Balanced gene retention, transcription and methylation levels between subgenomes produced by soybean-specific paleo-tetraploidization

Chendan Wei  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Jun Qin  
Cereal & Oil Crop Institute, Hebei Academy of Agricultural and Forestry Science

Zhenyi Wang  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Li Wang  
School of Life Science and Center for Genomics and Computational Biology, North China University of Science and Technology

Weina Ge  
School of Life Science and Center for Genomics and Computational Biology, North China University of Science and Technology

Tianyu Lei  
Center for Genomics and Computational Biology, School of Life Science, North China University of Science and Technology

Xinyu Li  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Pengchuan Sun  
School of Life Sciences, North China University of Science and Technology

Jiaqing Yuan  
School of Life Science and Center for Genomics and Computational Biology, North China University of Science and Technology

Fanbo Meng  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Zhikang Zhang  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology
Chao Liu  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

He Guo  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Sangrong Sun  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Ling Jin  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Ying Liu  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Yue Hou  
School of Life Science and Center for Genomics and Computational Biology, North China University of Science and Technology

Jigao Yu  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Xiuqing Li  
Fredericton Research and Development Centre, Agriculture and Agri-Food Canada

Jinpeng Wang  
School of Life Sciences and Center for Genomics and Computational Biology, North China University of Science and Technology

Mengchen Zhang  
Cereal & Oil Crop Institute, Hebei Academy of Agricultural and Forestry Science

Xiyin Wang  
North China University of Science and Technology  https://orcid.org/0000-0003-3454-0374

---

Research Article

**Keywords:** Balanced gene retention, conserved transcription, homoeologous DNA methylation, autotetraploid, soybean, genomics

**Posted Date:** June 21st, 2019

**DOI:** https://doi.org/10.21203/rs.2.10491/v1
Abstract

Background

Polyploidization creates duplicated sets of genomes in a plant. Dominance by a subgenome is often observed in a polyploid species. Here, we studied the duplicated genomic regions in soybean, which was affected by a tetraploidization 13 million years ago.

Results

By referring to eudicot relatives, we detailed that duplicated genes are retained in a balanced manner among homoeologs. Moreover, we found that duplicated genes in colinearity in extant genome are expressed at balance. Notably, the DNA methylation patterns of duplicated genes shared similar methylation levels among three different soybean lines (W931A, WR016, and ZAYOU1), suggesting that either the epigenetic information could be maintained transgenerationally for millions of years or homoeologs in this paleo-autopolyploid species have similar methylation patterns under similar environment.

Conclusions

These findings support the paleo-autotetraploidy of soybean. We proposed that, if taking maize as a model plant for studying paleo-allotetraploids, soybean is a model plant for studying paleo-autotetraploids.

Background

Soybean, *Glycine max*, is an important agricultural crop, providing vegetative oil, proteins, and fodder. The whole-genome sequence of *G. max* was deciphered producing a ~950 Mb genome sequence representing 85% of the predicted 1,115 Mb genome [1]. It is also a model plant to study the evolution of plant genomes and gene families. Its genome is considerably complex, derived from three rounds of deciphered polyploidization, sharing a hexaploid ancestor ~115-130 million years ago (mya) shared with core eudicots, a tetraploid ancestor ~59 mya common to the legume family, and a soybean specific tetraploid (SST) ancestor ~13 mya. The legume-common tetraploidization may have contributed to the establishment of the legume family, the third largest in flowering plants. Recently, by referring to the barrel medic (*Medicago truncatula*) genome and checking gene retention in the corresponding soybean subgenome regions produced by the SST, it was proposed that the SST seems to have an autotetraploid nature [2]. A higher number of chromosomal rearrangements and higher frequency of retention of duplicated genes in soybean than in maize [3]. A similar proposition was raised by checking genomic fractionation statistically [4].

