A map of genetic variation from 781 soybean genomes

Soon-Chun Jeong[1,*], Myung-Shin Kim[1], Roberto Lozano[2], Ji Hong Kim[1], Dong Nyuk Bae[1], Man Soo Choi[3], Soo-Kwon Park[3], Michael A. Gore[2], and Jung-Kyung Moon[4]
[1]Bio-evaluation center, Korea Research Institute of Bioscience and Biotechnology, Cheongju, Korea
[2]Plant Breeding and Genetics Section, School of Integrative Plant Science, Cornell University, USA
[3]National Institute of Crop Science, Rural Development Administration, Wanju, Korea
[4]Agricultural Genome Center, National Academy of Agricultural Sciences, Jeonju, Korea
[*]Corresponding author: scjeong@kribb.re.kr

Abstract

Soybean is an economically and environmentally important major crop worldwide. It is a predominant plant protein and oil source of both food and feed and has capacity to fix atmospheric nitrogen by intimate symbioses with microorganisms. Here we present a fine genome-wide variation map in 781 accessions including 418 domesticated (*Glycine max*) and 345 wild (*Glycine soja*) soybeans and 18 of their natural hybrids. We identified 31 million single nucleotide polymorphisms and 5.7 million small indels that contribute to within- and between-population variation. We describe a comprehensive characterization of the geographic and functional differentiation of rare and common genetic variants with insights into the domestication history of soybean and detection of domestication-selective sweeps. We show that this resource enables us to increase marker density of existing data sets for improving the resolution of association studies.

Keywords: association, domestication, soybean, variation

Introduction

Soybean (*Glycine max* [L.] Merr.) is an important crop species. The cultivation of soybean has been historically confined to East Asia. Its cultivation area has been recently expanded to North America, South America, and India, positioning it as one top crop in terms of growing area worldwide. After the release of the draft soybean genome sequence, efforts to map soybean genetic variation by single nucleotide polymorphism (SNP) array genotyping have resulted in the global picture of common and rare SNPs across the genome. However, those data have been poorly used as an integrated manner to serve as haplotype information by imputation approaches that enrich the above SNP genotype data with whole genome SNP data. The resultant genomic data set could also be leveraged as a reference panel to enrich sparse data sets generated from platforms such as low-density SNP array, genotyping-by-sequencing (GBS) or skim sequencing. In addition, genetic variation of wild soybean, which contains a large amount of untapped and unexplored soybean diversity, remains poorly characterized relative to that of domesticated soybean.

