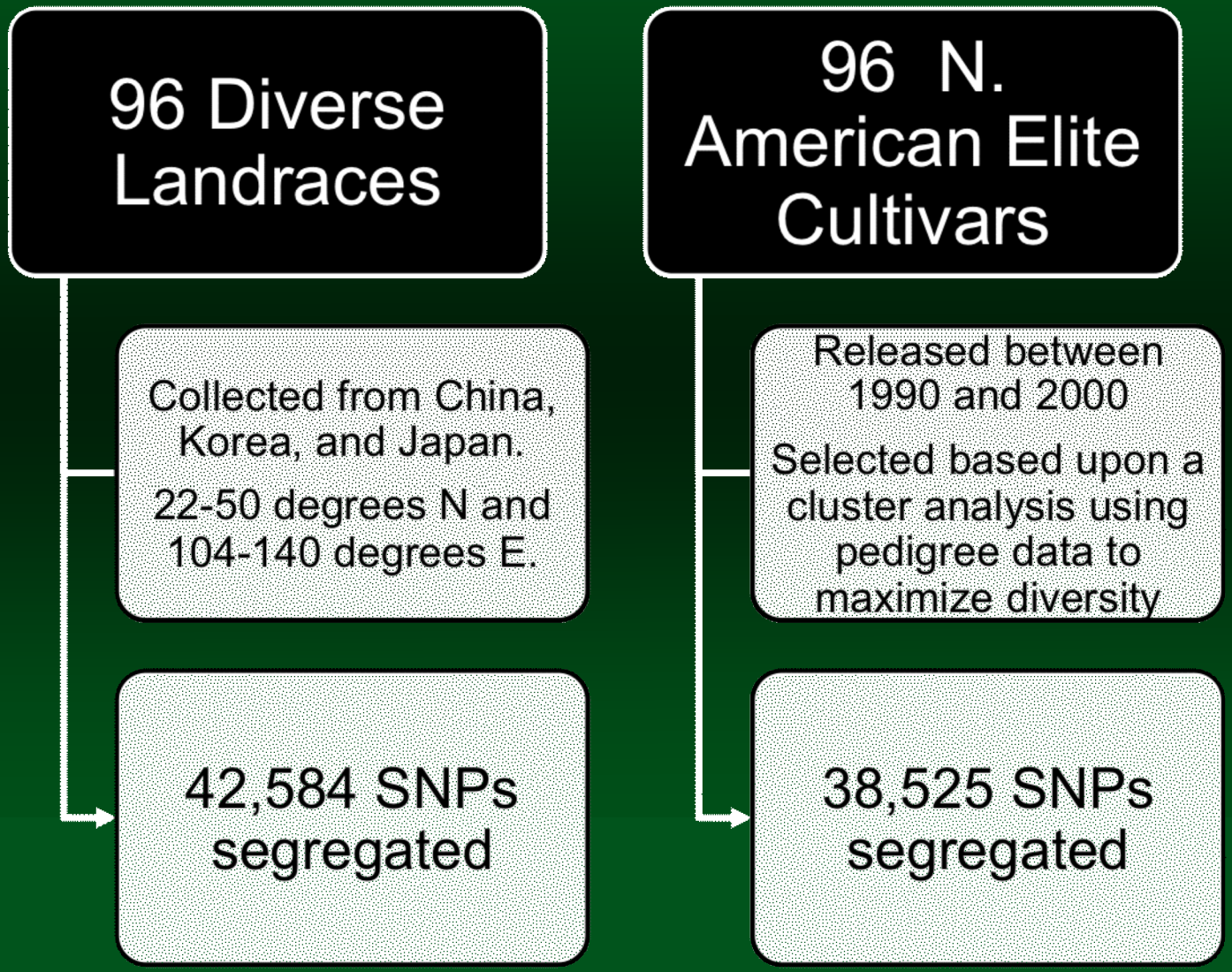
- Contains 52,041 SNPs (out of 60,800 submitted to Illumina) on a 24 DNA sample chip
  - 8,759 SNPs failed Illumina's manufacturing level (14%)
  - 46,735 gave polymorphic marker data



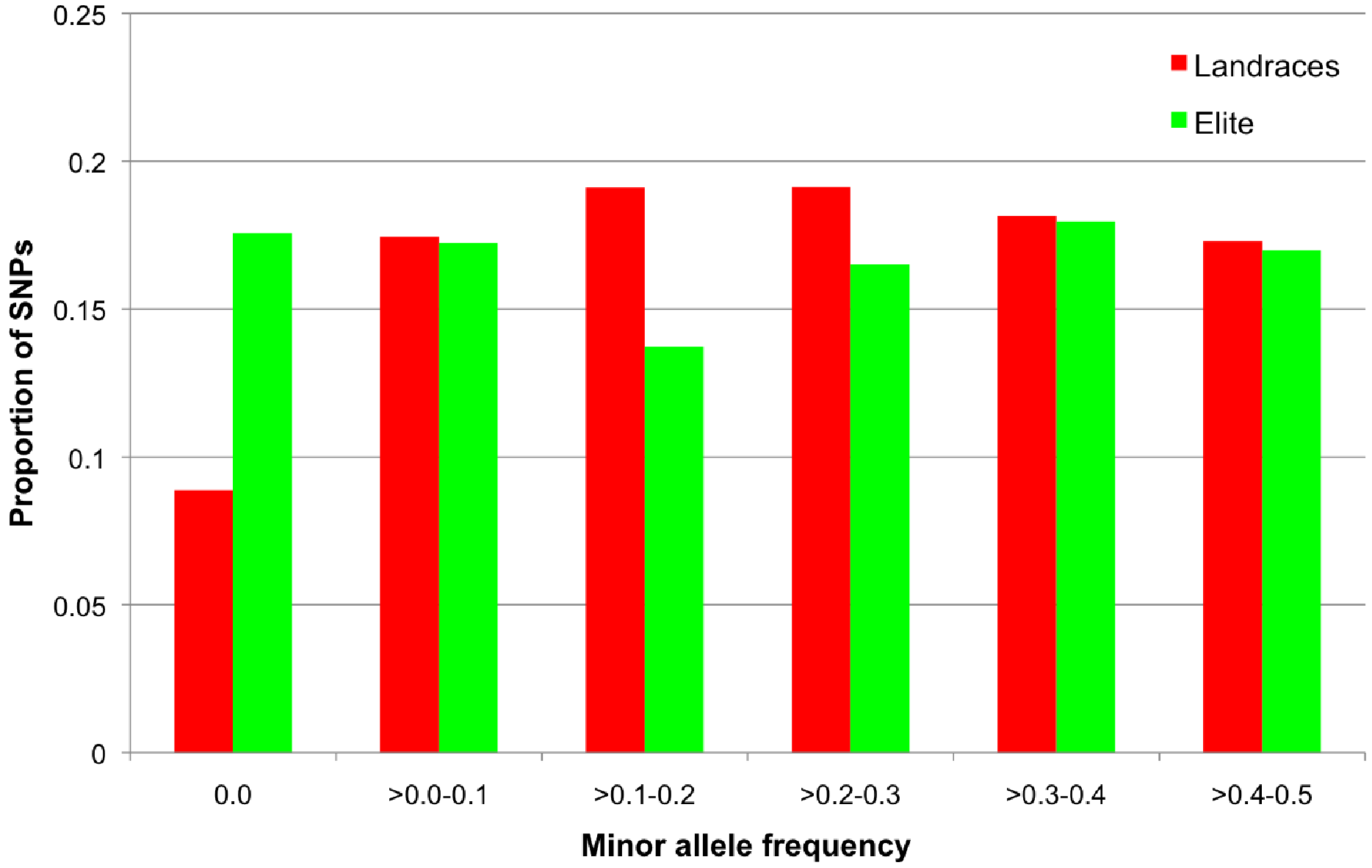
An overview of the soybean HapMap project at the USDA