The development of RNA-Seq technologies has greatly improved the sensitivity of transcriptomics research, contributing to understand the function of genes and pathways, the biotic or abiotic resistance,
flower heterosis, plant development, epigenetics, etc. [5-9]. So far, the technology has been successfully applied in different plants, especially the model plants, such as Arabidopsis and rice [10-17]. With soybean, RNA-Seq datasets were produced to study plant response to seed and nodule development, dehydration stress, salinity stress, drought, flooding, or ozone damage, nematode resistance, phytohormone biosynthesis, signal transduction pathways, and the novel and alternatively spliced transcripts [18-28]. A report proposed that one subgenome has subject to lower levels of gene expression in maize, whereas none of these features was observed in soybean [29].

A relatively high degree of nuclear DNA methylation is a specific feature of plant genomes. Targets for cytosine DNA methylation in plant genomes are CG, CHG and CHH (H is A, T, or C) sequences. Gene Body CG and CHG methylation and suppression of centromeric CHH methylation are mediated by DECREASE IN DNA METHYLATION1 in promoters affecting the splicing of genes in rice [30].

In contrast to the autotetraploid nature of SST in soybean, maize was proposed to have a paleo-allotetraploid nature. The maize duplicated regions display significant differences in gene loss and gene expression, transposable element accumulation, small interfering RNAs, and DNA methylation around genes [3, 31]. Recently, a study shows that he DNA methylation levels of the two pear subgenomes show little deviation, and there is no expression advantage between the two subgenomes [32]. Here, we would use new data to check the retention, transcription, and methylation difference between the SST duplicated genes, shedding light on the evolutionary divergence and genomic stability of them after evolution during millions of years.

**Methods**

**Materials**

Soybean and *Medicago truncatula* genomes and their gene annotations were downloaded from JGI, version 2.0, and [http://www.jcvi.org/medicago/](http://www.jcvi.org/medicago/), version 4.0, respectively [1, 33]. Transcriptome data of 11 soybean tissues generated by Wang et al. was obtained from the NCBI Short Read Archive ([http://www.ncbi.nlm.nih.gov/sra/?term=SRP040057](http://www.ncbi.nlm.nih.gov/sra/?term=SRP040057)) [34].

**Calculation of Gene Expression Levels**

For all expression sets, reads were aligned to the soybean genome using an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences (Bowtie2) [35]. The align results were sorted and transfer ‘bam’ to ‘sam’ format by SAMTOOLS [36]. Following the alignment, estimates the expression levels for each soybean gene model using Cufflinks [37]. Using the published annotations of the Williams 82 v2.0 working gene list.

**Methylation characterization**
To characterize DNA methylation, three soybean samples (ZAYOU1, WR016, and W931A) were sequenced on an Illumina Hiseq 2000/2500 platform, and 100 bp/50 bp single-end reads were generated. Image analysis and base calling were performed with the standard Illumina pipeline, and finally 100 bp paired-end reads were generated.

Bismark software (version 0.12.5) [38] was used to perform alignments of bisulfite-treated reads to the soybean (Williams82) genome with the default parameters. The reference genome was first transformed into a bisulfite-converted version (C-to-T and G-to-A converted) and then indexed using bowtie2 [35]. Sequence reads were also transformed into fully bisulfite-converted versions (C-to-T and G-to-A converted) before alignment to similarly converted versions of the genome in a directional manner. Then sequence reads that produced a unique best alignment from the two alignment processes (original top and bottom strand) were compared with the normal genomic sequence, and the methylation state of all cytosine positions in the read was inferred. Reads that aligned to the same regions of the genome were regarded as duplicated reads.

To identify the methylation site, we modeled the sum $s_{ij}$ of methylated counts as a binomial (Bin) random variable with methylation rate $r_{ij}$,

\[ \text{Formula 1} \]

Due to technical limitations the formula could not be printed here. You can find it in the manuscript file or in a separate supplemental file.

We employed a sliding-window approach, which is conceptually similar to the approaches that have been used for bulk bisulfite sequencing (http://www.bioconductor.org/packages/2.13/bioc/html/bsseq.html). With window size $w = 3,000$ bp and step size $600$ bp [39], the sum of methylated and unmethylated read counts in each window was calculated. Methylation level (ML) for each C site shows the fraction of methylated Cs, and is defined as:

\[ \text{Formula 2} \]

Due to technical limitations the formula could not be printed here. You can find it in the manuscript file or in a separate supplemental file.

