Genome Sequencing and Analysis of Wild Soybean (*Glycine soja* Sieb. and Zucc.)

Suk-Ha Lee[1,2], Moon Young Kim[1], Yang Jae Kang[1], Minyoung Yoon[1], Sue Kyung Kim[1], Hyun-Ju Jang[1], Kyujung Van[1]

[1] Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Korea
[2] Plant Genomics and Breeding Institute, Seoul National University, Seoul 151-921, Korea

Introduction

Wild soybean (*Glycine soja* Sieb. and Zucc.), together with cultivated soybean (*G. max* (L.) Merr.), belongs to the family Legumiosae. *G. soja* is generally considered to be closest wild relative of *G. max*. *G. soja* and *G. max*, both have 20 chromosomes (2n = 40), hybridize easily, exhibit normal meiotic chromosome pairing, and generate viable fertile hybrids. However, wild and cultivated soybeans differ with respect to several plant morphological characteristics.

The genomes of more than 6 crop species have been sequenced so far. However, the genome sequence of a single crop strain does not allow an understanding of the origins of separate genes involved in complex traits. Furthermore, it is challenging to decipher the processes involved in crop domestication as domestication involves developmental changes. Thus, the genome sequences of wild species will provide key information about the genetic elements involved in speciation and domestication.

Recently, remarkable advances in high-throughput DNA sequencing have enabled generation of several orders of magnitudes more sequence data within a relatively short time than traditional Sanger method. Among them, the two massive parallel sequencing (MPS) platforms, Illumina Genome Analyser (Illumina-GA) and Roche Genome Sequencer FLX (GS-FLX), using reversible terminators and pyrosequencing, respectively, are in widespread use for genomic, biological, and medical studies.

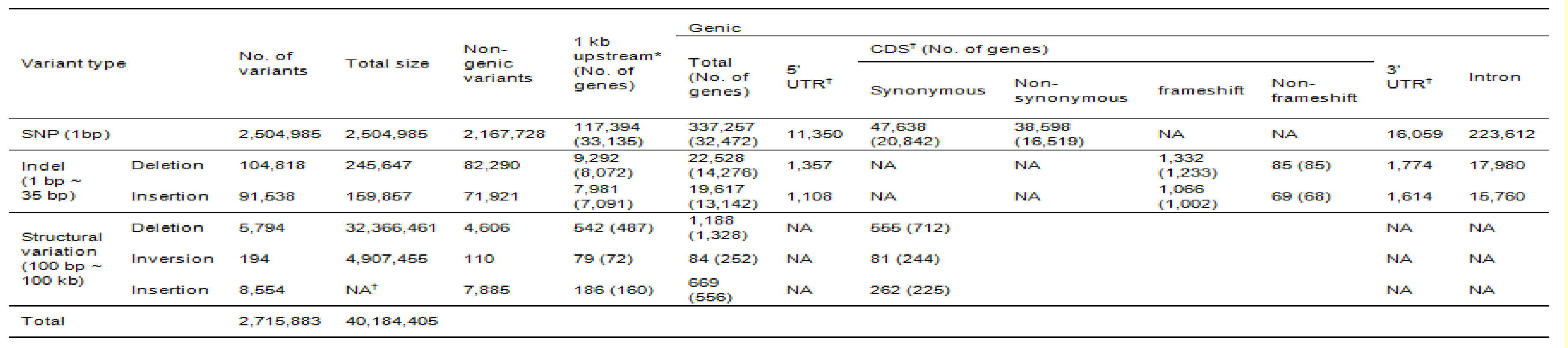
In this study, we sequenced the whole genome of wild soybean (*G. soja*) using two representative massive parallel sequencing platforms, Illumina-GA and GS-FLX, and analyzed the *G. soja* genome sequences in detail to catalogue the wealth of genomic variations between *G. max* and *G. soja*. This detailed analysis offers a primary glimpse of soybean domestication history, which suggested that divergence of *G. soja* and *G. max* predated the soybean domestication.