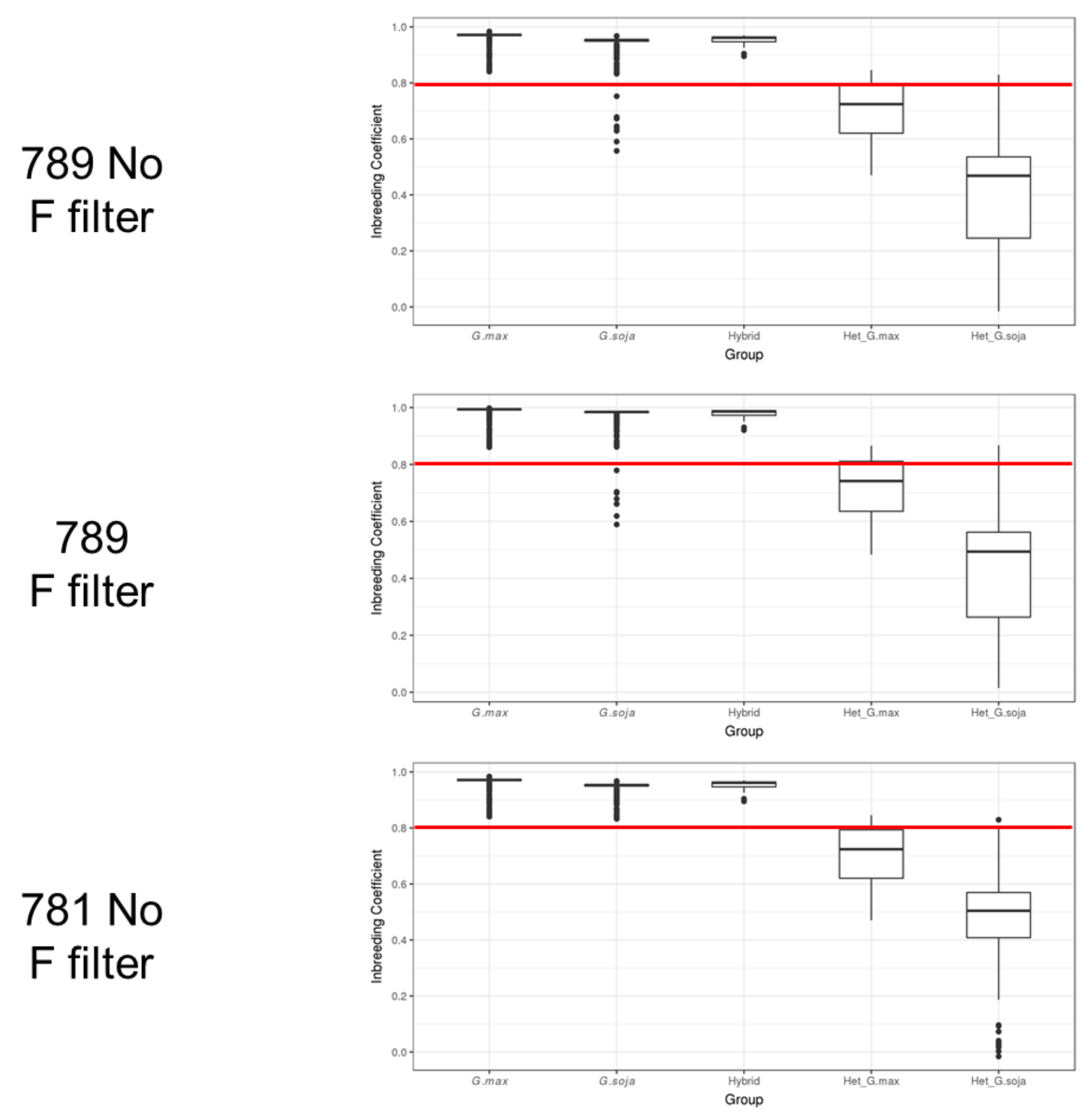
Here we report genomes of 781 soybean accessions consisting of 418 *G. max*, 345 *G. soja*, and 18 hybrid (*G. max* x *G. soja*) accessions obtained through high-coverage (> 13x) whole-genome sequence data. We conducted a comprehensive characterization of genetic variation including profiling and functional significance of rare and common genetic variants and detection of domestication-selective sweeps. We are then attempting to show the usefulness of our data in soybean genetics by imputing the SNP data set from 50K US SNP array genotyping using variants identified here for association analyses of protein and oil traits.

Materials and Methods

Resequencing data from a total of 855 samples were obtained for initial variant calling in this study. We then essentially followed procedures described in the Genome Analysis Toolkit (GATK) Best Practices for data pre-processing and variant calling. Principal component, fastStructure Bayesian clustering, and Neighbor-joining tree construction was used to infer population structure. Historical recombination rate, linkage disequilibrium decay, genomic evolutionary rate profiling, Sorting Intolerant From Tolerant 4G, cross-population composite likelihood ratio test, and genome-wide association studies were performed to infer nucleotide variation pattern, identify selective sweeps, and find association between genotypes and phenotypes.