The ML calculated was further corrected with the bisulfite non-conversion rate according to previous studies [40]. Given the bisulfite non-conversion rate $r$, the corrected ML was estimated as:

\[ \text{Formula 3} \]

Due to technical limitations the formula could not be printed here. You can find it in the manuscript file or in a separate supplemental file.

All P-values were calculated using cumulative binomial distributions assuming an equal chance of gene copies on two sets of homoeologous regions corresponding to a specified referring chromosome.

**Results**

**Homoeologous gene colinearity**
By using our established software ColinearScan [41], we revealed 25,302 pairs of gene blocks, each with 5 or more colinear genes, within soybean; 4,391 pairs in barrel medic; and 50,672 colinear gene pairs between soybean and barrel medic. Owing to the existence of more ancient polyploidizations, the homologous pairs in soybean may not all be produced by the SST. By checking homologous correspondence between barrel medic and soybean (1:2), synonymous substitution rates or Ks values of colinear gene pairs, chromosomal rearrangement breakpoints, and colinear block complementarity relative to barrel medic chromosomes (Additional file 1: Figure S1), we tried to locate SST-derived homologous blocks. Even for genes produced by the same event, the variation of Ks values among duplicated genes could be sufficiently large to interfere the inferring whether genes were derived from that specific event. We used the median Ks to help distinguish the blocks from different events. Further, blocks showing complementary mapping onto barrel medic chromosomes were inferred to be produced from the same ancestral chromosome. For each barrel medic chromosome, two sets of SST-derived homoeologous soybean regions were located (Additional file 2: Figure S2). A total of 330 SST-derived blocks were inferred, including 25,229 pairs of genes with average Ks value ~0.52 – this corresponds closely to the soybean-barrel medic split, suggesting that the SST might have led to the speciation of soybean from other legumes.

Similar homoeologous gene retention

Corresponding pairs of homoeologous soybean regions showed similar gene loss rates and therefore could not be divided into subgenomes. Among 59,578 barrel medic genes, 8,465 (14.1%) had two soybean (likely) orthologs, and 8,382 (14.2%) had only one. The remainder had no correspondence, possibly due to gene loss or gene translocation in soybean or to incongruity in gene prediction, which is not rare at present [42]. Regardless of the cause, such incongruity does not substantially affect our interpretation of soybean genome formation below.

An interesting observation is that SST homoeologous regions have very similar gene retention patterns (Fig. 1). With barrel medic chromosomes as reference, we calculated gene retention rates in sliding windows of 100 genes and steps of 1 gene. Here, for homoeologous regions mapped onto a barrel medic chromosome, there were prominent variations in gene retention. In some regions, only a tiny fraction of genes was preserved, while in other regions, gene preservation reached 40%. However, while the retention curves along the barrel medic chromosomes were volatile, we observed similar gene retention or loss between homoeologous/duplicated regions (Fig. 1): a total of 57.5% of homoeologous regions had less than 5% difference in gene retention rates (Additional file 3: Table S1). This precludes their division into two subgenomes, and is in sharp contrast to the two easily distinguished maize subgenomes, a dominant one and a sensitive one [31]. Homoeologous regions with different gene retention levels often involve large chromosome patches that are missing, possibly due to occasional DNA removal (Fig. 1).

Similar homoeologous gene expression
Corresponding pairs of homoeologous soybean regions also showed similar gene expression (Additional file 4: Table S2). There were 8,205 SST-derived homoeologous gene pairs expressed, and 56.7% (4,653 pairs) had similar expression levels (Fig. 2a; Additional file 5: Figure S3a-l). Along the barrel medic reference chromosomes, as with gene retention, there were prominent differences between certain peaks and troughs, but expression levels of preserved homoeologous genes showed similar patterns (Fig. 2b). Similar to the analysis of gene retention rate, we divided the homoeologous chromosomes into sliding windows and checked for differences in gene expression (see Methods for details). We found that the expression activity levels between homoeologous genes in nearly 80% of homoeologous regions for all chromosomes were not significantly different (p < 5%).

