Large-Scale Genome Re-sequencing

Henry T. Nguyen

National Center for Soybean Biotechnology
University of Missouri

Funding support:
USDA-NIFA, the United Soybean Board,
and the Missouri Soybean Merchandising Council

Applying Next Generation Sequencing to Molecular Mapping

- **Sequencing-based genotyping approaches**
  - whole genome resequencing in rice (WGR)
    (Huang et al: 2009; Xie et al. 2010)
  - reduced representation of sequencing
    A. Maize: genotyping by sequencing (GBS). (Elshire et al. 2011)
    B. D. *simulans*: multiplexed shotgun genotyping (MSG) (Andolfatto et al. 2011)
    C. Stickleback: sequencing of restriction-site associated DNA tag (RAD) (Baird et al. 2008)
- **Advantages of sequencing-based genotyping approaches**
  - genotyping large amount of loci in many individuals
  - Identified SNPs can be validated by their segregation in progenies
  - more accurate
  - cost efficient
  - fast

Objectives of this Study

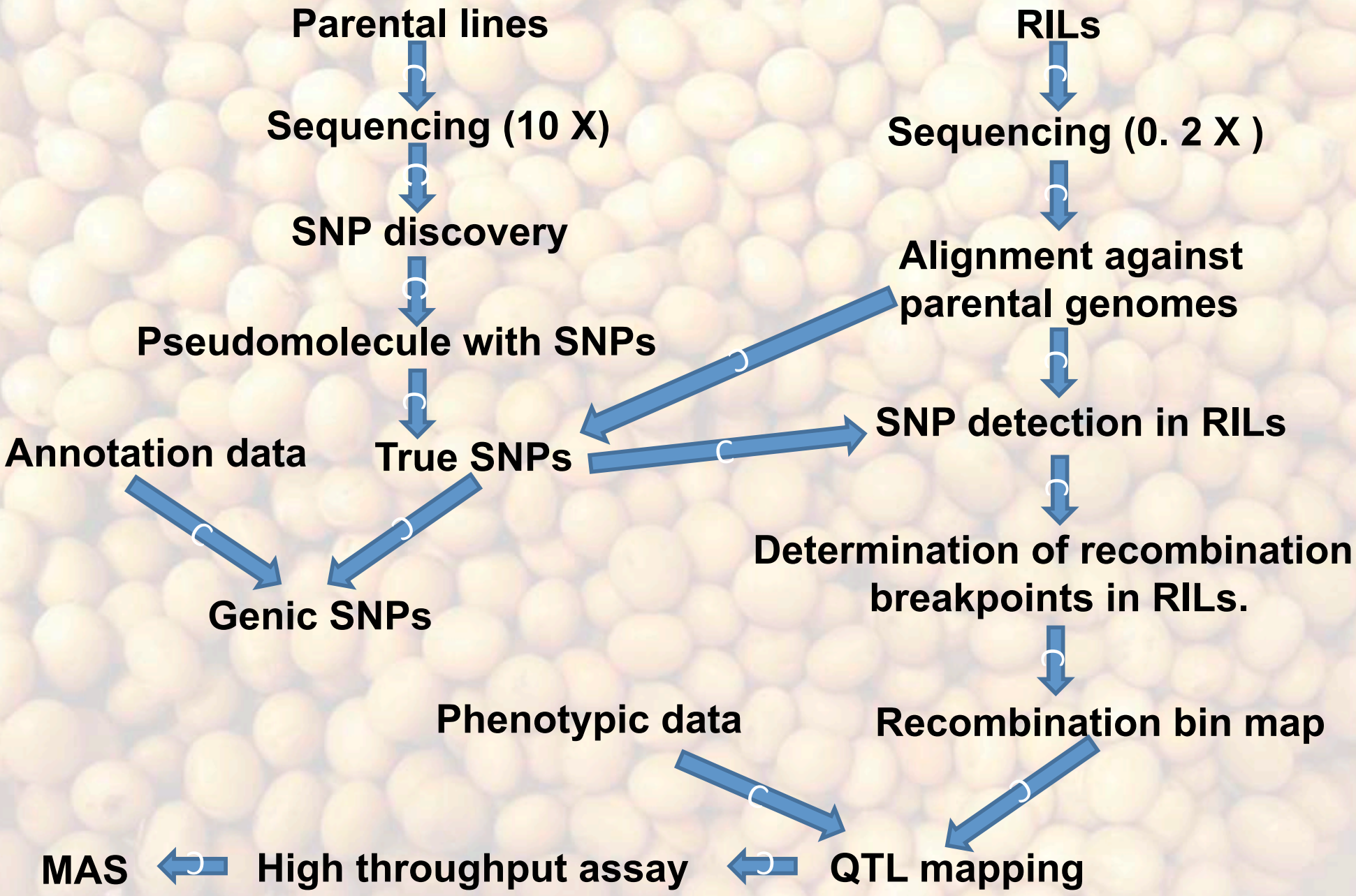
- Identify high quality SNPs/Indels and reveal landscapes of SNP distribution
- Develop a genotyping approach based on low coverage sequencing for soybean QTL mapping
- Construct a high density maps leading to identifying and cloning genes/QTLs for traits of agronomic importance

Materials

- Forrest x Williams 82 (FW82) population
  499 RILs
  Target traits: soybean cyst nematode (SCN)
  seed size

- Magellan x PI 438489B (MPB) population
  246 RILs
  Target traits: soybean cyst nematode (SCN)
  root knot nematode (RKN)
  Reniform nematode (RN)
  protein and oil content
  root morphology
  isoflavones
  seed size
- Ten additional RIL populations targeting different traits

Experimental Design



Forrest x Williams 82 Genetic Map

- Forrest and Williams 82 represent the southern and northern US soybean base, respectively.
- Williams 82 is the reference genome.
- Many genomic and genetic resources are available for both genotypes.
- 1,025 recombinant inbred lines (RILs) are available.
- A core set of 376 RILs were used to construct the framework map which includes a total of 986 markers; 471 SSRs and 515 SNP markers.
- About 5,000 SNP markers are being added to this genetic map using Illumina SNP array, in collaboration with Dow AgScience.

Wu et al. 2011.

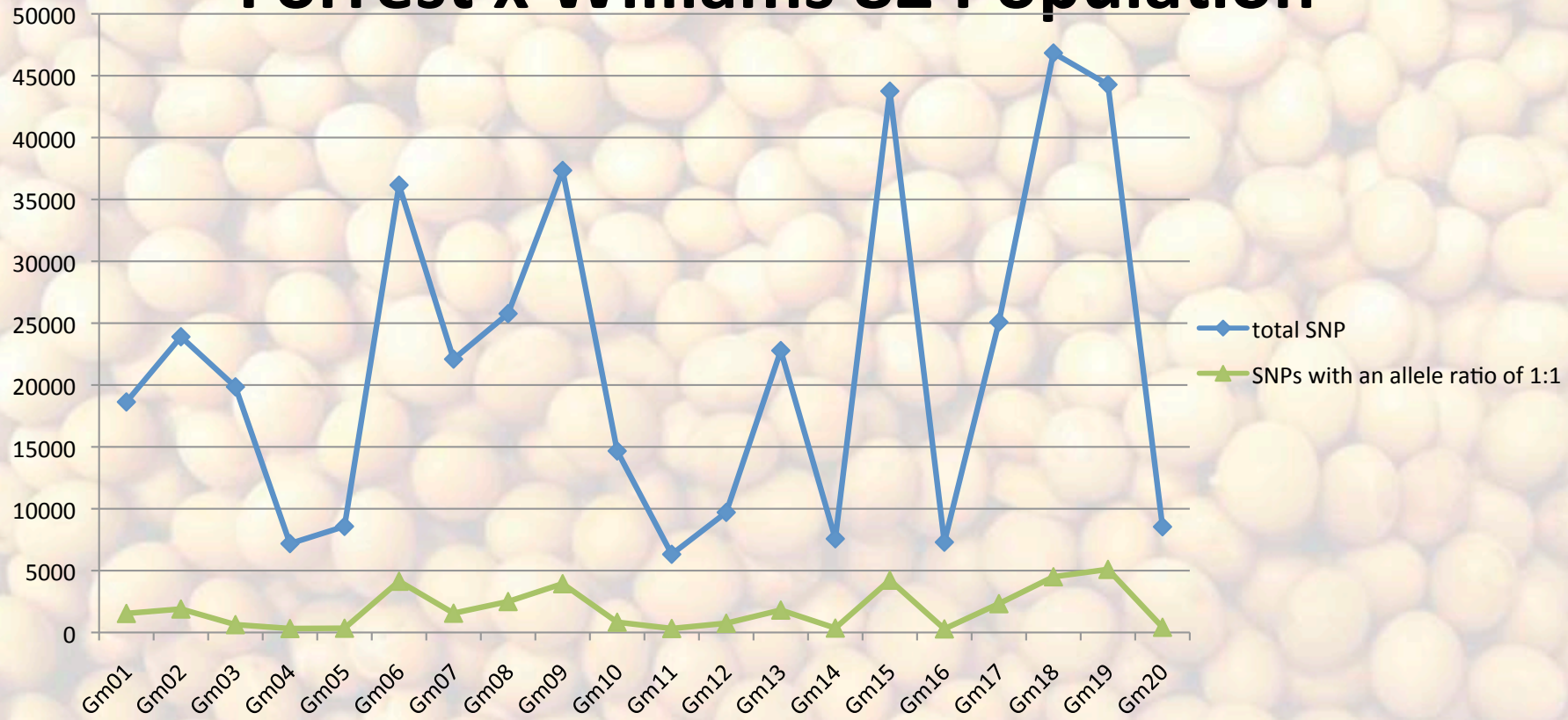
Re-sequencing of Parental Lines (Forrest, W82)