Results and Discussion

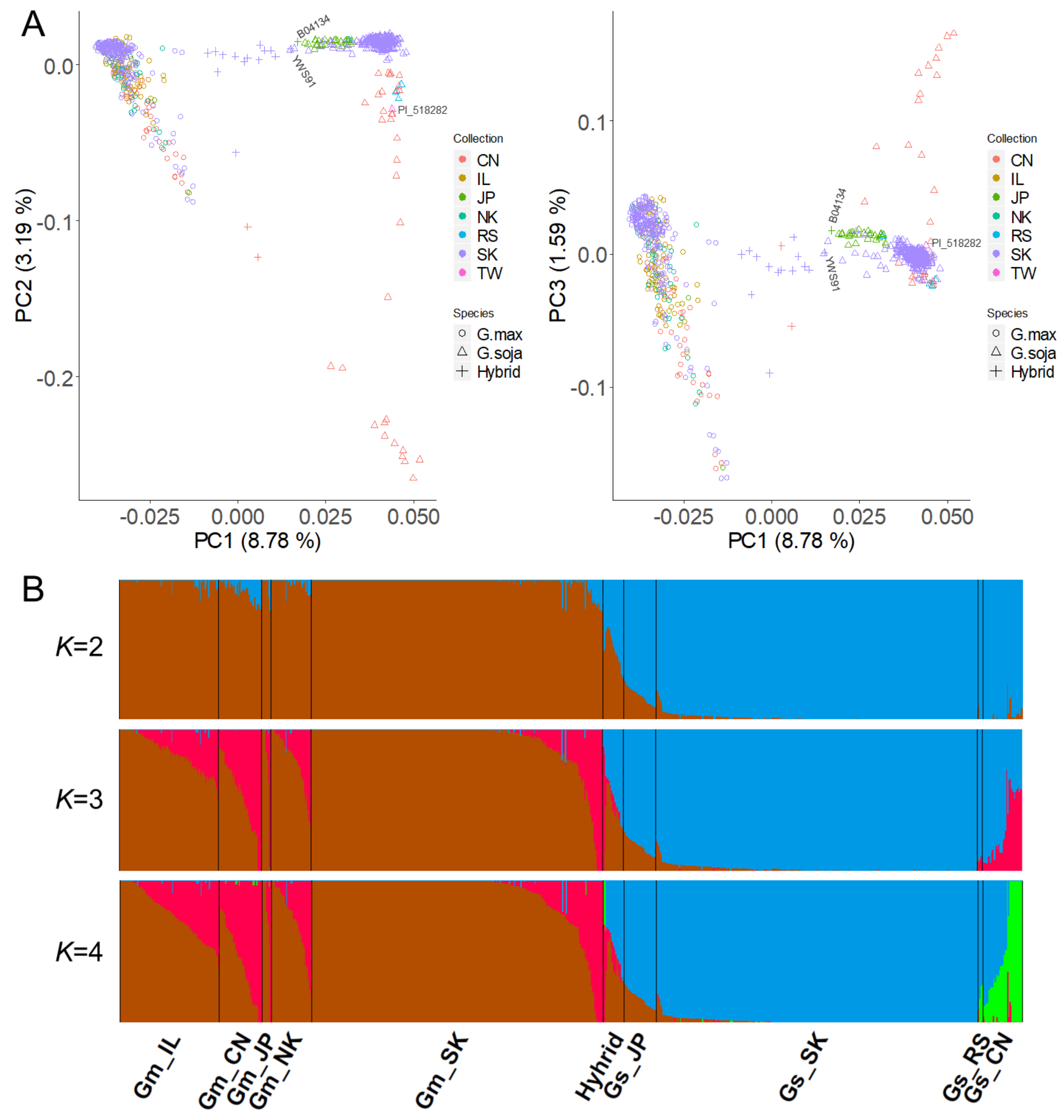
High heterozygous samples were excluded from further analyses. Distributions of inbreeding coefficient per individuals in subgroups divided by species and heterozygosity indicated that most samples excluded by the high heterozygosity showed an inbreeding coefficient of less than 0.8. Final 781 accessions consisted of 418 *G. max*, 345 *G. soja*, and 18 hybrid (*G. max* x *G. soja*) accessions.



<Distributions of Inbreeding coefficient per individual in *Glycine max, Glycine soja*, hybrid, high heterozygous *G. max* and high heterozygous *G. soja* groups>

