Genetic Analysis of Genomic Locations Underlying Domestication Transitions in Soybean

Soon-Chun Jeong
Korea Research Institute of Bioscience and Biotechnology
scjeong@kribb.re.kr

I. Introduction

Ways to ensure sustainable food production

- Genetic diversity of cultivars should be maintained or increased to combat with new pests and diseases and the change of local climates.

Sources of diversity

- Wild progenitors are usually more diverse than their cultigens, however utility of wild progenitors in breeding programs are hindered by undesirable domestication alleles.
- Therefore, understanding of domestication process is essential for the harnessing of untapped genetic diversity from wild progenitors for enhancement of elite cultivars.
- An ultimate point exploring the wild progenitors is re-domestication of the cultigen.

Means to discover and utilize the domestication alleles:

- The combination of quantitative trait locus mapping, genome-wide association study, and genome scan

II. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes

III. Current efforts for genetic analysis of genomic locations underlying domestication transitions in soybean

- Re-analysis of genomic locations underlying domestication transitions using a larger number of soybean accessions
- Development and use of a large soybean SNP genotyping array

How many genomes and what depths of resequencing are appropriate for population-level analysis of plant genetic resources?

<table>
<thead>
<tr>
<th>ARTICLE</th>
<th>Genes</th>
<th>Resource</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu et al. (2012)</td>
<td>305</td>
<td>Resequencing 30 accessions of cultivated and wild soybean cultivars for identifying applications</td>
</tr>
<tr>
<td>Gan et al. (2011)</td>
<td>15</td>
<td>Multiple reference genomes and transcriptomes for Arabidopsis thaliana</td>
</tr>
<tr>
<td>Hufford et al. (2012)</td>
<td>40</td>
<td>Comparative population genetics of marine domestication and improvement</td>
</tr>
<tr>
<td>Huang et al. (2012)</td>
<td>410</td>
<td>A map of rice genome variation reveals the origin of cultivated rice</td>
</tr>
</tbody>
</table>

Regardless of the number, all 4 datasets were equally powerful for pinpointing candidate genes

Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes

The Next Generation Sequencing (NGS) platforms facilitate generation of reference genome sequences as well as full resequencing of population samples of genomes.
Soybean is an important worldwide plant source of dietary protein and oil and its capability of nitrogen fixation during symbiosis with rhizobia plays an important role in establishing sustainable agriculture systems.

Cultivated soybean (Glycine max) is thought to have been domesticated from wild soybean (Glycine soja) with distribution in East Asia, including Korea, Japan, and most parts of China, as early as 7000–9000 years ago. Soybean has undergone radical phenotypic changes in the seed size, colour, shattering, seed dormancy, flowering time, and plant architecture during the process of domestication.

However, while localization of the major causative genes, responsible for these traits, would facilitate improvements in soybean using marker-assisted breeding, Genome-wide diversity measurements have consistently indicated an approximately two times higher genetic variation in wild than that in cultivated soybeans (Byten et al. 2006; Laim et al. 2010; Chuah et al. 2014).

The main objective of this study was to provide comprehensive resequencing data of a diverse group of 10 cultivated and 6 wild soybean genotypes.

DNA Res. (2014) 21: 153

Population structure and domestication of soybean

The wild soybeans were collected from different provinces in South Korea. Non-duplicated cultivated accessions were obtained from the National Soybean Germplasm Collection (Korea), and 6 wild collections were collected from wild soybean (G. soja) populations in different provinces in South Korea.

Sequenced 16 accessions consisting of 6 pure breeds, 4 landraces, and 6 wild collections.

Breeding history of pure bred accessions

Population structure and domestication

Data of resequencing

Population structure and domestication of soybean

Theme of this study

- Have been interested in establishing a research framework to harness the untapped genetic diversity from wild soybean
- The main objective of this study was to provide comprehensive genome variation and genotyping data of wild soybean across the entire genome
- The comprehensive data was further used to investigate evolutionary aspects of soybean including population structure and domestication transitions by population genomics analysis (genomic scans)

Wild and cultivated soybeans.
Cultivated soybean population is a subclade of wild soybean population rather than an independent species or subspecies.

Incorporation of data from 17 wild and 14 cultivated soybean genomes resequenced at a low depth

- Cultivated soybeans formed three well-defined subgroups, each of which contained at least one accession out of the 16 accessions examined in this study.

Linkage disequilibrium patterns

- LD decayed to half of its maximum value at about 150 kb for wild soybeans and at about 350 kb for cultivated soybeans.
- The LD distance of the wild soybeans was extremely longer than those of wild plants analyzed at the genome level to date, which are usually within <100 kb.
- Association studies will be less challenging.