Results & Discussion

1. Identification of SNP, indel and structural variation

About 2.5 million SNPs between *G. max* and *G. soja* were predicted (supported by more than four reads) via multiple stringent filtering criteria (Table 1). Non-coding genic regions contained 251,021 SNPs with 27,409 in 5' and 3' untranslated regions and 223,612 in introns. A total of 86,236 SNPs were classified as coding sequence variants. Frequency of SNPs in both wild and cultivated soybeans was 2.67 SNPs per 1 kb. The *G. soja* genome contains 196,356 indels (-35 to +14-bp) compared with Glyma1.01 (Table 1). Single-base-pair indels are the most frequent type. Only 21.46% of the indels are positioned within genic boundaries (Table 1). The 2,398 indels in coding sequences cause frameshifts in 2,235 genes and indels were located throughout the *G. soja* genome at a density of 1 indel per 4.8 kb. We detected 5,794 deletions and 194 inversions in the range of 0.1~100-kb and predicted the presence of 8,554 insertions in the *G. soja* genome (Table 1). About 48.11% of the deleted regions in the *G. soja* genome contained repetitive elements. In the range of 1~2-kb and 10~20-kb, about 50% of the deletion events involved retrotransposons. The 32 Mb fragments present in *G. max* but absent in *G. soja* harbours 712 coding sequences in 555 deletion events (Table 1).

Table 1. Summary of differences between the G. soja and G. max genomes



Variant type		No. of variants	Total size	Non-genic variants	1 kb upstream* (No. of genes)	Genic							
						Total (No. of genes)	5' UTR†	CDS† (No. of genes)				3' UTR†	Intron
								Synonymous	Non-synonymous	frameshift	Non-frameshift		
SNP (1bp)		2,504,985	2,504,985	2,167,728	117,394 (33,135)	337,257 (32,472)	11,350	47,638 (20,842)	38,598 (16,519)	NA	NA	16,059	223,612
Indel (1 bp ~ 35 bp)	Deletion	104,818	245,647	82,290	9,292 (8,072)	22,528 (14,276)	1,357	NA	NA	1,332 (1,233)	85 (85)	1,774	17,980
	Insertion	91,538	159,857	71,921	7,981 (7,091)	19,617 (13,142)	1,108	NA	NA	1,066 (1,002)	69 (68)	1,614	15,760
Structural variation (100 bp ~ 100 kb)	Deletion	5,794	32,366,461	4,606	542 (487)	1,188 (1,328)	NA	555 (712)				NA	NA
	Inversion	194	4,907,455	110	79 (72)	84 (252)	NA	81 (244)				NA	NA
	Insertion	8,554	NA†	7,885	186 (160)	669 (556)	NA	262 (225)				NA	NA
Total		2,715,883	40,184,405										

2. Effect of mapping depth and chromosomal distribution of SNPs/indels

The chromosomal distribution of SNPs and indels was non-uniform and less SNPs and indels occurred in pericentromeric regions that are highly repetitive in soybean (Fig. 1A). We filtered out nucleotide variants for which the mapping depth was above 200 in the pericentromeric regions to raise the quality of variation detection. Genome coverage was directly proportional to the mapping depth of short DNA reads (Fig. 1B). For a read threshold ≥ 1, the genome coverage reached a plateau of 98% at a mapping depth of 20-fold, while read thresholds of ≥ 3 and ≥ 5 showed the genome coverage plateaus at a mapping depth of 30-fold. Most of the SNPs were identified at a mapping depth of 30-fold for both of ≥ 3 and ≥ 5 read thresholds but SNP discovery increased gradually up to a mapping depth of 42-fold (Fig. 1B). These results indicate the theoretical maximum of genome coverage and SNP calling at effective mapping depth reached in this *G. soja* genome sequencing using MPS technology.