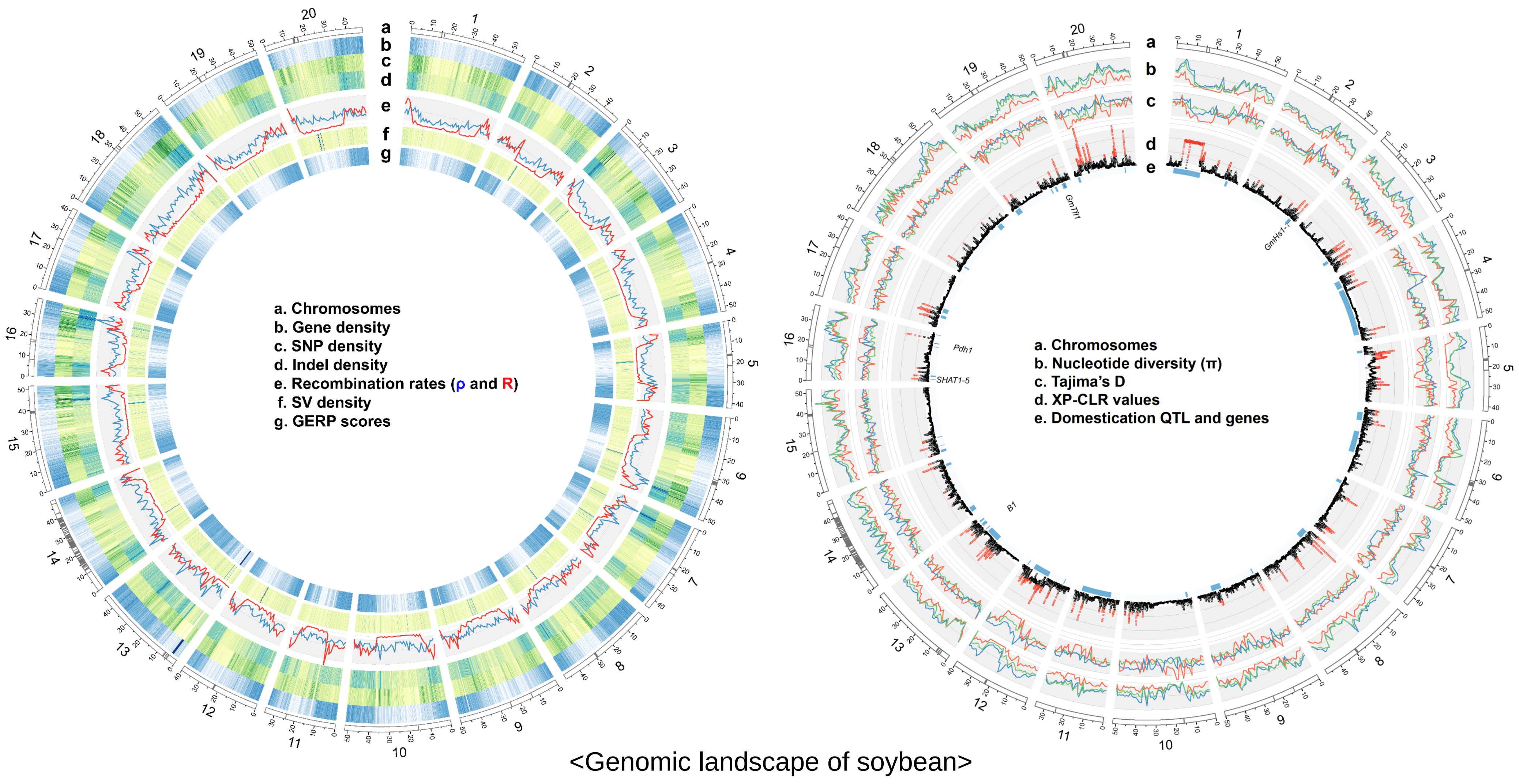
Results and Discussion

As observed in the previous study, the 781 accessions were clearly divided into one *G. max* group and two *G. soja* groups with a distinct subgroup of *G. soja* accessions collected from the middle region of the Yellow River basin.



<Population structure of the 781 haplotype soybean accession set. **A**. Principal components (PC) of SNP variation. The plots show the first three principal components. **B**. fastSTRUCTURE plots>

All chromosomes had lower recombination near the centromere repeat regions, which are presumed to be percentromeric regions spanning more than 10 Mbp, relative to that in euchromatin regions. Overall chromosomal distribution patterns of gene density, SNP density, indel density, and Genomic evolutionary rate profiling (GERP) scores were similar to that of recombination rates. A total of 183 domestication-selective sweep regions were detected. In this selfing species, overall deleterious alleles among landraces relative to wild soybean accessions have been drastically reduced by up to almost 35%. Purging of deleterious alleles from the domesticated soybean has been further enhanced in selective sweep regions. Finally, we have shown that our high-quality map of genome variation in soybean could be used as a reference panel for the imputation of genotypes to improve the existing GWAS for oil and protein traits with 50K US SNP array genotyping data.



<Genomic landscape of soybean>