	Sequence (Gb)	Coverage	Depth	SNP			Genic SNP				
									CDS		
				Total	homo	hete	5'-UTR	Intron	syn	non-syn	3'-UTR
Williams 82_MO	9.4	94.4%	7.7X	86,281	13,172	73,109	0	1,643	575	307	0
Forrest	15.9	87.6%	12.4X	681,282	505,237	176,045	0	64,184	13,506	18,655	0

homo: homozygous; hete: heterozygous; syn: synonymous; non-syn: non-synonymous

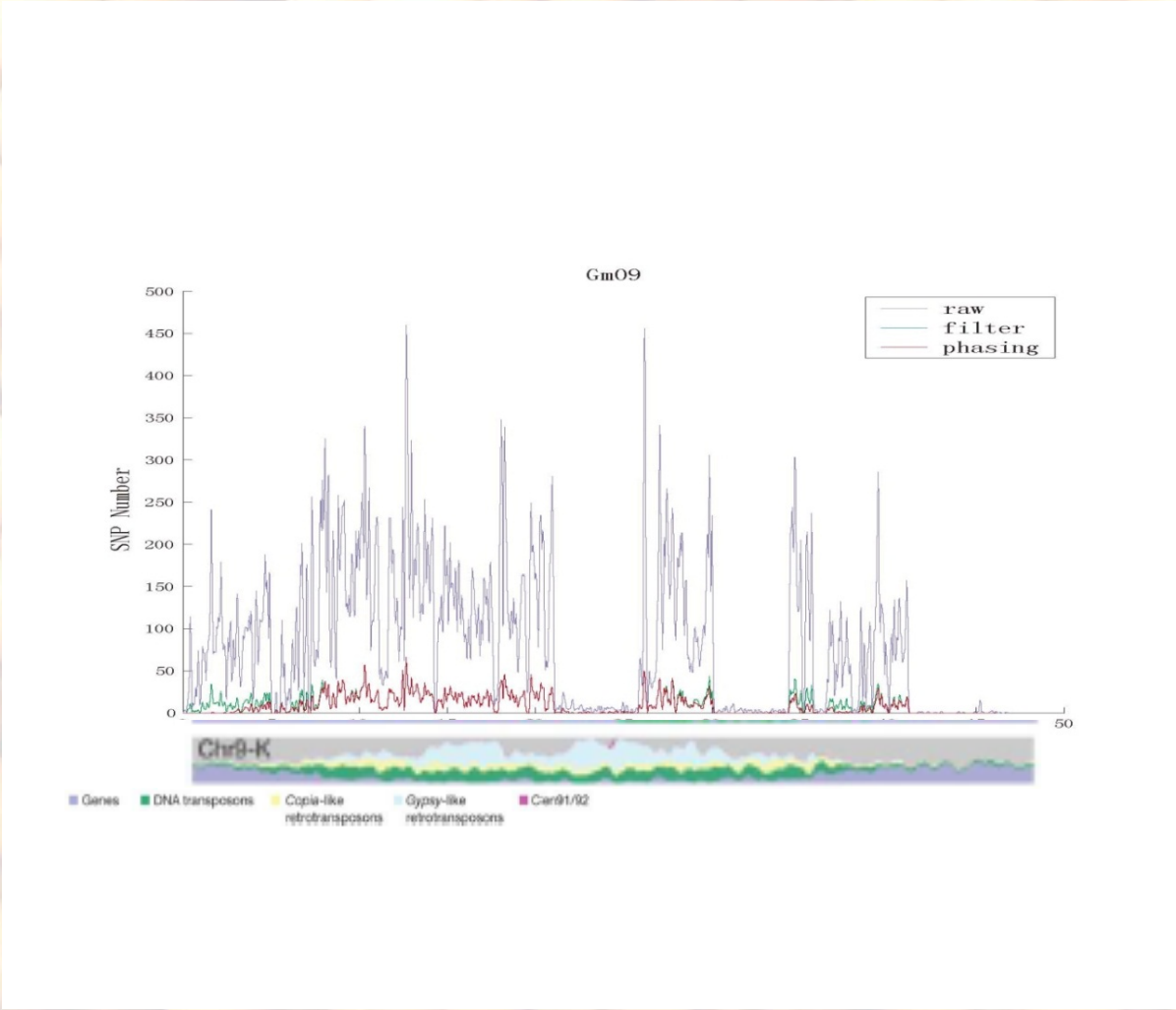
- 86,281 SNPs were identified between Williams 82_MO and the Williams 82 reference genome.
- Considerable intra-cultivar variances (nucleotide, structure, and gene content) were documented in Williams 82 (Haun et al. 2011).
- More than 600,000 SNPs identified between the two parents of the mapping population.
- 14% of total SNPs were genic and 4.7% of them resided in coding regions.

SNPs Distribution in Soybean Genome
Forrest x Williams 82 Population



- Uneven distribution of SNPs
- SNPs per chromosome range from 6,320 (Chr. 11) to 46,835 (Chr. 18)
- 37,820 SNPs segregating with an allele ratio of 1:1 were used to construct bin map

Distribution of SNP Markers on Chromosome 9



- Window size: 200 Kb
- Few SNPs found in centromere, telomere, and highly repetitive regions

Bin Map of 500 RILs from FW82 Population

Total bins: 3,356 Total SNPs: (37,820 SNPs) Median bin size: 157 Kb

Summary – Forrest x W82 Population

- A total of 681,282 SNPs were identified which were not evenly distributed in soybean genome.
- Using 37,820 SNPs that segregated with an 1:1 allele ratio in the FW82 RIL population, 3,356 bins were identified and used to construct the bin map.
- Average bin size is 157 kb.
- Recombination cold spot and hot spots were identified.
- This genotyping by sequencing approach enables the mapping of QTLs/genes to a smaller region (here 157 Kb) verses traditional mapping methods that typically produce regions of 10-20 cM (equivalent to 2-4 Mb).

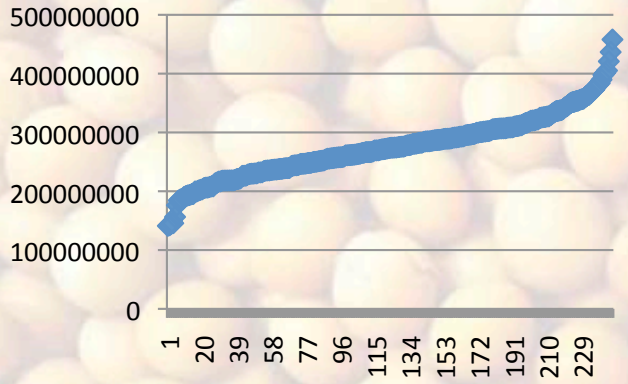
Re-sequencing of Parental Lines

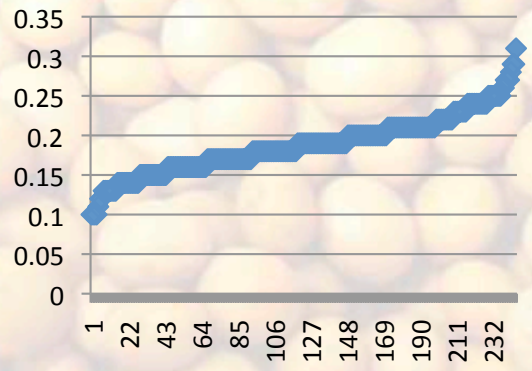
	Sequence (Gb)	Coverage	Depth	SNP			Genic SNP				
									CDS		
				Total	homo	hete	5'-UTR	Intron	syn	non-syn	3'-UTR
Williams 82_MO	9.4	94.4%	7.7X	86,281	13,172	73,109	0	1,643	575	307	0
Forrest	15.9	87.6%	12.4X	681,282	505,237	176,045	0	64,184	13,506	18,655	0
Magellan	16.4	95.2%	13.5X	661,386	463,662	197,724	2,108	51,047	10,058	13,392	4,220
PI438489B	17.1	93.1%	13.5X	1,237,794	1,006,361	233,433	4,781	118,867	22849	30,139	9,814

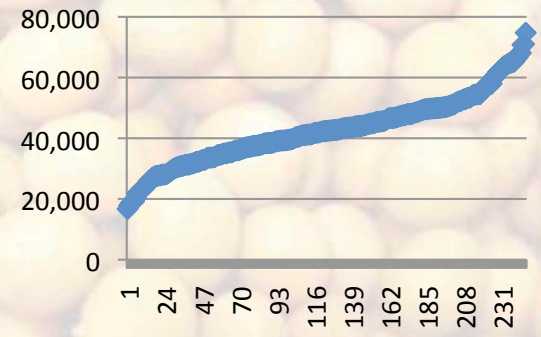
homo: homozygous; hete: heterozygous; syn: synonymous; non-syn: non-synonymous

- Considerable intra-cultivar variances were identified in Williams 82.
- Significantly more SNPs (2-fold) were found in PI438489B
- 12.0-15.1% of total SNPs were genic and 2-2.4% of them are nonsynonymous (which may change the functions of genes).