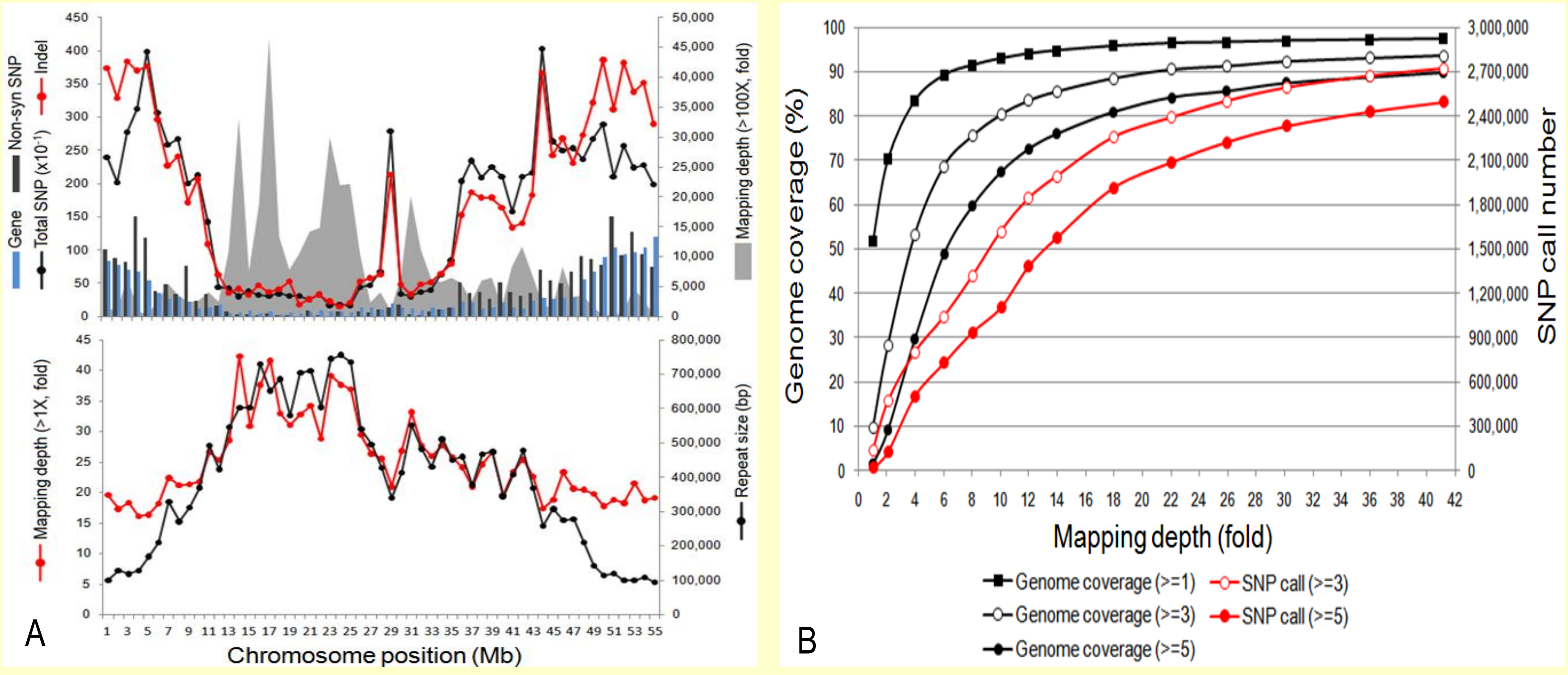


Fig. 1. Distribution of sequence variation on chromosome 1 of G. soja. (A) Black and red lines indicate total SNPs and the number of indels, respectively. To fit the lines and bars in one graph using a binning unit of 1 Mb on the x-axis, the SNP number was scaled to 1/10 and the repeat size was scaled to 1/50. (B) Effect of mapping depth (to reference genome) on genome coverage and SNP detection. The numbers of detected SNPs according to mapping depth are indicated by red lines.

3. Genomic difference and divergence between G. max and G. soja

The sequence difference between *G. max* and *G. soja* was 35.2 Mb (3.76% of 937.5 Mb), consisting of 2.5 Mb (0.267%) of substituted bases, 406 kb (0.043%) of inserted/deleted bases, and 32.3 Mb (3.45%) of large deleted sequences (Table 1). In addition, we calculated the theoretical divergence time between the genomes of IT182932 (*G. soja*) and Williams 82 (*G. max*) and found that *G. soja* and *G. max* diverged at 0.267 ± 0.03 MYA based on a *Ks* distribution with 6,780 synonymously changed neutral genes, suggesting that divergence between *G. soja* and *G. max* predated the domestication of soybean. Thus, our data suggests that the *G. soja*/*G. max* complex is at least 270 thousand years old (Fig. 2), although it is widely accepted that there is no undomesticated *G. max* without domestication, which is estimated to have occurred 6,000-9,000 years ago.