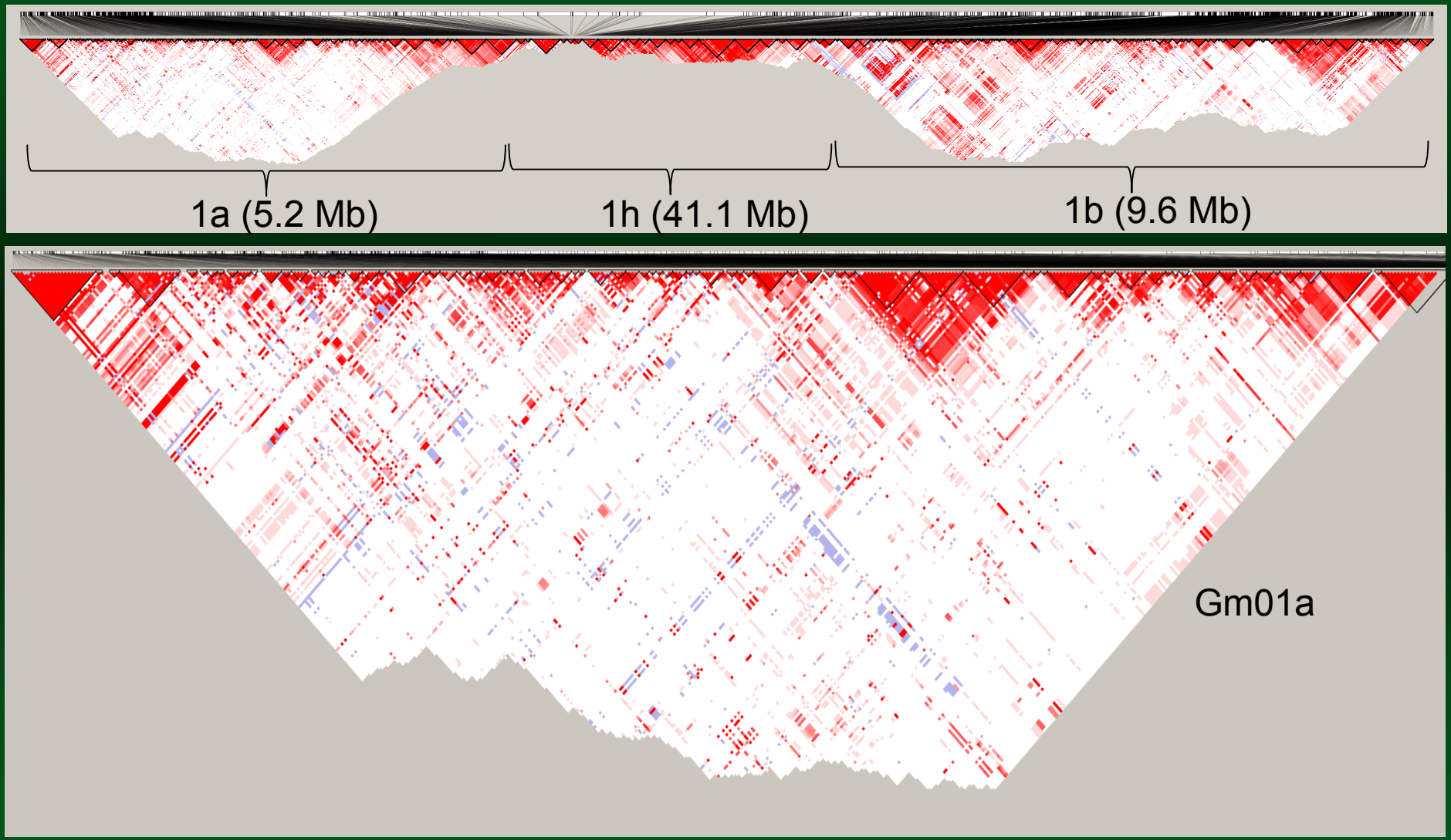
SoySNP50 chip run on diverse Exotic and Elite soybean germplasm lines to create a Phase I HapMap