Re-sequencing of the MPB Recombinant Inbred Line (RIL) Population





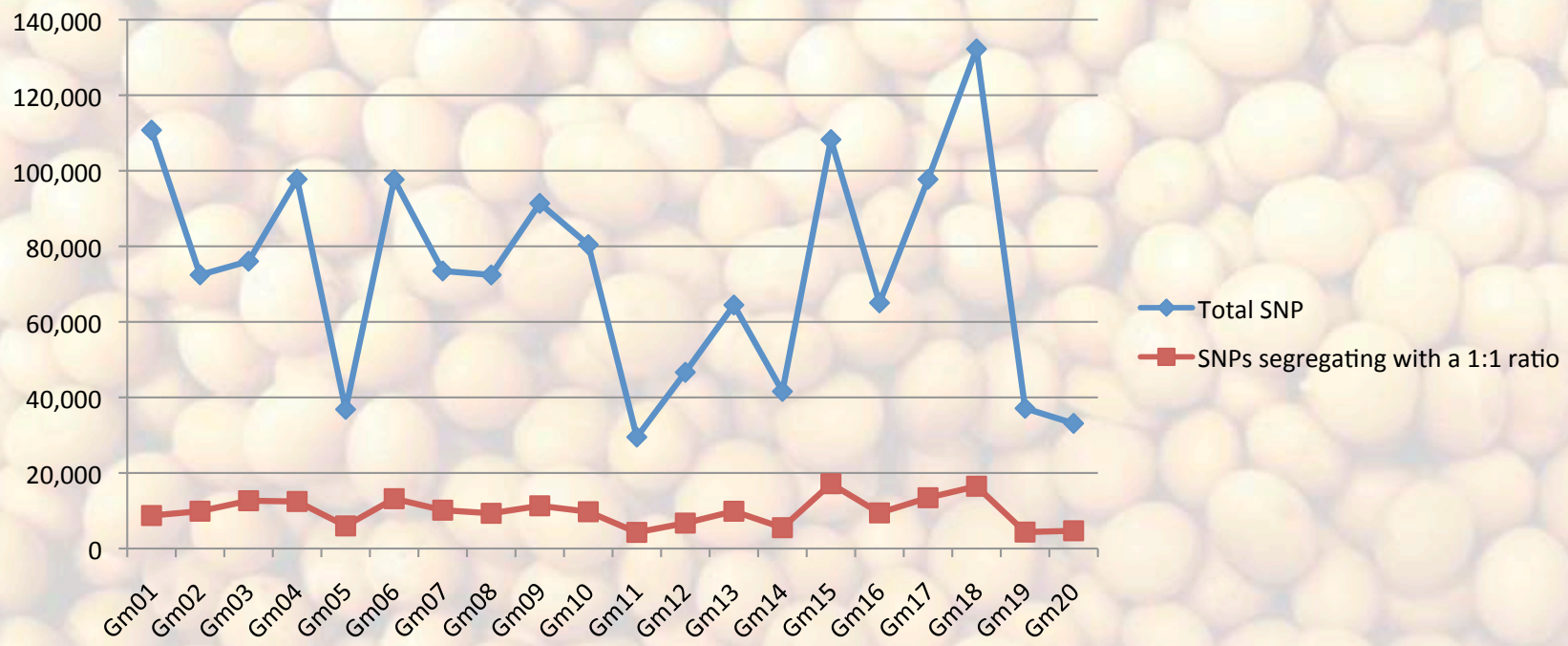


67.6 Gb sequence
Range: 141 – 458Mb/RIL
Mean: 275Mb/RIL

Depth: 0.1-0.31X
(Coverage: 8.8-24.8%)
Mean: 0.19X (16.2%/RIL)

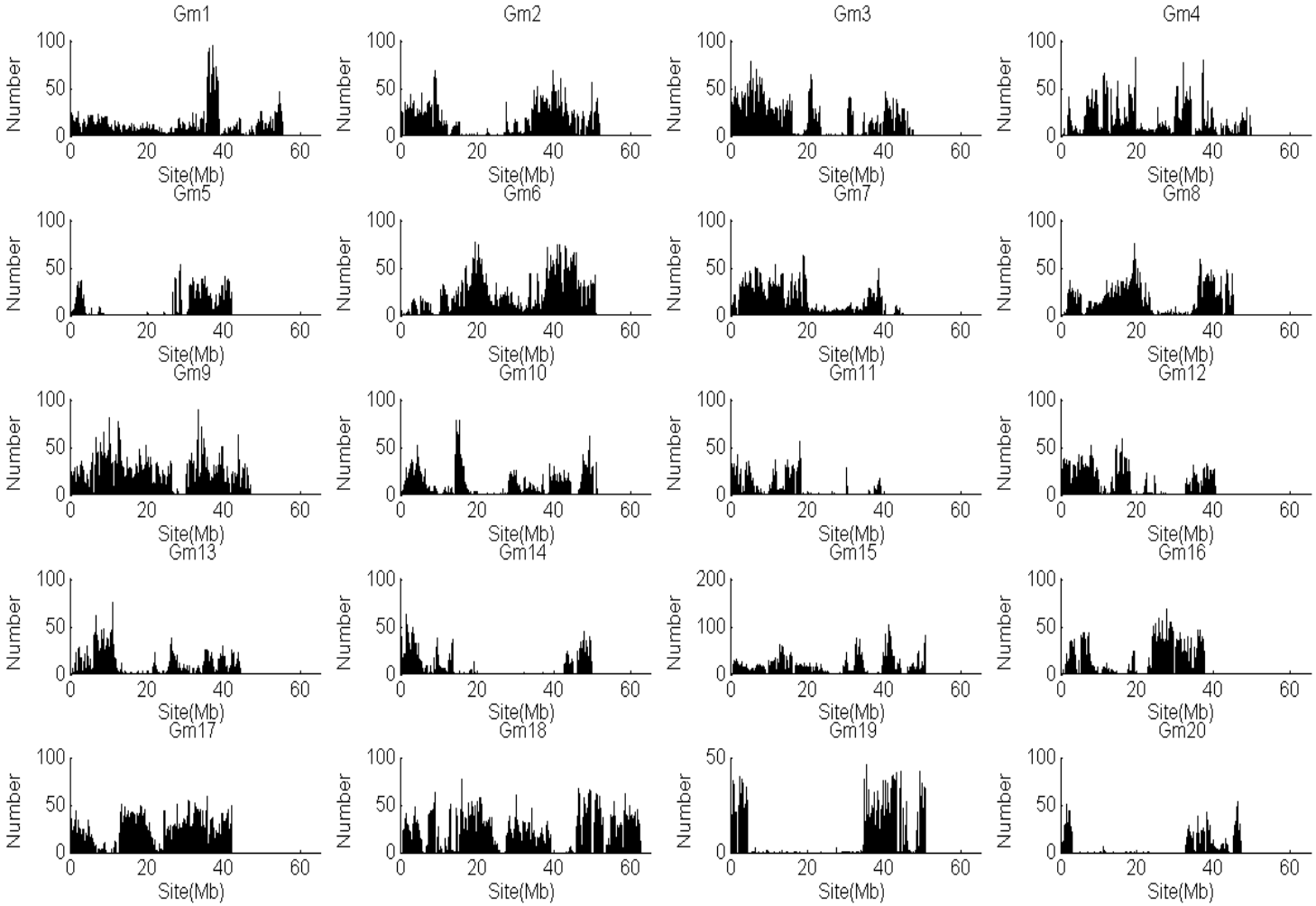
Homozygous SNPs
Range:16,705-74,756
Mean: 42577

SNPs Distribution in Soybean genome (MPB Population)



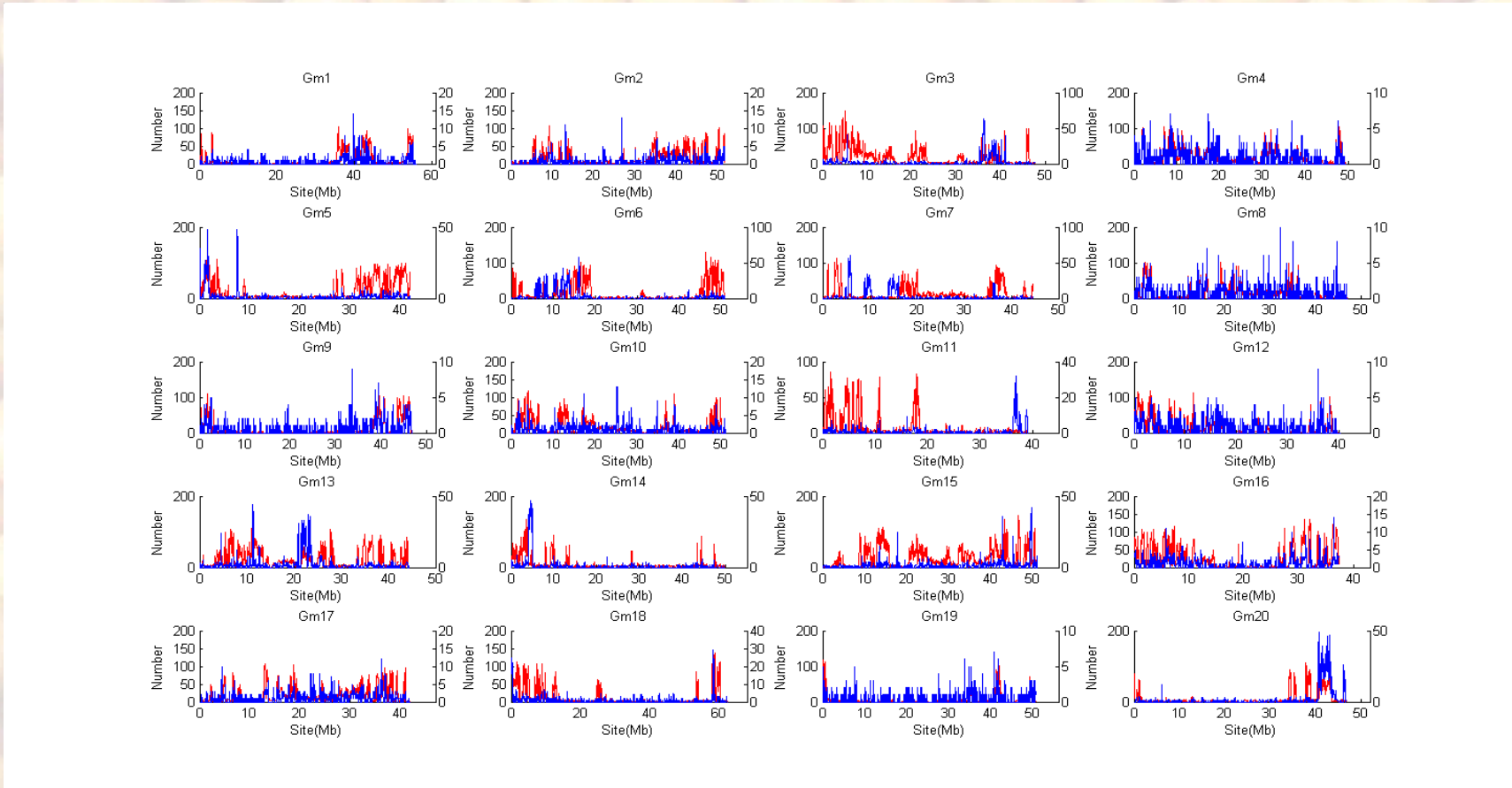
- Total SNPs between two parents (Magellan x and PI 438489B): 1,464,938
- Uneven distribution of SNPs.
- SNPs per chromosome range from 29,500(Chr. 11) to 110,760 (Chr. 18).
- 195,375 SNPs segregating with an allele ratio of 1:1 were used to construct the bin map.

