BMC Genomic Data

DATA NOTE

Open Access

Comprehensive analysis of the genetic variation dataset among wild soybean (*Glycine soja*) in Shandong Province, China

Lei-Lei Li[1†], Ru-Mei Tian[1†], Yan-Yan Pu[1], Yun-Zhe Cong[1], Xue Chen[2], Kai-Hua Jia[1*] and Na-Na Li[1*]

Abstract

Objectives Wild soybean (*Glycine soja*), the ancestor of domesticated soybean, retains a higher level of genetic diversity and adaptability to harsh environments, making it highly valuable for breeding. Here, we re-sequenced 69 wild soybean individuals collected by the Shandong Academy of Agricultural Sciences and identified 1,613,162 high-quality SNPs which not only enriches our understanding of the genetic structure of wild soybean, but also provides valuable resources for further genomic research and genetic improvement of soybean.

Data description In this study, we collected 69 wild soybean accessions from Shandong Province, China, and performed re-sequencing on the DNBSEQ platform, followed by SNPs identification. We then integrated ADMIXTURE, neighbor-joining tree, and principal component analysis to illustrate population characteristics. The results showed that these wild soybean accessions could be divided into three distinct subpopulations, exhibiting significant genetic differences.

Keywords Wild soybean, Population genetics, Structure, Re-sequencing

Objective

Cultivated soybean (*Glycine max*) was domesticated from wild ancestor (*G. soja*) in China approximately 6000 to 9000 years ago. Currently, soybean breeding research is significantly restricted by the narrow genetic variation present in *G. max* [1]. As the wild ancestor of cultivated soybean, wild soybean possesses higher genetic diversity and extensive adaptability [2]. Additionally, there is no reproductive isolation between wild and cultivated soybeans, and the genetic exchange can significantly promote the process of soybean genetic improvement, making wild soybean a valuable gene pool for cultivated soybean [3].

In this study, we report the population characteristics of 69 wild soybean accessions collected from Shandong Province, China, which were re-sequenced on the DNBSEQ platform with an average depth of nearly 26×. Then we identified a total of high-quality 1,613,162 SNPs, ultimately obtained 714,52 SNPs after removing those in linkage. We performed ADMIXTURE, neighbor-joining tree and principal component analysis (PCA) to explore the population structure. The results showed that the 69 individuals were divided into three subgroups.

†Lei-Lei Li and Ru-Mei Tian contributed equally to this work.

*Correspondence:
Kai-Hua Jia
kaihuajia_saas@163.com
Na-Na Li
qyz1127@163.com
[1] National Saline-Alkali Tolerant Crop Germplasm Resources Nursery (Dongying); Shandong International Joint Laboratory of Agricultural Germplasm Resources Innovation; Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences, Jinan 250100, China
[2] Weifang Academy of Agricultural Sciences, Weifang 261017, China

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data file 1	Population structure of 69 wild soybean accessions	Portable document format (.pdf)	Figshare (https://doi.org/10.6084/m9.figshare.27193413.v2) [12]
Data set 1	Resequencing of 69 Wild Soybeans	Fastq file (fastq.gz)	CNGB (https://doi.org/10.26036/CNP0005998) [13]

Data description

In 2022, we collected 69 wild soybean seeds samples from different ecological regions in Shandong Province, China. In 2023, these seeds were germinated indoors to the seedling stage and then collected leaves from the plants. The tissues were immediately frozen in liquid nitrogen and stored at -80℃.

A plant genomic DNA rapid extraction kit (Beijing Biomed Gene Technology Co. Ltd., Beijing, China) was used to extract genomic DNA according to the manufacturer's instruction. DNA integrity was evaluated using Femto Pulse. Libraries were constructed using the MGIEasy universal DNA library prep kit and paired-end sequencing was performed on the DNBSEQ platform.