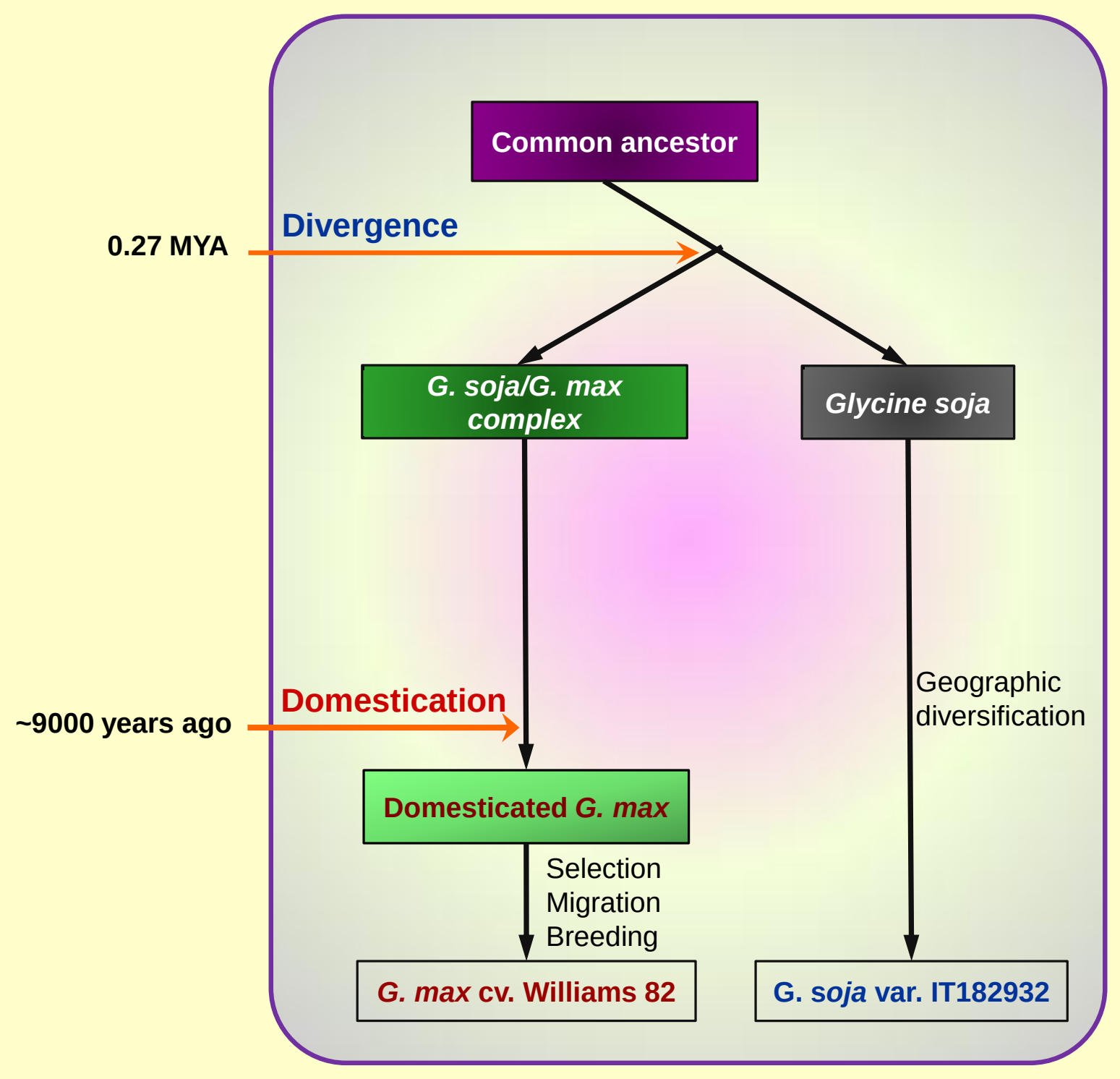
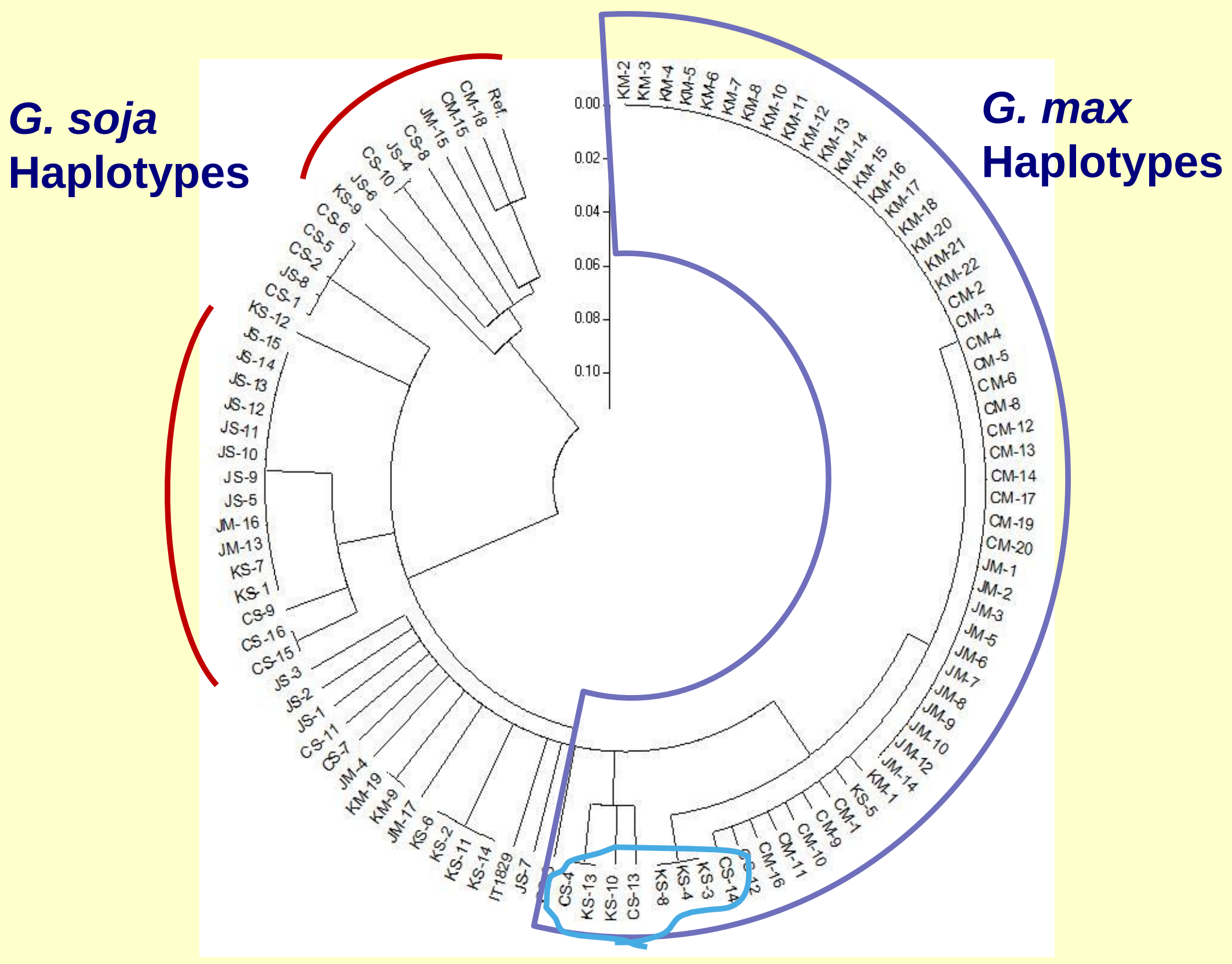