Distribution of SNP Markers (MPB pop)



Window size: 100Kb on the physical map of W82 reference genome

Distribution of Indel Markers (MPB Pop)



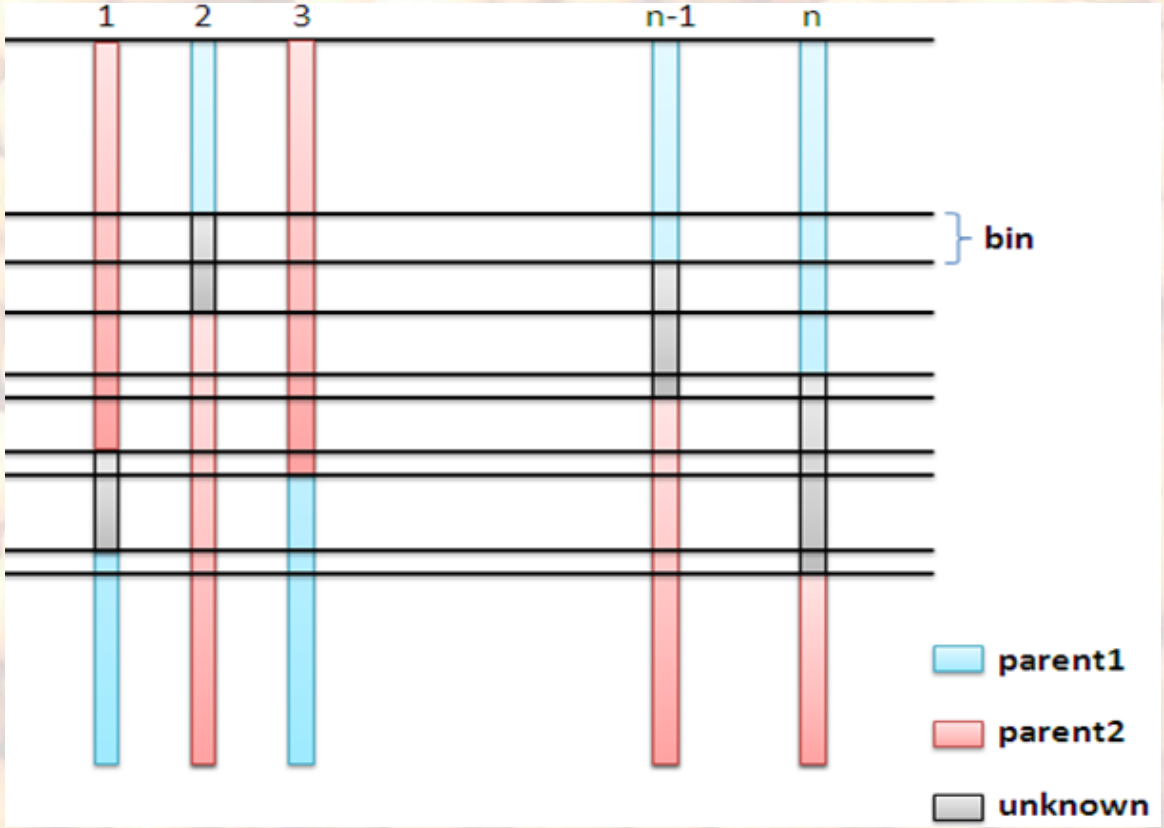
Red: homozygous ; blues: heterozygous

Indels:

PI438489B: 317,118

Magellan: 156,385

Construction of Bin Map

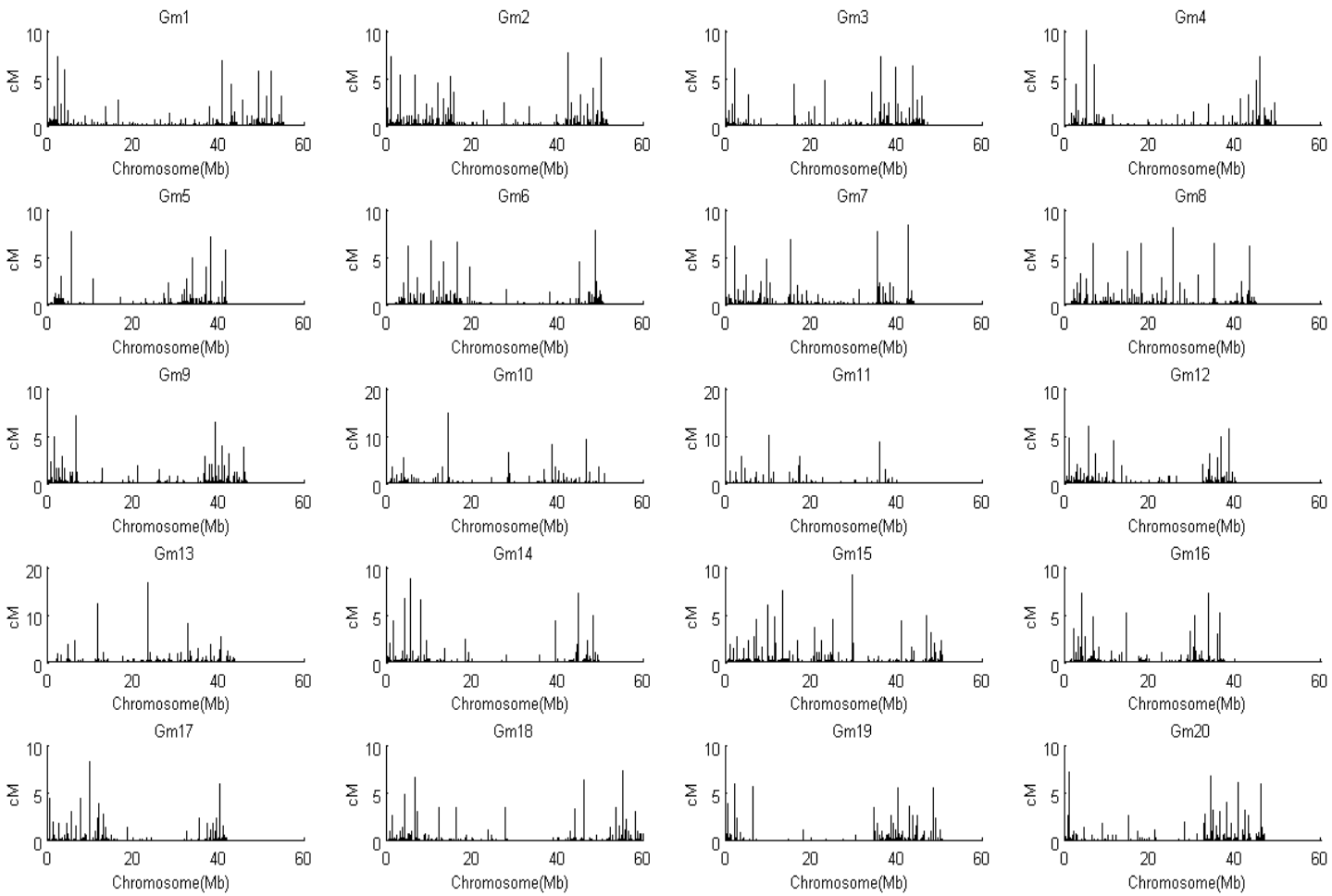


- Bin is the chromosome interval between 2 adjacent recombination breakpoints.