**Similar homoeologous DNA methylation**

Corresponding homoeolog pairs of homoeologous soybean regions were also similarly methylated. We characterized DNA methylation of three different soybean lines (W931A, WR016, and ZAYOU1). By referring to the barrel medic chromosomes, DNA methylation levels were statistically indistinguishable between most corresponding homoeologous regions (Fig. 3a; Additional file 6: Figure S4a-b; Additional file 7: Figure S5a-b). More than 45.0% (3,354 pairs) of homoeologous genes had methylation levels that were less than twofold different (Additional file 8: Table S3). Along the barrel medic reference chromosomes, the DNA methylation levels of homoeologous regions showed similar zigzagging patterns (Fig. 3b; Additional file 6: Figure S4c-d). Just like gene retention and expression, we divided homoeologous chromosomes into sliding windows (see Methods), finding for all chromosomes that nearly 43–44% of regions had methylation levels that were indistinguishable at a significance level < 0.05 (Additional file 9: Table S4a); and 69%–71% regions were indistinguishable at a significance level $P < 0.1$ (Additional file 9: Table S4b).

**Discussion**

Polyploidization would produce duplicated copies of genomes, or subgenomes in a plant. Often non-equivalence is observed between subgenomes. Maize was an ancient allotetraploid, and it was found that the duplicated regions in maize genome are significantly unbalanced in terms of the preservation of duplicated genes [31]. Actually, maize duplicated regions that are orthologous to each sorghum chromosome showed distinctive gene retention differences, suggesting existence of a dominant subgenome and a sensitive one. Besides, genes constituting the dominant subgenome show a higher expression level. This makes a good proof for its paleo-allotetraploidy origin.

Here, we found that corresponding pairs of homoeologous soybean regions are similar in gene retention, expression, and methylation. This is distinctively different from the observation in maize. If the maize pattern is typical of an ancient allotetraploid, the situation in soybean shows an ancient tetraploid with two largely similar subgenomes, and therefore, can only explain by autotetraploidy. This supports the previous classification of soybean as having a class I polyploid ancestor, or autotetraploid [4]. Taking
maize as a model plant to study ancient allotetraploidy, soybean can be a model plant to explore the effect of ancient autoploidy.

As a typical epigenetic feature, DNA methylation was well documented in recent years at a whole-genome level [43, 44]. As compared to genetic patterns, one may expect that the patterns of DNA methylation might be more easily lost over time. Although there has been recent evidence showing transgenerational inheritance of epigenetic changes imprinted by environmental factors [45], none has ever expected how long epigenetic changes to pass through time. In the present study, we found a high similarity of homoeologs in DNA methylation among the three soybean genotypes. Two hypotheses can be proposed to interpret the similarity of DNA methylation among homoeologs (duplicated gene copies): The first hypothesis is that the DNA methylations transgenerationally passed since the paleo-autotetraploidization ~13 million years ago, which means that epigenetic features could be kept for millions of years, even eternally. The epigenetic changes may be imprinted into genetic materials to translate into genetic information. If this is the case, the finding may obscure the boundary, to certain extent, between epigenetic and genetic information. The second hypothesis is that the similarity among homoeologs is due to sequence-based determination of DNA patterns. The homoeologs in subgenomes of this ancient autotetraploid (soybean) are highly similar in sequence and gene order conservation and therefore could also have similar DNA methylation patterns since these three soybean genotypes are all from soybean breeding programs and their plants were all grown under similar conditions, a very different climate condition from 13 millions year ago. The high similarity of DNA methylation among homoeologs in soybean is a very interesting discovery, regardless whether the DNA methylation was preserved transgenerationally for 13 million years or occurred similarity due to the sequence and gene order similarity.