Fig. 2. Soybean domestication history. G. max is generally believed to have been domesticated from its wild relative, G. soja, 6,000-9,000 years ago. We calculated that the G. soja/G. max complex diverged from the common ancestor of the two Glycine species at 0.27 MYA. Divergence between G. soja and G. max thus predated domestication, indicating that cultivated soybean was domesticated from pre-existing G. soja/G. max complex.

4. Understanding relationship between G. max and G. soja and soybean domestication history

We collected 104 *G. max* and *G. soja* samples that were distributed from China, Korea to Japan for population genetic approach. Seventeen single-copy nuclear genes were selected and sequenced. *G. soja* from both China and Korea showed higher variations than other samples, which was also supported by phylogenetic analysis . After chloroplast (cp) genome of *G. soja* (var. IT182932) was sequenced, cp-specific primers were designed based on two cp genome sequences of *G. max* (PI 437654) and *G. soja* (IT182932) (Fig. 3). These nucleotide variations in nuclear and chloroplast sequences provided valuable information on the geographical divergence process of soybean and its wild relatives and soybean domestication history.



Wild soybean having haplotypes similar to *G. max*

Fig. 3. Phylogenetic analysis of soybean chloroplast (cp) genome with 104 G. max and G. soja samples. Seven cp haplotypes were identified and some G. soja samples belonged to the haplotype similar to G. max.

Acknowledgements

This work was supported by a grant from the BioGreen 21 Project (code no. 20080401034010), Rural Development Administration, the Republic of Korea. Dr. K. Van is the recipient of a fellowship from the BK21 program granted by the Ministry of Education & Human Resources Development (ME & HRD), the Republic of Korea.