Bin Map of 246 RILs from MPB Population

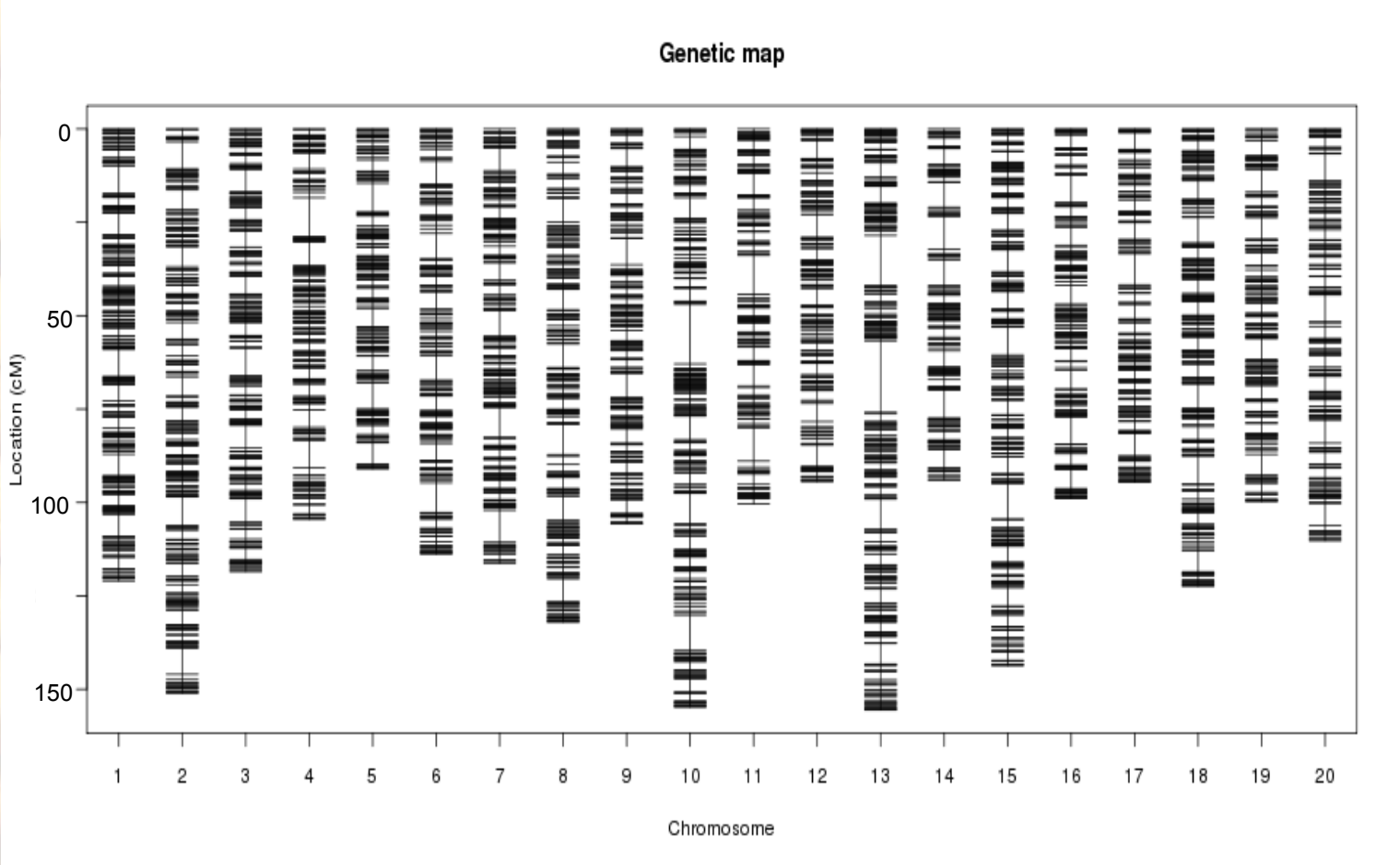
Total bins: 3510 ; Total SNPs: 195,375 **Median bin size: 98.2Kb**

Recombination Hot/Cold Spots



Window size: 200KB

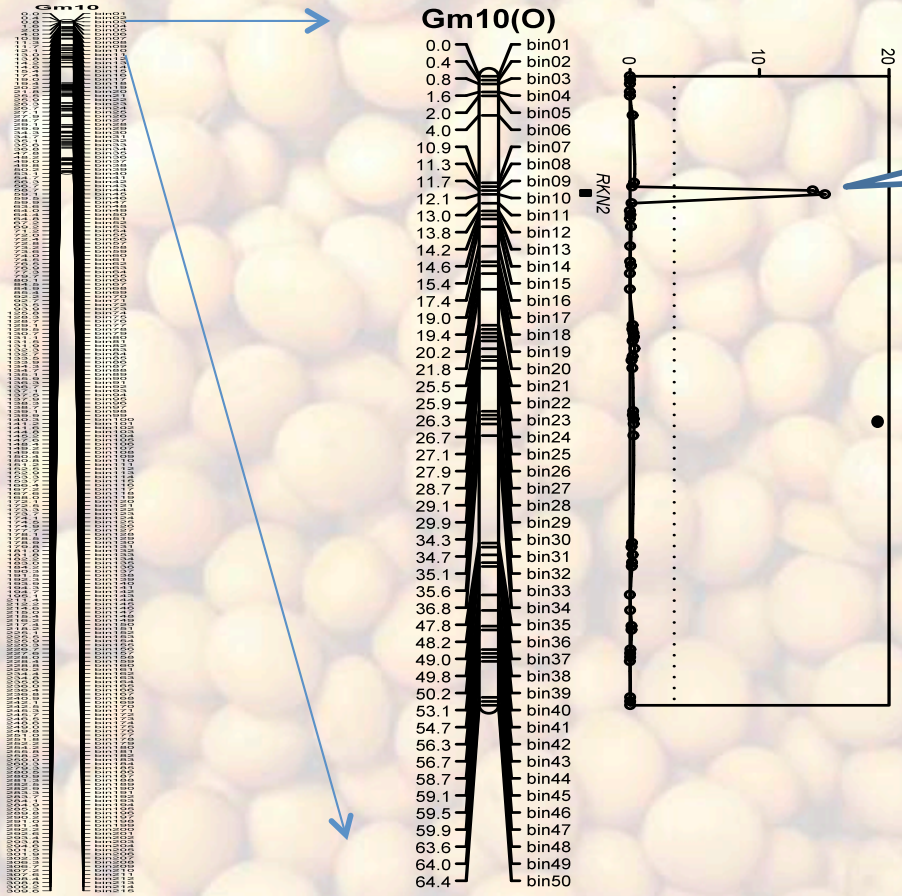
Genetic Maps with Bins Serving as Markers



Total genetic distance: 2314 cM

Positions of bins are in perfect agreement with their physical locations

Mapping QTL for Root Knot Nematode Resistance



LOD: 15.1
R^2 : 23.2 %
Additive effect : 0.75

- **Coincide with a QTL previously mapped to an interval of 9.1 cM (between Satt358 and Satt492) (Li et al. 2001)**

Chr.	Bin	Bin size	LOD	R^2	Add. Effect
8	113	340 Kb (~ 1.7 cM)	5.3	7.4	0.42
10	10	29.7 Kb (~ 0.15 cM)	15.1	23.2	0.75
13	131	7.9 Kb (~ 0.04 cM)	4.1	5.6	0.4

Pinpoint the Gene Underlying the Major QTL for Root Knot Nematode

- Four genes in Chr10_bin10
  - three genes were annotated as “cell-wall-associated enzyme”
  - one gene was annotated as “protein of unknown function”
- Identification of the candidate gene
  - 97 SNPs/Indels were identified in Chr10_bin10
  - 15 genic SNPs/Indels were identified in the CDS regions of two genes
  - Candidate SNPs (nonsynonymous SNP) are identified
- Next:
  - Cloning the genes
  - Functional analysis

Summary

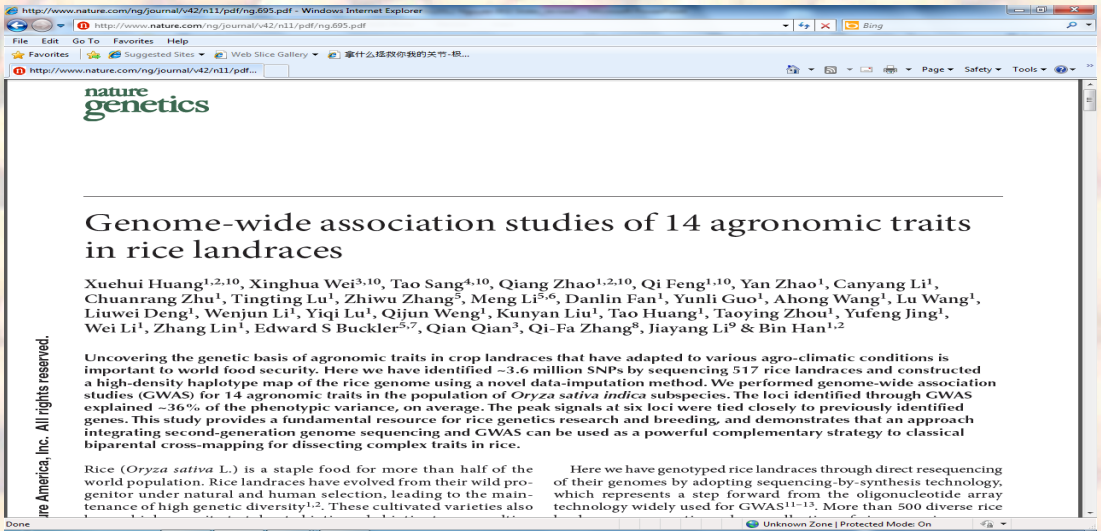
- The genotyping approach developed in this study is reliable and highly efficient. It is accurate, fast and cost-effective.
- Recommendations for sequencing-based genotyping in soybean:

  - cross pattern: cultivar x PI
  - sequencing of parental lines: 2X coverage
  - sequencing of RILs: ~0.2X coverage
  - multiplex: 48 RILs/library

Genome-wide association study (GWAS)

- HapMaps constructed with microarray-based methods suffered substantial ascertainment bias
- Parallel and comprehensive SNP discovery and genotyping by sequencing
  - Huang et al. sequenced a set of rice germplasm at low coverage (~1X).
  - QTL for important traits were mapped