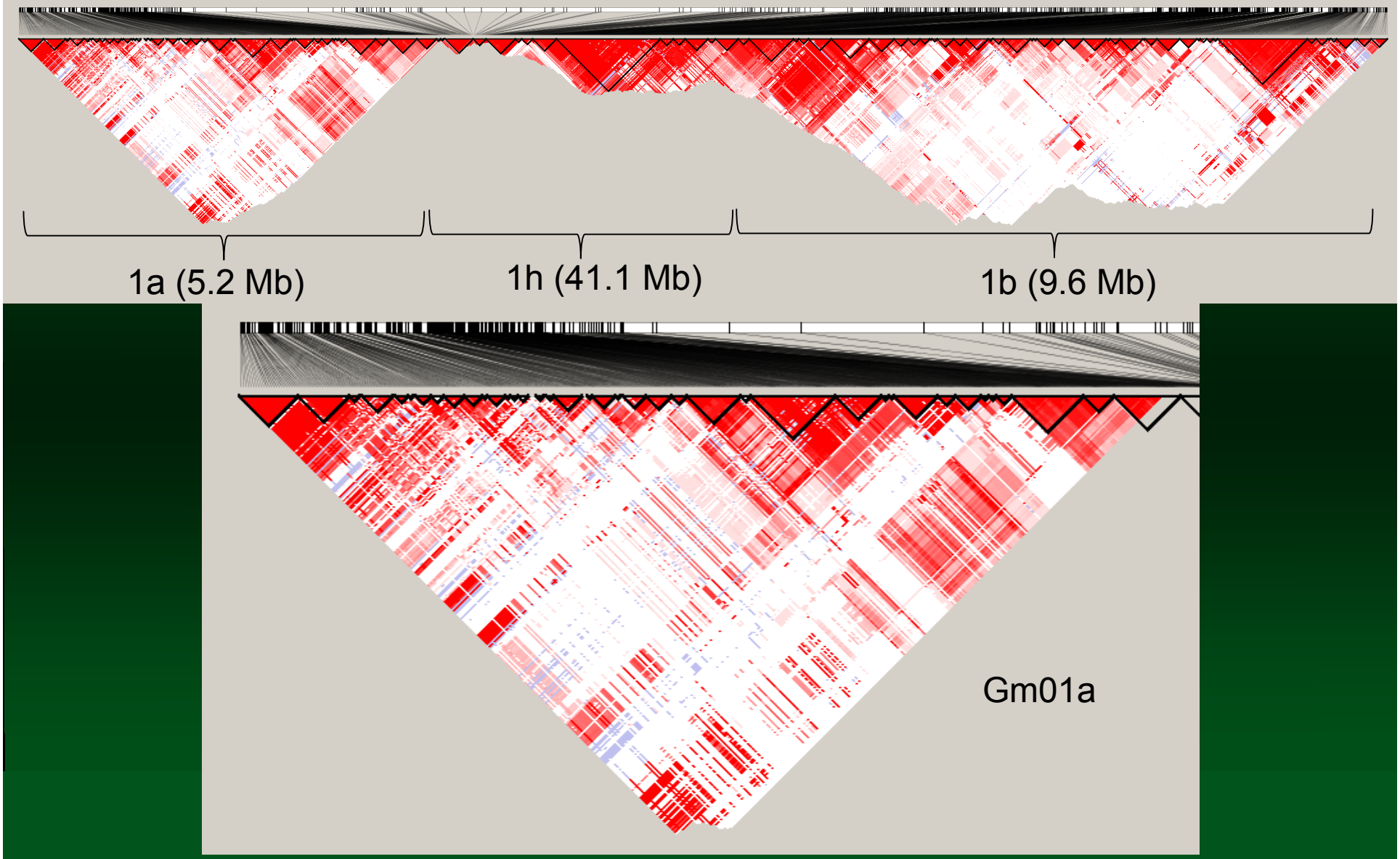
The 46,735 polymorphic SNPs on the chip had an even minor allele freq. distribution



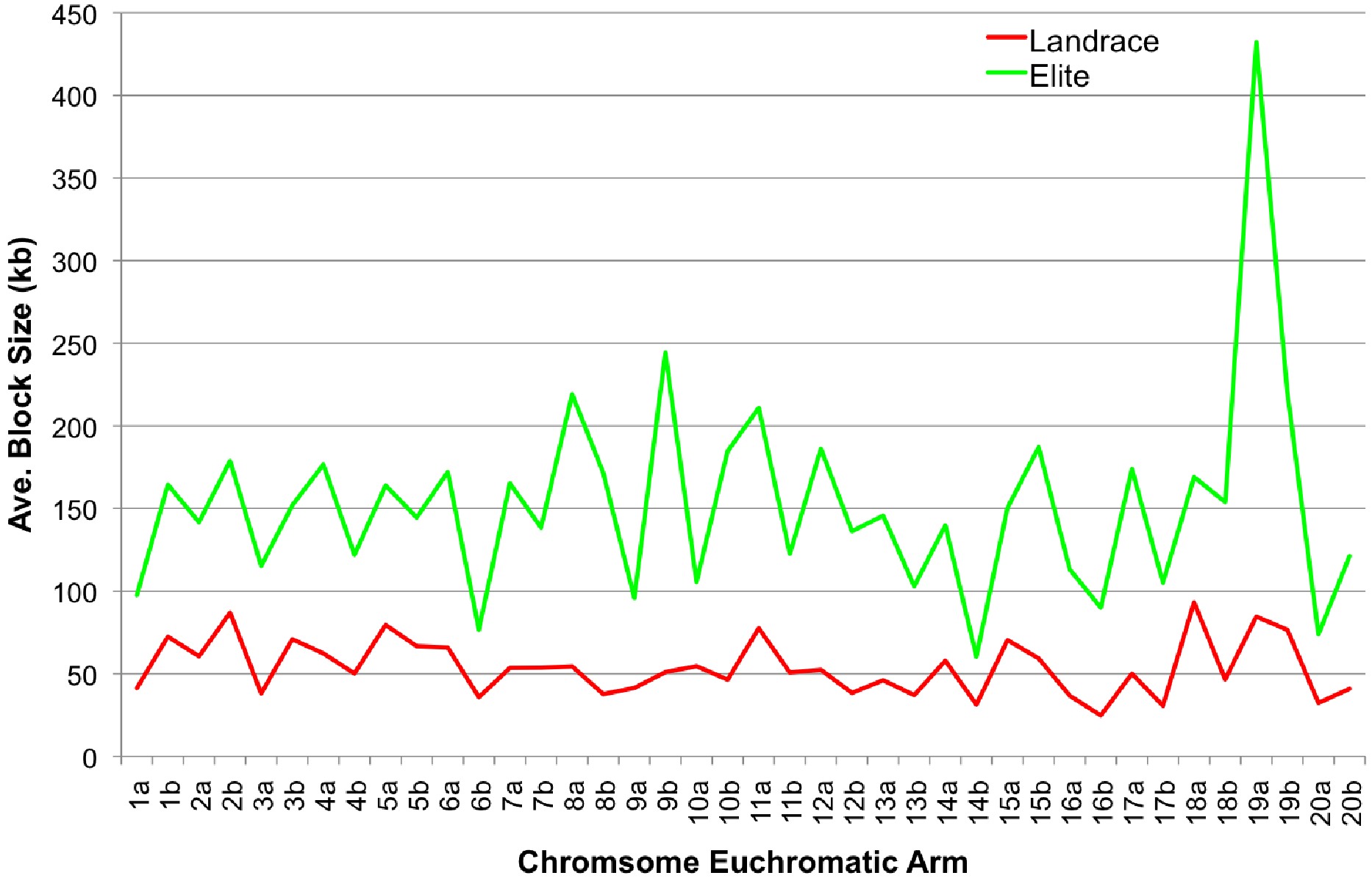
Chromosome 1 (1,757 SNPs in 96 Diverse Landraces)



Chromosome 1 (1,757 SNPs in 96 Diverse Elite Cultivars)