Detection of selective sweeps

- Distribution of ROD and \( F_{ST} \) for cultivated soybeans relative to wild soybeans across 20 chromosomes.

\[
\text{ROD} = \frac{\pi_{\text{wild}} - \pi_{\text{cultivated}}}{\pi_{\text{wild}}}
\]

\[
F_{ST} = \frac{\pi_{\text{within}} - \pi_{\text{between}}}{\pi_{\text{within}}}
\]

- 206 CDRs
- Total of 3,068 genes
- An average of 14.9 genes
- The longest fragment extended to 880 kb
- The median was 140 kb
- Successfully identified homologs of several canonical domestication genes such as \( \text{tb1} \), \( \text{tga1} \), \( \text{sh4} \), and \( \text{fw2.2} \) in our putative artificial selection gene set.
Features of candidate domestication genes

- Genes underlying critical domestication traits often encode transcription factors, and the causal mutations are more likely functional modification rather than the loss or gain of function.
- Selective sweep for loss-of-function mutations was not as strong or complete as critical functional modifications.

Current efforts for genetic analysis of genomic locations underlying domestication transitions in soybean

- Re-analysis of genomic locations underlying domestication transitions using a larger number of soybean accessions
- Development and use of a large soybean SNP genotyping array

Population structure of the expanded population with 39 resequenced soybeans

- Present 16 genome resequencing data + 23 additional high-depth genome resequencing data
- The phylogenetic tree indicates that this population likely represents worldwide soybean diversity.

Genomic locations underlying domestication transitions in soybean

- Direct comparison between the previous and current analyses is not possible because the two analyses used different numbers of soybean genomes, different cut-off criteria, and different versions of reference genome sequences.
- The number (2,049) of candidate domestication genes were reduced to two thirds of that from the analysis of the previous small population.
- More than 50% of the 2,049 were also detected in the previous small population.
Validate the SoyaSNP array, which contains the 180,961 SNPs, using 222 diverse soybean lines.

170,233 of them are high-quality suitable for further genetic analyses.

Distribute the converted 170,233 SNPs on the Williams 82 reference sequence (Wm82.a2.v1).

In general, markers were well-distributed over the chromosomes with a lower number of markers in many centromeric regions and a higher number of markers near the telomeres.

Overall, >77% of the 170,233 SNPs resided within genes or within 5 kb upstream or downstream of genes and at least one SNP was put on the array for >86% of >46,000 genes initially predicted from the Williams 82 genome sequence.

The cultivated and wild soybeans formed two clearly separated subclades as previously shown.

Intermediate morphology types between cultivated and wild soybeans are not an separated group but just a natural hybrid by the following two evidences:

- Twelve of the putative hybrid soybeans were grouped between the cultivated and wild soybean clades.
- Without the putative hybrid soybeans, private SNPs from direct comparison between the cultivated and wild soybeans increased approximately one and half times.

Distribution of SNPs on soybean genome

Utilization of SoySNP array for Genome-wide association studies (GWAS)

Both wild and cultivated soybean populations showed the strongest association at the E1 locus.

Genetic distances between 133 soybean accessions

The validation results suggest that this array is a rare example that has such a density of higher than 100,000 SNPs for a plant species (in press in The Plant Journal).

High-density SNP array enables us to perform various genetic analyses such as constructing diversity assessments of diverse populations, high-resolution genetic maps, GWAS, and improvement of reference genome.

We are currently analyzing several biparental populations for constructing high-resolution genetic linkage maps as well as thousands of wild and cultivated soybeans deposited in the National Agrobiodiversity Center in Korea for GWAS.

Utilizability of the SoyaSNP array

Applicability of the SoyaSNP array

Both wild and cultivated soybean populations showed the strongest association at the E1 locus.

Contributors

Namhee Jeong, Woo Kyu Lee, Ji Hong Kim, Soon-Chun Jeong
Bio-Evaluation Center, Korea Research Institute of Bioscience and Biotechnology

Yun-Gyeong Lee, Won-Hyoung Chung, Jiwon Kim, Yun-Gyeong Lee, Sung-Heon Lee, Nanwuk Kim
Korean Bioinformation Center, Korea Research Institute of Bioscience and Biotechnology

Ali Pirani, Shantanu Kaushikkar
Affymetrix Inc.

Sung-Taek Kang
Dankook University

Bo-Kern Ha
Chonnam National University

Ki Hynn Kim, Jung-Kyung Moon
National Institute of Crop Science, Rural Development Administration

Thank you for your attention !!!!