(Huang et al. 2010; 2011)



nature genetics

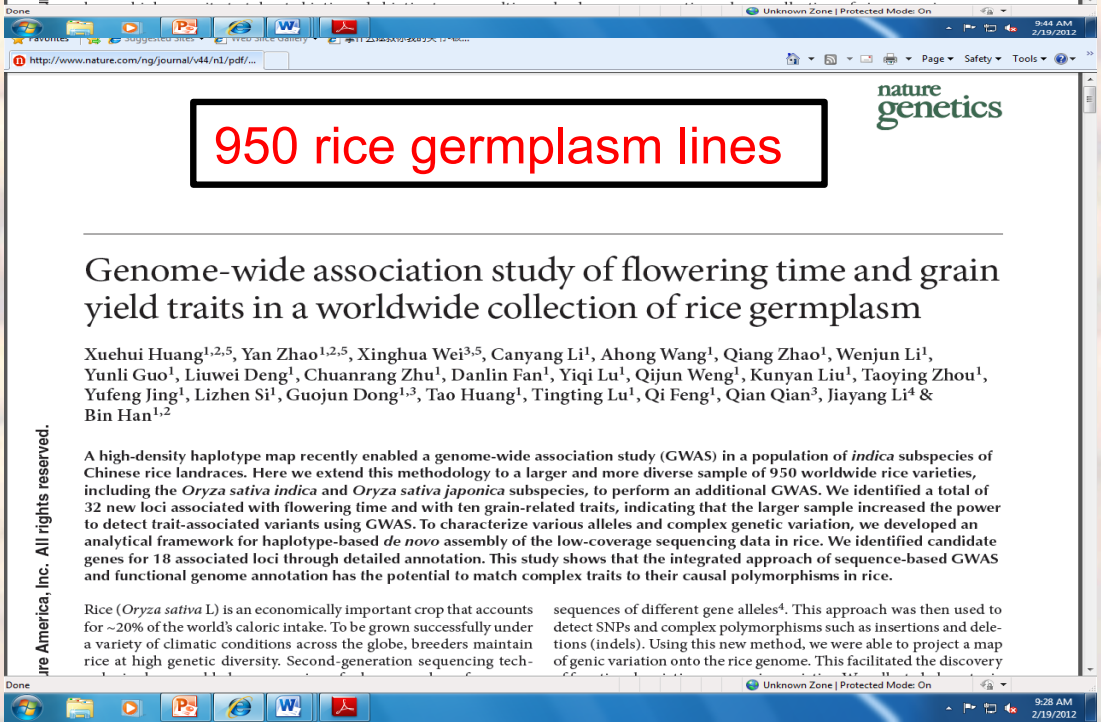
Genome-wide association studies of 14 agronomic traits in rice landraces

Xuehui Huang[1,2,10], Xinghua Wei[3,10], Tao Sang[4,10], Qiang Zhao[1,2,10], Qi Feng[1,10], Yan Zhao[1], Canyang Li[1], Chuanrang Zhu[1], Tingting Lu[1], Zhiwu Zhang[5], Meng Li[5,6], Danlin Fan[1], Yunli Guo[1], Ahong Wang[1], Lu Wang[1], Liuwei Deng[1], Wenjun Li[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Tao Huang[1], Taoying Zhou[1], Yufeng Jing[1], Wei Li[1], Zhang Lin[1], Edward S Buckler[5,7], Qian Qian[3], Qi-Fa Zhang[8], Jiayang Li[9] & Bin Han[1,2]

Uncovering the genetic basis of agronomic traits in crop landraces that have adapted to various agro-climatic conditions is important to world food security. Here we have identified ~3.6 million SNPs by sequencing 517 rice landraces and constructed a high-density haplotype map of the rice genome using a novel data-imputation method. We performed genome-wide association studies (GWAS) for 14 agronomic traits in the population of *Oryza sativa indica* subspecies. The loci identified through GWAS explained ~36% of the phenotypic variance, on average. The peak signals at six loci were tied closely to previously identified genes. This study provides a fundamental resource for rice genetics research and breeding, and demonstrates that an approach integrating second-generation genome sequencing and GWAS can be used as a powerful complementary strategy to classical biparental cross-mapping for dissecting complex traits in rice.

Rice (*Oryza sativa* L.) is a staple food for more than half of the world population. Rice landraces have evolved from their wild progenitor under natural and human selection, leading to the maintenance of high genetic diversity[1,2]. These cultivated varieties also

Here we have genotyped rice landraces through direct resequencing of their genomes by adopting sequencing-by-synthesis technology, which represents a step forward from the oligonucleotide array technology widely used for GWAS[11–13]. More than 500 diverse rice



950 rice germplasm lines

nature genetics

Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm

Xuehui Huang[1,2,5], Yan Zhao[1,2,5], Xinghua Wei[3,5], Canyang Li[1], Ahong Wang[1], Qiang Zhao[1], Wenjun Li[1], Yunli Guo[1], Liuwei Deng[1], Chuanrang Zhu[1], Danlin Fan[1], Yiqi Lu[1], Qijun Weng[1], Kunyan Liu[1], Taoying Zhou[1], Yufeng Jing[1], Lizhen Si[1], Guojun Dong[1,3], Tao Huang[1], Tingting Lu[1], Qi Feng[1], Qian Qian[3], Jiayang Li[4] & Bin Han[1,2]

A high-density haplotype map recently enabled a genome-wide association study (GWAS) in a population of *indica* subspecies of Chinese rice landraces. Here we extend this methodology to a larger and more diverse sample of 950 worldwide rice varieties, including the *Oryza sativa indica* and *Oryza sativa japonica* subspecies, to perform an additional GWAS. We identified a total of 32 new loci associated with flowering time and with ten grain-related traits, indicating that the larger sample increased the power to detect trait-associated variants using GWAS. To characterize various alleles and complex genetic variation, we developed an analytical framework for haplotype-based *de novo* assembly of the low-coverage sequencing data in rice. We identified candidate genes for 18 associated loci through detailed annotation. This study shows that the integrated approach of sequence-based GWAS and functional genome annotation has the potential to match complex traits to their causal polymorphisms in rice.

Rice (*Oryza sativa* L) is an economically important crop that accounts for ~20% of the world's caloric intake. To be grown successfully under a variety of climatic conditions across the globe, breeders maintain rice at high genetic diversity. Second-generation sequencing tech-

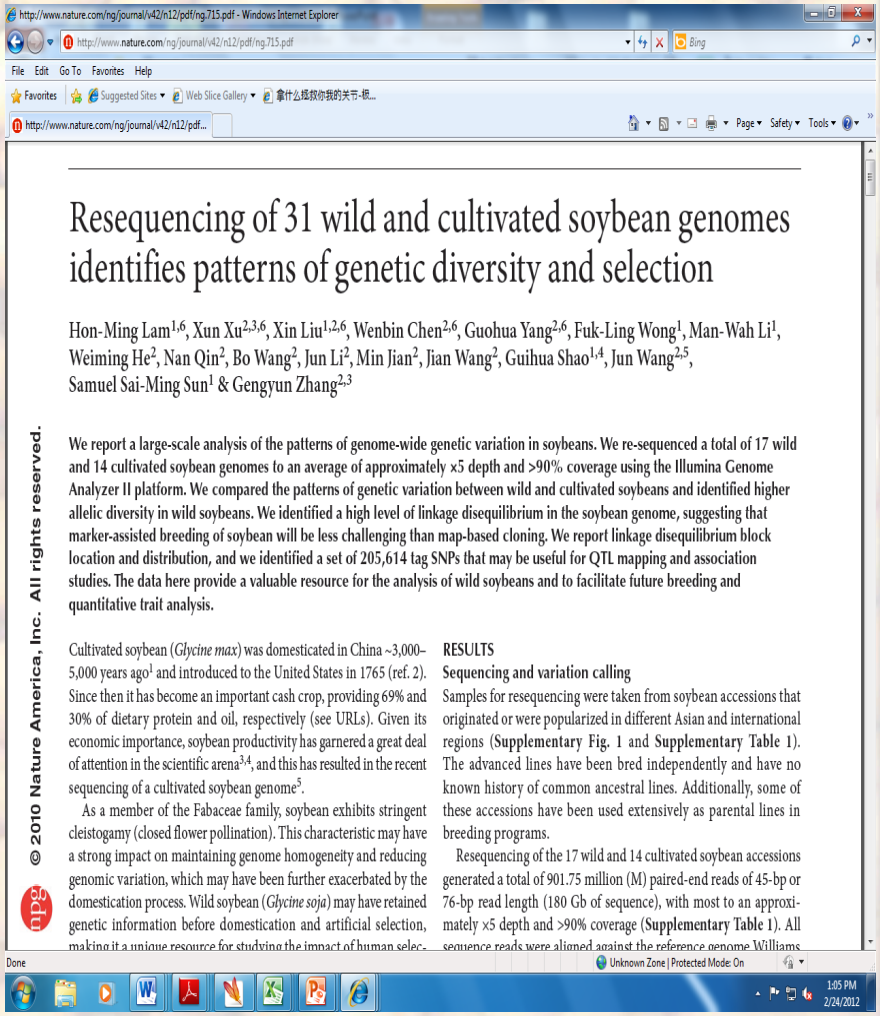
sequences of different gene alleles[4]. This approach was then used to detect SNPs and complex polymorphisms such as insertions and deletions (indels). Using this new method, we were able to project a map of genic variation onto the rice genome. This facilitated the discovery

What Can a Large Set of Sequenced Germplasm Lines Do for Soybean Breeders?