Ave. Block Size for the landraces and elite cultivars in Euchrom. DNA



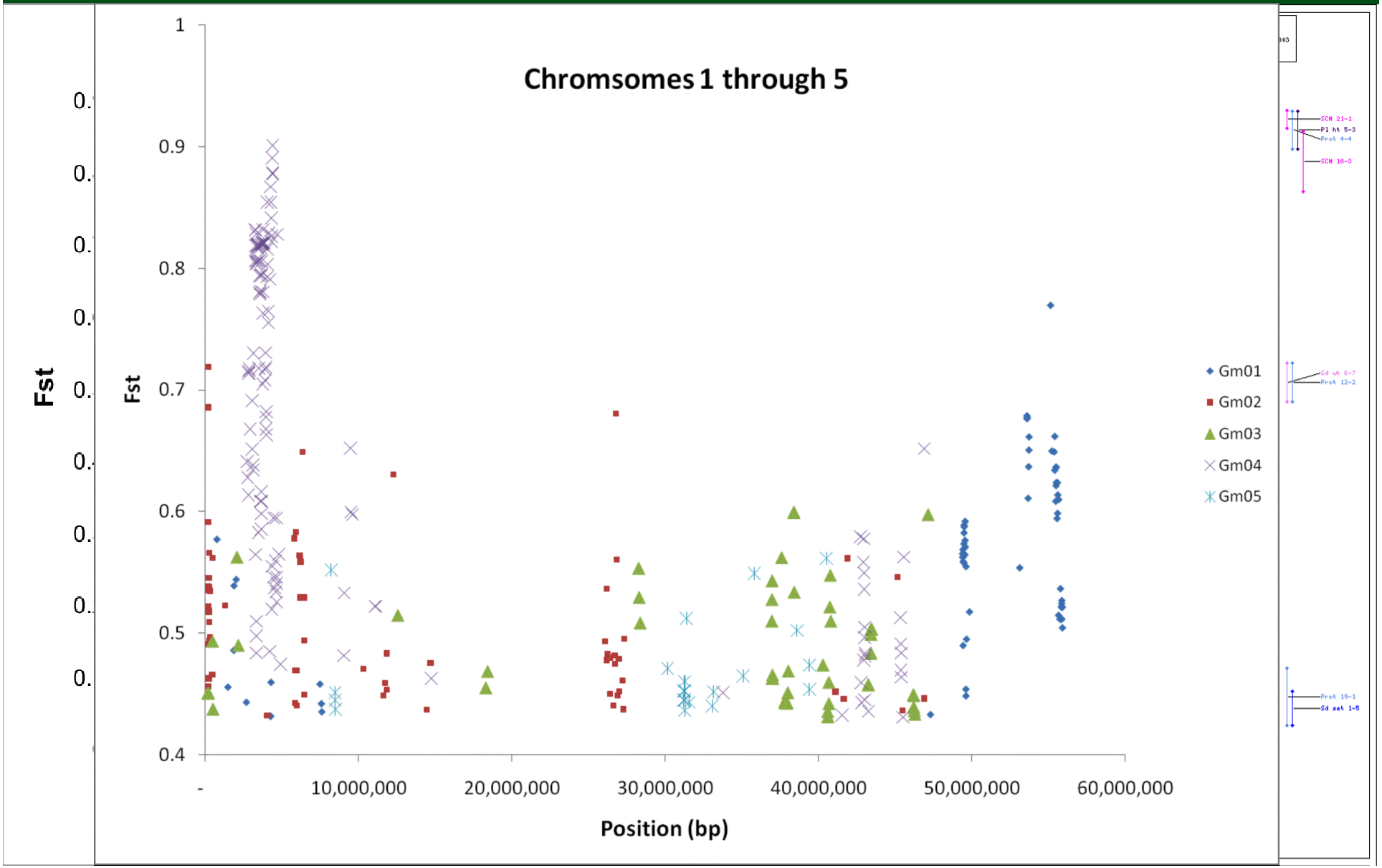
Landraces have 65% of the genome covered in LD blocks

Euchromatic DNA	• 232 Mb covered in LD blocks (50%) • Ave. LD block – 55 kb
Pericentromeric DNA	• 384 Mb covered in LD blocks (78%) • Ave. LD block – 1.2 Mb

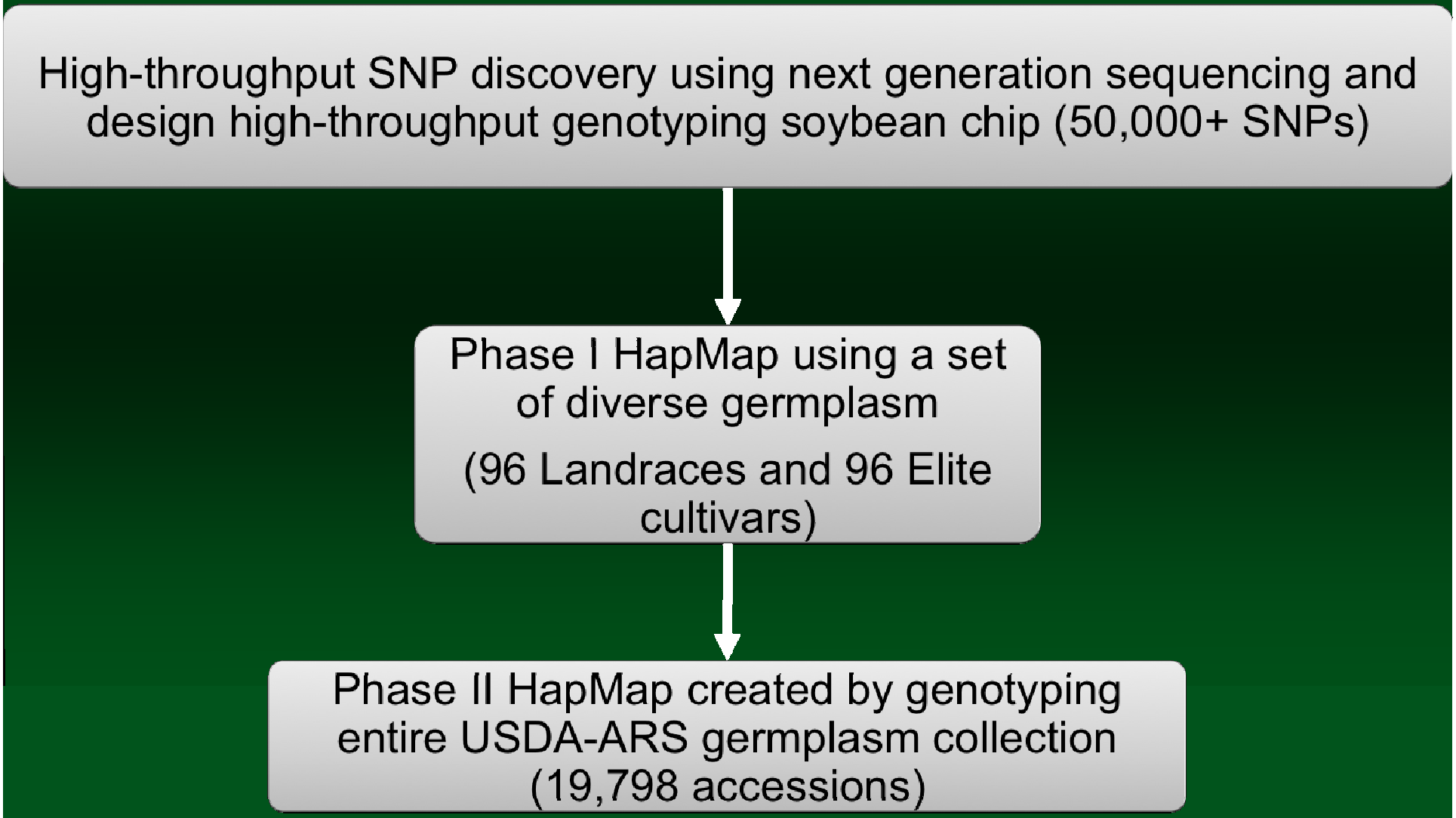
Elites have 75% of the genome covered in LD blocks