Conclusion

Polyploidization creates duplicated sets of genomes in a plant. Dominance by a subgenome is often observed in a polyploid species. Here, we studied the duplicated genomic regions generated by tetraploidization 13 million years ago in soybean. By referring to barrel medic species, we detailed that duplicated genes are retained in balance among the same duplicated region. Duplicated genes in colinearity in extant genome are also expressed in balance. Notably, the DNA methylation levels of duplicated genes are similar in three different soybean lines (W931A, WR016, and ZAYOU1). Putting together, these findings support the ancient autotetraploid nature of soybean. We proposed that, if taking maize as a model plant for studying ancient allotetraploids, soybean can be a model plant for studying ancient autotetraploids.

Declarations

Acknowledgements
We thank the helpful discussion with researchers at iGeno Co. Ltd, China.

**Funding**

We appreciate financial support from the Ministry of Science and Technology of the People's Republic of China (2016YFD0101001), China National Science Foundation (31510333 to JW and 31371282 to XW), Natural Science Foundation of Hebei Province (C2015209069 to JW and C2014209201 to LW), and Tangshan Key Laboratory Project to XW.

**Authors’ contributions**

Contributed equally: CW, JQ, ZW, LW. Conceived and led the research: XW. Collected and analyzed the data: CW, JQ, ZW, WG, XL, YL, PS, JY, FM, ZZ, CL, HG, SS, LJ, YL, YH, JY. Contributed analyzing tools: TL, PS. Participated in preparing and writing the manuscript: XW, CW, JW, JQ, ZW, LW, MZ. Performed the analysis with constructive discussion: LW, WG, ZW, CW, JQ, XW.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare no competing financial interests.

**Publisher’s Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Author details**

1 School of Life Sciences, North China University of Science and Technology, Tangshan, Hebei 063210, China. 2 Center for Genomics and Computational Biology, North China University of Science and
Technology, Tangshan, Hebei 063210, China. 3 Cereal & Oil Crop Institute, Hebei Academy of Agricultural and Forestry Science, No. 162, Hengshan Street, Shijiazhuang, 050035, China. 4 State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China. 5 University of Chinese Academy of Sciences, Beijing 100049, China. 6 Fredericton Research and Development Centre, Agriculture and Agri-Food Canada, Fredericton, New Brunswick, E3B 4Z7, Canada.

References

1. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J et al: Genome sequence of the palaeopolyploid soybean. Nature 2010, 463(7278):178-183.

2. Wang J, Sun P, Li Y, Liu Y, Yu J, Ma X, Sun S, Yang N, Xia R, Lei T et al: Hierarchically Aligning 10 Legume Genomes Establishes a Family-Level Genomics Platform. Plant Physiol 2017, 174(1):284-300.

3. Zhao M, Zhang B, Lisch D, Ma J: Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. Plant Cell 2017, 29(12):2974-2994.

4. Garsmeur O, Schnable JC, Almeida A, Jourda C, D'Hont A, Freeling M: Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol 2014, 31(2):448-454.

5. Brito LF, Irla M, Kalinowski J, Wendisch VF: Detailed transcriptome analysis of the plant growth promoting Paenibacillus riograndensis SBR5 by using RNA-seq technology. BMC genomics 2017, 18(1):846.

6. Zeng F, Biligetu B, Coulman B, Schellenberg MP, Fu YB: RNA-Seq Analysis of Plant Maturity in Crested Wheatgrass (Agropyron cristatum L.). Genes 2017, 8(11).

7. Li Y, Huang J, Song X, Zhang Z, Jiang Y, Zhu Y, Zhao H, Ni D: An RNA-Seq transcriptome analysis revealing novel insights into aluminum tolerance and accumulation in tea plant. Planta 2017, 246(1):91-103.

8. Van Campenhout J, Vanreusel A, Van Belleghem S, Derycke S: Transcription, Signaling Receptor Activity, Oxidative Phosphorylation, and Fatty Acid Metabolism Mediate the Presence of Closely Related Species in Distinct Intertidal and Cold-Seep Habitats. Genome biology and evolution 2016, 8(1):51-69.

9. Gupta S, Garg V, Kant C, Bhatia S: Genome-wide survey and expression analysis of F-box genes in chickpea. BMC genomics 2015, 16:67.