- **Genome-wide association study (GWAS)**
  - GWAS enables the mapping of causative loci to single nucleotide changes (QTN or QTIndel)
  - The resulting HapMap of this study will provide a complete view of the haplotype structure of soybean, pave the way for GWAS studies, and identify alleles underpinning phenotypic diversity across the whole genome
- **Broad impacts on soybean breeding**
  - Distinguish all genetic differences between any two lines. Make more precise crosses.
  - Identify rare alleles associated with valuable variants/mutations
  - Identify diagnostic SNPs/Indels for marker-assisted selection

Resequencing Studies in Soybean

- Lam et al. re-sequenced a total of 17 wild and 14 cultivated soybean genomes (5X depth; 90% coverage)
  - 5.9 million SNPs in wild soybean
  - 4.2 million SNPs in cultivated soybean
- Qiu et al. sequenced 25 accessions (per. comm.)
- Shoemaker et al. (87 ancestral lines and milestone cultivars in the US)



Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection

Hon-Ming Lam[1,6], Xun Xu[2,3,6], Xin Liu[1,2,6], Wenbin Chen[2,6], Guohua Yang[2,6], Fuk-Ling Wong[1], Man-Wah Li[1], Weiming He[2], Nan Qin[2], Bo Wang[2], Jun Li[2], Min Jian[2], Jian Wang[2], Guihua Shao[1,4], Jun Wang[2,5], Samuel Sai-Ming Sun[1] & Gengyun Zhang[2,3]

We report a large-scale analysis of the patterns of genome-wide genetic variation in soybeans. We re-sequenced a total of 17 wild and 14 cultivated soybean genomes to an average of approximately ×5 depth and >90% coverage using the Illumina Genome Analyzer II platform. We compared the patterns of genetic variation between wild and cultivated soybeans and identified higher allelic diversity in wild soybeans. We identified a high level of linkage disequilibrium in the soybean genome, suggesting that marker-assisted breeding of soybean will be less challenging than map-based cloning. We report linkage disequilibrium block location and distribution, and we identified a set of 205,614 tag SNPs that may be useful for QTL mapping and association studies. The data here provide a valuable resource for the analysis of wild soybeans and to facilitate future breeding and quantitative trait analysis.

Cultivated soybean (*Glycine max*) was domesticated in China ~3,000–5,000 years ago[1] and introduced to the United States in 1765 (ref. 2). Since then it has become an important cash crop, providing 69% and 30% of dietary protein and oil, respectively (see URLs). Given its economic importance, soybean productivity has garnered a great deal of attention in the scientific arena[3,4], and this has resulted in the recent sequencing of a cultivated soybean genome[5].

As a member of the Fabaceae family, soybean exhibits stringent cleistogamy (closed flower pollination). This characteristic may have a strong impact on maintaining genome homogeneity and reducing genomic variation, which may have been further exacerbated by the domestication process. Wild soybean (*Glycine soja*) may have retained genetic information before domestication and artificial selection, making it a unique resource for studying the impact of human selec-

RESULTS

Sequencing and variation calling

Samples for resequencing were taken from soybean accessions that originated or were popularized in different Asian and international regions (**Supplementary Fig. 1** and **Supplementary Table 1**). The advanced lines have been bred independently and have no known history of common ancestral lines. Additionally, some of these accessions have been used extensively as parental lines in breeding programs.

Resequencing of the 17 wild and 14 cultivated soybean accessions generated a total of 901.75 million (M) paired-end reads of 45-bp or 76-bp read length (180 Gb of sequence), with most to an approximately ×5 depth and >90% coverage (**Supplementary Table 1**). All sequence reads were aligned against the reference genome Williams

Vision for the future …

Re-sequencing a Core Set of USDA Soybean Germplasm

- Goals
  - sequence 3,000 germplasm lines/cultivars with 10 to 15 X redundancy and construct a genome-wide HapMap
  - extend to 5,000 soybean lines
  - multiple reference genomes: sequence, assemble and annotate 10 diverse soybean lines (80-100X) for structural and functional studies
  - Public-private partnership

Current Status

- 10 accessions have been sequenced (>10X depth).
- 100 PIs/cultivars will be sequenced soon.
- 300 PIs will be sequenced for WGAS study.
- Large scale resequencing project (3,000 – 5,000 soybean germplasm lines) to be started in 2013.
- User-friendly tools have been developed by Prof. Dong Xu's group to browse, search and visualize sequence variations (SoyKB).

SoyKB: Soybean Knowledge

dev.soykb.org/search/snp.php

soykb

Logged in as **joshitr** | Logout

Home

Optional: Geno

52000

Search

SoyKB: Soybean Knowledge

dev.soykb.org/snp.php?chromosome_number=01&start=52000&end=57000

SNP Search Results

Quick Search | Gene Card | Go

Gm01
Range: 52000 - 57000

Data Source: Williams vs Forrest (7947)

SNPID	Chromosome	Scaffold	Position	Refer Base	Consensus Base	Consensus Quality	Read Depth	Highest Quality	Hit Start	Hit End	Gene(s) in Range
SNP10250	Gm01	scaffold_166	52424	A	R	58	32	63	3704522	3704554	
SNP10251	Gm01	scaffold_166	52467	C	Y	48	45	63	3704576	3704608	

Data Source: Magellan vs PI (2631)

SNPID	Chromosome	Position	Refer Base	Consensus Base	Consensus Quality	Read Depth	Highest Quality
SNP01	Gm01	52601	G	C	36	3	63

Data Source: GWAS

Chromosome	Position	Ref	W01	W02	W03	W04	W05	W06	W07	W08	W09	W10	W11	W12	W13	W14	W15	W16	W17	C01	C02	C08	C12	C14	C16	C17	C19	C24	C27	C30	C33	C34	C35	Gene(s) in Range
Gm01	52271	T	A	T	T	A	T	A	T	A	T	A	T	T	T	A	A	A	T	W	T	T	T	T	A	T	T	T	T	T	T	T	T	Glyma01g00270.1
Gm01	52470	G	G	G	G	G	C	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	Glyma01g00270.1
Gm01	52565	G	G	G	G	G	C	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	Glyma01g00270.1
Gm01	52601	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	C	C	G	G	G	G	G	-	G	C	G	C	G	G	Glyma01g00270.1
Gm01	52646	G	G	A	R	G	G	G	G	G	G	G	A	A	A	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	Glyma01g00270.1
Gm01	52781	A	A	A	A	A	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	Glyma01g00270.1
Gm01	54332	C	C	C	C	C	T	-	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	Glyma01g00270.1
Gm01	54525	A	A	G	R	A	A	-	A	A	A	A	G	G	G	G	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	Glyma01g00270.1
Gm01	54532	A	A	M	M	A	A	-	A	A	A	A	C	C	C	C	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	Glyma01g00270.1
Gm01	54567	C	C	C	C	C	C	C	C	C	C	C	A	A	C	C	C	M	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	Glyma01g00270.1
Gm01	54679	A	G	R	R	G	A	G	A	G	A	G	A	A	A	A	G	G	A	A	A	A	A	A	G	A	A	A	A	A	A	A	A	Glyma01g00270.1
Gm01	55052	T	T	W	-	T	T	T	T	T	T	T	-	A	A	A	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	Glyma01g00270.1
Gm01	55085	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	C	C	C	-	C	C	C	C	C	C	C	C	-	C	C	C	C	Glyma01g00270.1
Gm01	55133	T	T	T	T	T	A	-	T	T	T	T	T	T	T	T	T	T	T	-	T	T	T	T	-	T	T	T	T	T	T	T	-	Glyma01g00270.1
Gm01	55218	G	G	G	G	G	G	G	G	G	G	G	-	T	K	-	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	Glyma01g00270.1
Gm01	55512	C	-	C	T	T	T	T	C	T	T	T	T	T	T	-	-	T	C	T	T	C	C	C	T	C	C	Y	T	C	T	C	Y	Glyma01g00270.1
Gm01	55560	G	G	G	G	G	T	G	G	G	G	G	G	G	G	-	G	G	-	G	G	-	G	G	G	G	G	G	G	G	G	G	G	Glyma01g00270.1
Gm01	56113	C	C	C	C	C	T	C	C	C	C	C	C	C	Y	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	Glyma01g00270.1

Data Source: Soja

SNPID	Chromosome	Position	Refer Base	Consensus Base	Gene(s) in Range
GSSNP220	Gm01	52271	T	A	Glyma01g00270.1
GSSNP221	Gm01	52369	T	A	Glyma01g00270.1
GSSNP222	Gm01	52470	G	C	Glyma01g00270.1
GSSNP223	Gm01	52565	G	C	Glyma01g00270.1
GSSNP224	Gm01	54474	G	A	Glyma01g00270.1
GSSNP225	Gm01	54561	A	T	Glyma01g00270.1
GSSNP226	Gm01	54574	G	A	Glyma01g00270.1
GSSNP227	Gm01	55164	G	A	Glyma01g00270.1
GSSNP228	Gm01	55297	T	C	Glyma01g00270.1
GSSNP229	Gm01	55369	C	T	Glyma01g00270.1
GSSNP230	Gm01	55512	C	T	Glyma01g00270.1
GSSNP231	Gm01	56057	A	G	Glyma01g00270.1
GSSNP232	Gm01	56113	C	T	Glyma01g00270.1
GSSNP233	Gm01	56326	A	C	Glyma01g00270.1
GSSNP234	Gm01	56699	C	A	Glyma01g00270.1