Euchromatic DNA	• 300 Mb covered in LD blocks (65%) • Ave. LD block – 149 kb
Pericentromeric DNA	• 414 Mb covered in LD blocks (84%) • Ave. LD block – 2.3 Mb

Chromosome 4 has a 2.1 Mb (~17cM) region which is one of the strongest signatures of selection in the genome



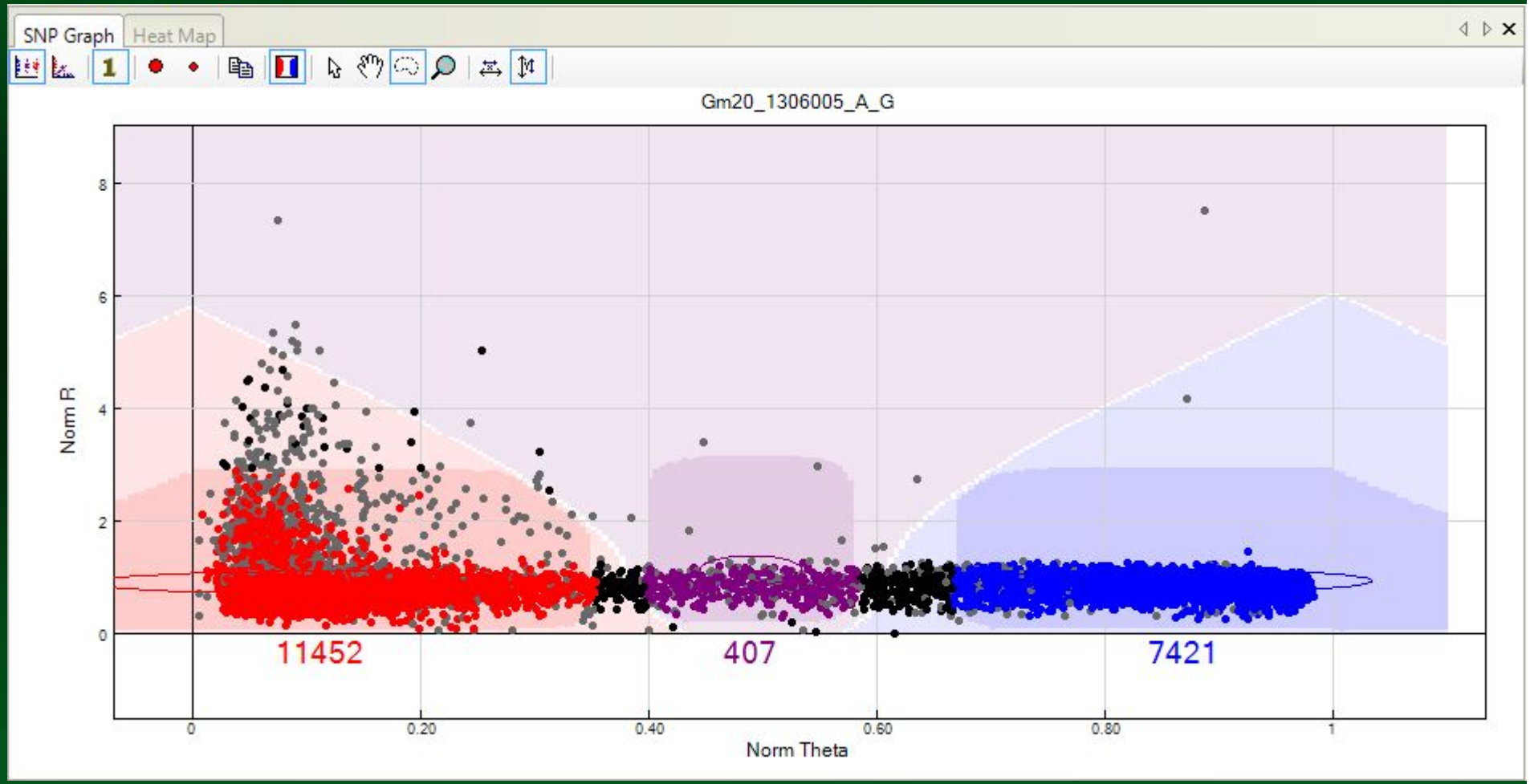
An overview of the soybean HapMap project at the USDA