Quality control of the raw data was conducted using fastp (v.0.23.2) [4], both ends of low-quality sequences were trimmed, minimum read length set to 36 base pairs. The cleaned data were then aligned to the soybean reference genome [5] using BWA (v.0.7.12) [6]. PCR duplicates were removed using Picard Toolkit (https://broadinstitute.github.io/picard/). Variant calling for each individual was conducted with HaplotypeCaller, single-sample GVCF files were imported into GenomicsDB using GenomicsDBImport, followed by multi-sample joint calling with GenotypeGVCFs, these tools are all part of GATK (v 4.4.0.0) [7]. SNPs with a genotype quality (GQ) below 20 and a genotype depth (DP) below 5 were redefined as missing. After SNP filtering with the following conditions: QC < 20, MQ < 40, MAF < 0.05, missing > 0.2, 1,613,162 SNPs were obtained. Using PLINK (v.1.9.0) [8] to remove linkage (--indep-pairwise 50 10 0.2), a total of 714,52 high-quality SNPs were obtained. All these steps is assisted by vcftools (v.0.1.15) [9].These high-quality SNPs were used to analyze the population structure by ADMIXTURE v.1.30 [10], determining the optimal number of ancestral populations (k) from 1 to 9. The most likely subpopulation classification for all wild soybeans was $k = 3$, all 69 cultivars was divided into three subgroups, named G1 to G3. The three subpopulations, G1 to G3, include 34, 23, and 12 individuals, respectively. Among them, G1 is the largest subpopulation. We also constructed a neighbor-joining tree of wild soybeans and visualized it using the ggtree package [11] in R (v.4.4.1), which revealed three major clusters corresponded to the results of the population analysis. This result mutually validated the accuracy of the population structure. Then perform PCA, which similarly grouped these wild soybeans into three subpopulations (Table 1).

Limitations

We have only collected and re-sequenced wild soybeans from the Shandong region, which may not represent the population structure of wild soybeans under different environmental and climatic conditions.

Acknowledgements

We thank the grassroots agricultural institutions in Shandong province for granting permission to collect resources essential to this project.

Authors' contributions

NNL and KHJ conceived the project. LLL, RMT, YYP, XC, and KHJ contributed tissue sampling, LLL and KHJ contributed to the data analysis and LLL wrote the original draft. KHJ reviewed and edited the manuscript. All authors read and approved the final.

Funding

This research was supported by the Key R&D Program of Shandong Province (2024TZXD052,2021LZGC025, 2022LZGC022, and 2023LZGC001).

Data availability

The Data file 1 in this Data Note can be freely and openly accessed on FigShare (https://figshare.com/). Sequence data that support the findings of this study has deposited into CNGB Sequence Archive (CNSA) of China National GeneBank DataBase (CNGBdb) with accession number CNP0005998.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 28 August 2024 Accepted: 29 October 2024
Published online: 11 November 2024

References

1. Nawaz MA, Lin X, Chan TF, Lam HM, Baloch FS, Ali MA, Golokhvast KS, Yang SH, Chung G. Genetic architecture of wild soybean (*Glycine soja* Sieb. And Zucc.) Populations originating from different east Asian regions. Genet Resour Crop Evol. 2021;68:1577–88.
2. Guo J, Liu Y, Wang Y, Chen J, Li Y, Huang H, Qiu L, Wang Y. Population structure of the wild soybean (*Glycine soja*) in China: implications from microsatellite analyses. Ann Botany. 2012;110:777–85.
3. Tirnaz S, Zandberg J, Thomas WJ, Marsh J, Edwards D, Batley J. Application of crop wild relatives in modern breeding: an overview of resources, experimental and computational methodologies. Front Plant Sci. 2022;13:1008904.
4. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:884–90.
5. Jia KH, Zhang X, Li LL, Shi TL, Liu D, Yang YY, Cong YZ, Li RF, Pu YY, Gong YC, Chen X, Si YJ, Tian RM, Qian ZY, Ding HF, Li NN. Telomere-to-telomere genome assemblies of cultivated and wild soybean provide insights into evolution and domestication under structural variation. Plant Commun. 2024;5:100919.
6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
7. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*. 2013;43:11.10. 11-11.10. 33.
8. Slifer SH. PLINK: key functions for data analysis. Curr Protocols Hum Genet. 2018;97:e59.
9. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
10. Liu CC, Shringarpure S, Lange K, Novembre J. Exploring population structure with admixture models and principal component analysis. Methods Mol Biol. 2020;2090:67–86.
11. Yu G. Using ggtree to visualize data on tree-like structures. Curr Protocols Bioinf. 2020;69:e96.
12. Li LL. Population structure of 69 wild soybean accessions. Figshare. Figure 2024. https://doi.org/10.6084/m9.figshare.27193413.v2.
13. Li LL. Resequencing of 69 Wild Soybeans. CNGBdb. Dataset. 2024. https://doi.org/10.26036/CNP0005998.

Publisher's note