SNPs: Forrest & Williams 82

SNPs: Magellan & PI2631

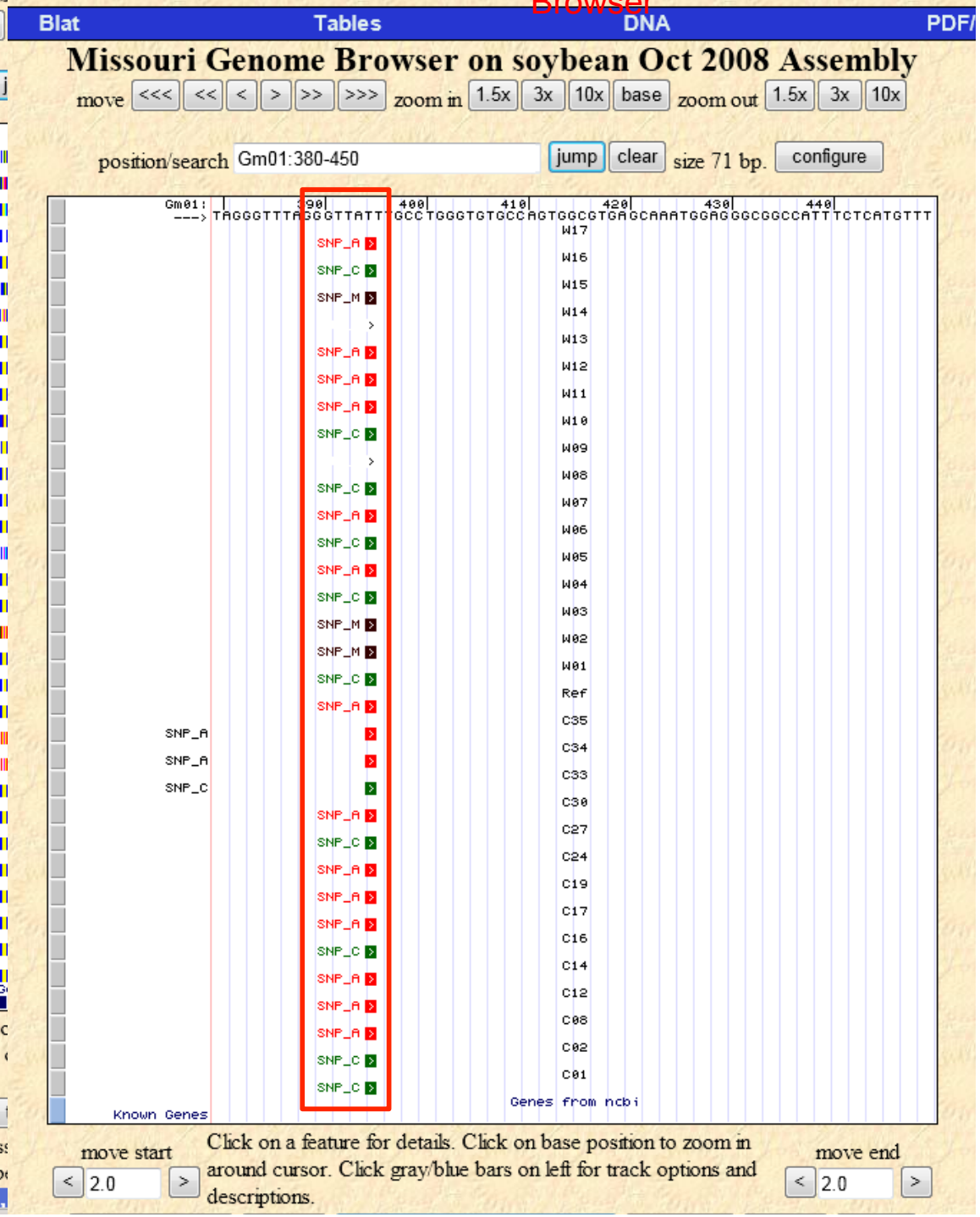
SNPs: GWAS 31 genotypes from BGI

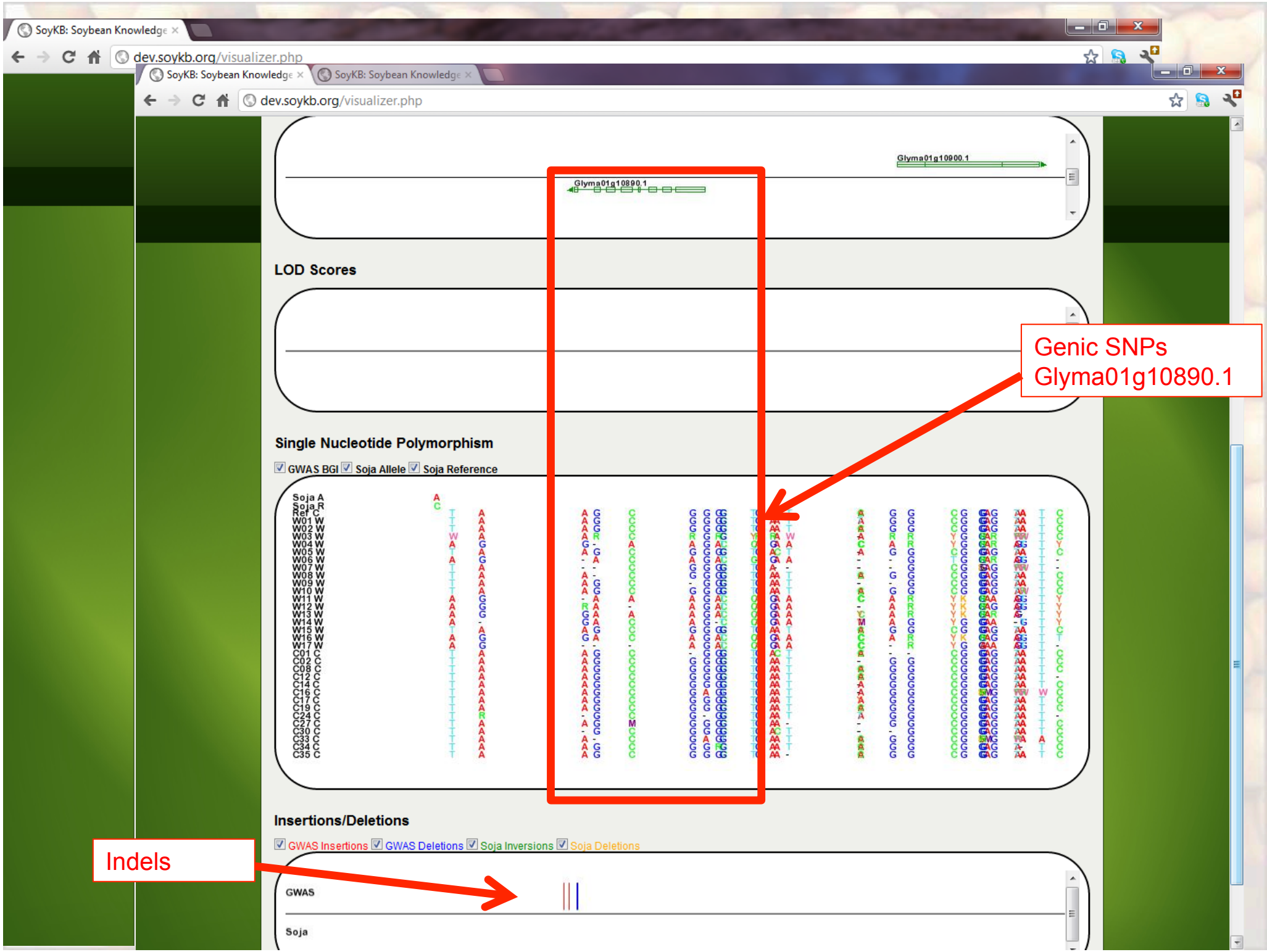
SNPs: G. Soja

Missouri Genome Browser

GWAS SNP Data for 31 Genotypes from BGI in Genome Browser

Color Code:
A Red
T Yellow
C Green
G Blue
K Purple
W Cyan
M Brown
Y Pink
R Lime
S Grey
- Black




SoyKB: Soybean Knowledge
dev.soykb.org/visualizer.php
Glyma01g10900.1
Glyma01g10890.1
LOD Scores
Single Nucleotide Polymorphism
GWAS BGI
Soja Allele
Soja Reference
Genic SNPs
Glyma01g10890.1
Insertions/Deletions
GWAS Insertions
GWAS Deletions
Soja Inversions
Soja Deletions
Indels
GWAS
Soja

Acknowledgements

University of Missouri

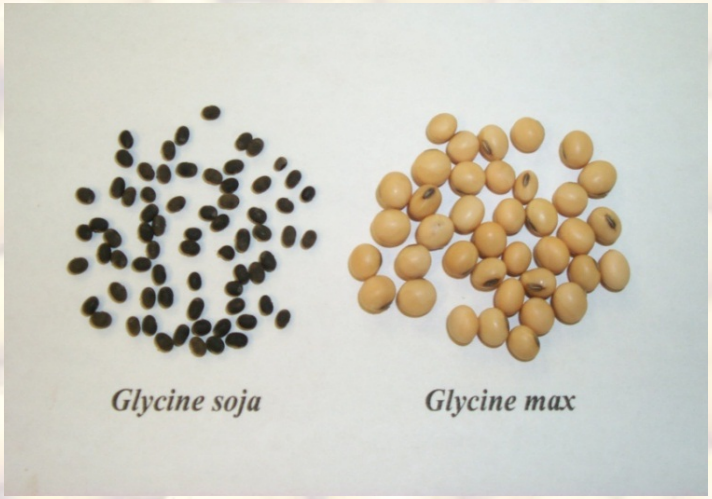
- Xiangyang Xu
- Dong Xu
- Trupti Joshi
- Tri Vuong
- Jessica Frank
- Grover Shannon
- M.S. Pathan

Beijing Genome Institute

- Ye Tao
- Liang Zeng

University of Georgia

- Roger Boerma
- Steve Finnerty



Next ?
Nanopore sequencers

48 KB genome of bacteriophage could be sequenced as a complete fragment

Future vision is to sequence an entire human genome in 15 minutes