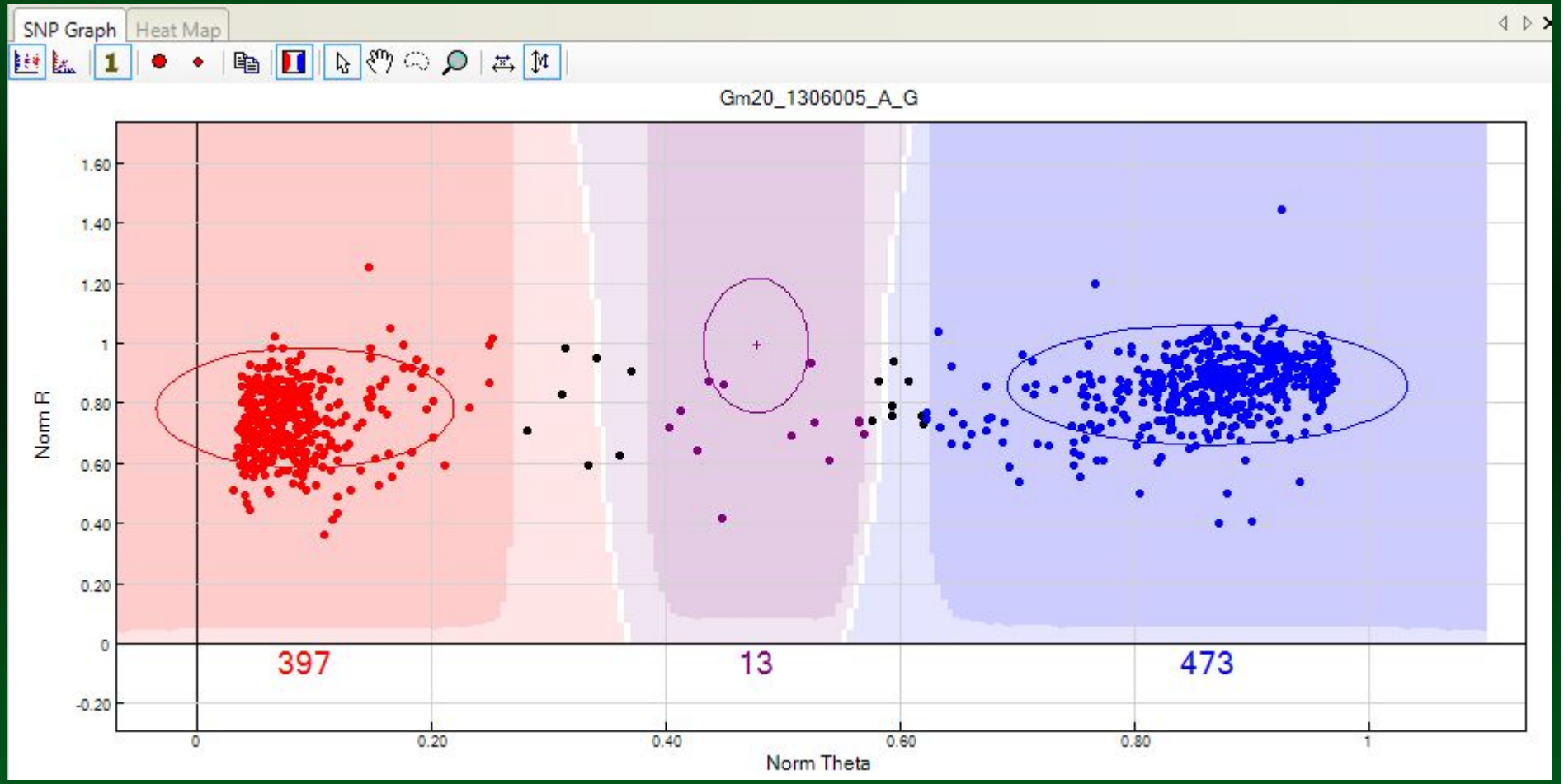
Soy HapMap Phase II

- Genotype the USDA soybean germplasm collection of 19,000+ wild and cultivated soybeans with 52,041 SNP DNA markers.
  - Have completed the initial lab portion of the genotyping
  - Of the 19,798 lines, 15552 have been examined for failures:
    - Total failures, 1344 or 8.6%
      - Failures from chip failure, 336 or 2.2%
  - Currently working to obtain allele calls for the accessions

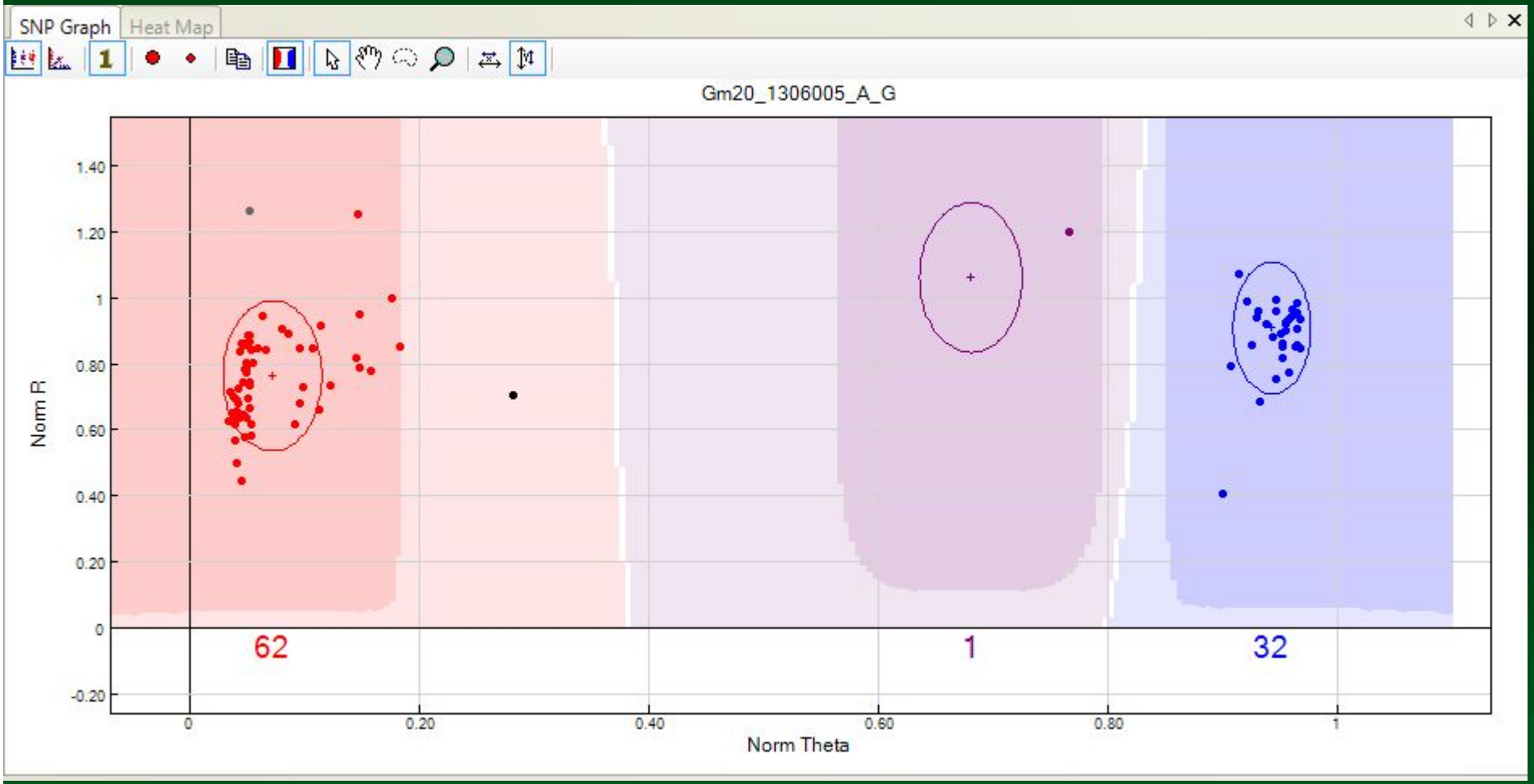
An example of a SNP with 19,768 accessions



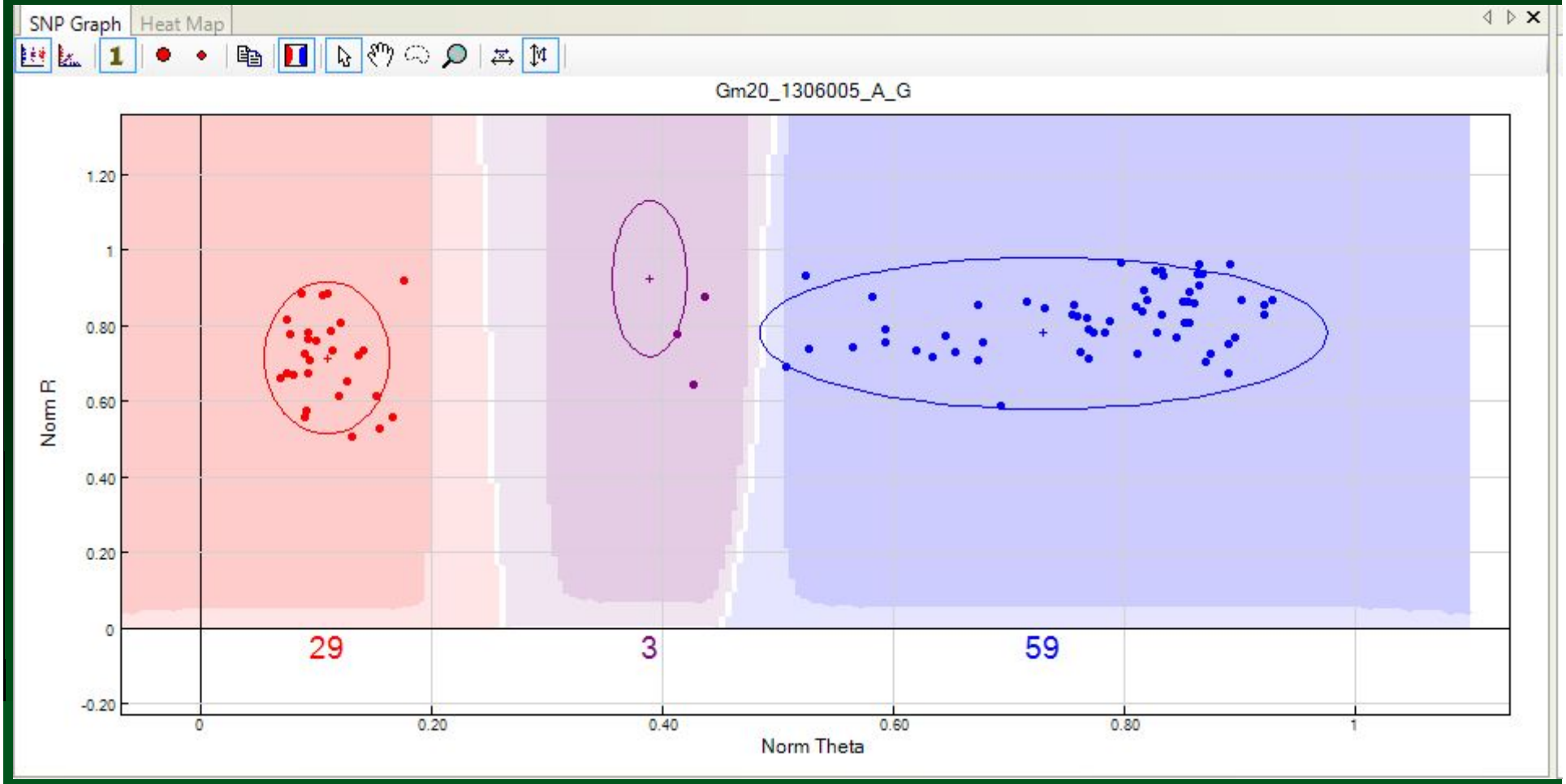
An example of 960 samples



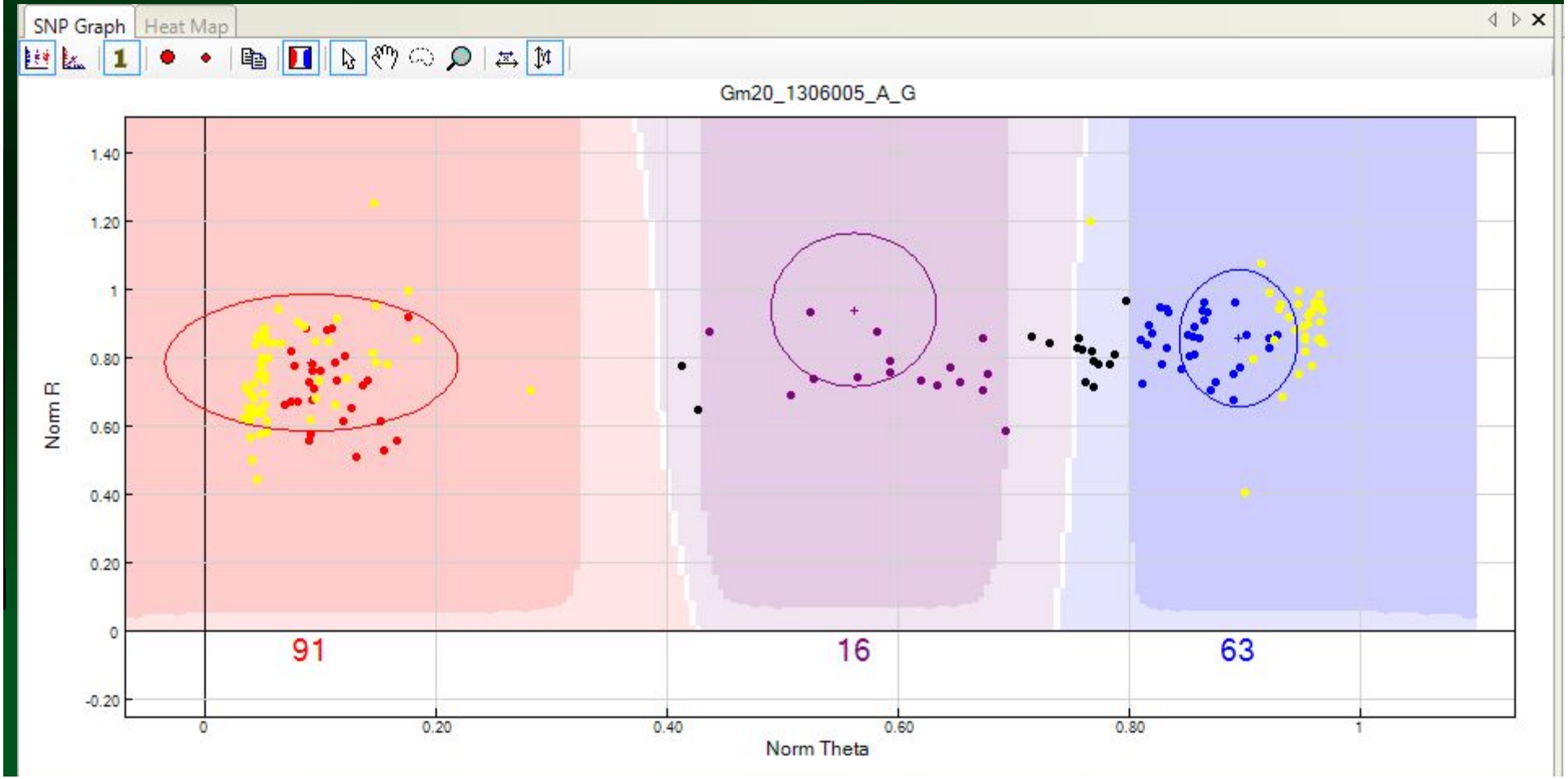
96 samples from same DNA plate



Same SNP, different set of 96 samples



The two plates (192 samples) put together (automatic allele calling)



Haplotyping 800+ lines associated with soybean rust resistance for 5 known *Rpp* genes

Accession number	Known Resistant gene	Haplotype	Genotype of Haplotype
PI 200492 (Rpp1 source)	Rpp1	Rpp1-Hap1	CTCGGATTGCCGCGCTCGATA
PI547875 (L85-2378)	Rpp1	Rpp1-Hap1	CTCGGATTGCCGCGCTCGATA
PI 594538A	Rpp1-b	Rpp1-Hap2	TGCATGGTGTCACGTCCAAGA
PI 587886	Rpp1	Rpp1-Hap3	CTAAGATCACTGNGCCCGCTA
PI 587880A	Rpp1	Rpp1-Hap3	CTAAGATCACTGNGCCCGCTA
PI 587905	Rpp1	Rpp1-Hap3?or6	CTAANATCACTGAGCTCGNTA
PI 561356	Rpp1	Rpp1-Hap4	TGCAGGTCATCGCACCCGAGC
PI594760B	Rpp1 or rpp1	Rpp1-Hap5	CTCGGGGTACCGCNCCCANNA
PI230970 (Rpp2 source)	Rpp2	Rpp2-Hap1	TGGACTTTCTACAGCATGTCGGTCCGCACTG CCCAGTTGCCGAAGTTCTT
PI230971	Rpp2	Rpp2-Hap2	TAGACGTCTTACAGCATGTCGGTCTTAGCTGT CCAGTTGCCGAAGTTCTT
PI417125	Rpp2	Rpp2-Hap2	TAGACGTCTTACAGCATGTCGGTCTTANCTGT CCAGTTGNCGAAGTTCTT
PI224270	rpp2	Rpp2-Hap2	TAGACGTCTTACAGCATGTCGGTCTTAGCTGT CCAGTTGNCGAAGTTCTT
[illegible]	[illegible]	[illegible]	GGAGTTTTCCGTGATGCTCTAACTCTAGTCAC [illegible]

Rpp3-5

Accession number	Known Resistant gene	Haplotype	Genotype of Haplotype
PI462312 (Original Rpp3 source)	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI506764 (Hyuuga)*	Rpp3&Rpp5	Rpp3-Hap1	TGTCNGAAACCNT
PI417503	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI605829*	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI605838	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI507259	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI506947	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI417089B	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI567024	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI567059	Rpp3	Rpp3-Hap1	TGTCGGAAACCGT
PI567046A	Rpp3	Rpp3-Hap1	TGTCGGAAACNGT
PI416873B	Rpp3	Rpp3-Hap2	TACCGAAGACCGT
PI615437	Rpp3	Rpp3-Hap3	TATTAAAAATTTT
PI605854B	Rpp3	Rpp3-Hap4	TGTCAAAAACCGT
PI605865B	Rpp3	Rpp3-Hap4	TGTCAAAAACCGT
PI605891A	Rpp3	Rpp3-Hap4	TGTCAAAAACNGT
PI606405*	Rpp3	Rpp3-Hap5	TGTTAAAAACCGT
PI459025B (Rpp4 source)	Rpp4	Rpp4-Hap1	GCAGTTCCGATAATCACCGTCAGC
PI567104B	Rpp4	Rpp4-Hap2	GCAGTTCCGATAGCTACCGTCAGC
PI417120	Rpp4	Rpp4-Hap3	GCAGCCCCGAGAATCACCGTCAGT
PI200456	Rpp5	Rpp5-Hap1	AATAGCAGTAGA
PI471904	Rpp5	Rpp5-Hap1	AATAGCAGTAGA
PI200487	Rpp5	Rpp5-Hap1	AATAGCAGTAGA
PI506764 (Hyuuga)	Rpp3&Rpp5	Rpp5-Hap1	AATAGCAGTAGA
PI200526	rpp5	Rpp5-Hap2	AATATTGGTAGA

Summary of 800+ lines with known haplotypes

Rpp locus	Number of lines
Rpp1	120
Rpp2	315
Rpp3	59
Rpp4	99
Rpp5	548
None	151

Number of genes pyramided	Number of lines
1	304
2	249
3	98
4	11
5	0

Conclusions

- Next generation sequencing has greatly accelerated SNP discovery
  - 177,347 SNPs discovered in Soybean
- The Soy 52k SNP Infinium assay is providing an order of magnitude greater genotyping capacity
  - 45,735 SNPs were successfully polymorphic
    - Landraces 74% SNPs with minor allele frequency >10%
    - Elites 65% SNPs with minor allele freqeuncy >10%
- Soybean HapMap Phase I will cover the majority of the genome in haplotype blocks
  - Landraces – 65% genome covered in haplotype blocks
  - Elites –75% genome covered in haplotype blocks
- Obtaining genotyping calls for the entire USDA Soybean Germplasm Collection (19,000^{+} genotypes) with the 52k chip is currently in progress

Thanks!!!!

- Perry Cregan, Ed Fickus, Chuck Quigley, Karen Williams, Gaofeng Jia, Eun-Young Hwang, Sophie Zebell, Christine Rajnes, USDA-ARS, Beltsville, MD
- Qijian Song, Univ. of Maryland
- James Specht, Univ. of Nebraska
- Randall Nelson, USDA-ARS, Urbana, IL
- Tommy Carter, Jr. USDA-ARS, Raleigh, NC
- Vince Pantalone, Univ. of Tennessee

Funding Support United Soybean Board, USDA-ARS