10. Wang Z, Tang K, Zhang D, Wan Y, Wen Y, Lu Q, Wang L: High-throughput m6A-seq reveals RNA m6A methylation patterns in the chloroplast and mitochondria transcriptomes of Arabidopsis thaliana. PloS one 2017, 12(11):e0185612.
11. Basbouss-Serhal I, Soubigou-Taconnat L, Bailly C, Leymarie J: Germination Potential of Dormant and Nondormant Arabidopsis Seeds Is Driven by Distinct Recruitment of Messenger RNAs to Polysomes. *Plant physiology* 2015, 168(3):1049-1065.

12. Gunther T, Lampei C, Schmid KJ: Mutational Bias and Gene Conversion Affect the Intraspecific Nitrogen Stoichiometry of the Arabidopsis thaliana Transcriptome. *Molecular biology and evolution* 2013, 30(3):561-568.

13. Cohen SP, Liu H, Argueso CT, Pereira A, Vera Cruz C, Verdier V, Leach JE: RNA-Seq analysis reveals insight into enhanced rice Xa7-mediated bacterial blight resistance at high temperature. *PloS one* 2017, 12(11):e0187625.

14. Oono Y, Yazawa T, Kanamori H, Sasaki H, Mori S, Matsumoto T: Genome-wide analysis of rice cis-natural antisense transcription under cadmium exposure using strand-specific RNA-Seq. *BMC genomics* 2017, 18(1):761.

15. Renuka P, Madhav MS, Padmakumari AP, Barbadikar KM, Mangrauthia SK, Vijaya Sudhakara Rao K, Marla SS, Ravindra Babu V: RNA-Seq of Rice Yellow Stem Borer Scirpophaga incertulas Reveals Molecular Insights During Four Larval Developmental Stages. *G3 (Bethesda)* 2017, 7(9):3031-3045.

16. Hsu SK, Tung CW: RNA-Seq Analysis of Diverse Rice Genotypes to Identify the Genes Controlling Coleoptile Growth during Submerged Germination. *Frontiers in plant science* 2017, 8:762.

17. Yoo YH, Nalini Chandran AK, Park JC, Gho YS, Lee SW, An G, Jung KH: OsPhyB-Mediating Novel Regulatory Pathway for Drought Tolerance in Rice Root Identified by a Global RNA-Seq Transcriptome Analysis of Rice Genes in Response to Water Deficiencies. *Frontiers in plant science* 2017, 8:580.

18. Zhang H, Kjemtrup-Lovelace S, Li C, Luo Y, Chen LP, Song BH: Comparative RNA-Seq Analysis Uncovers a Complex Regulatory Network for Soybean Cyst Nematode Resistance in Wild Soybean (Glycine soja). *Scientific reports* 2017, 7(1):9699.

19. Zhang C, Lin C, Fu F, Zhong X, Peng B, Yan H, Zhang J, Zhang W, Wang P, Ding X *et al.*: Comparative transcriptome analysis of flower heterosis in two soybean F1 hybrids by RNA-Seq. *PloS one* 2017, 12(7):e0181061.

20. Waldeck N, Burkey K, Carter T, Dickey D, Song Q, Taliercio E: RNA-Seq study reveals genetic responses of diverse wild soybean accessions to increased ozone levels. *BMC genomics* 2017, 18(1):498.

21. Yuan SL, Li R, Chen HF, Zhang CJ, Chen LM, Hao QN, Chen SL, Shan ZH, Yang ZL, Zhang XJ *et al.*: RNA-Seq analysis of nodule development at five different developmental stages of soybean (Glycine max) inoculated with Bradyrhizobium japonicum strain 113-2. *Scientific reports* 2017, 7:42248.

22. Chen W, Yao Q, Patil GB, Agarwal G, Deshmukh RK, Lin L, Wang B, Wang Y, Prince SJ, Song L *et al.*: Identification and Comparative Analysis of Differential Gene Expression in Soybean Leaf Tissue under
Drought and Flooding Stress Revealed by RNA-Seq. *Frontiers in plant science* 2016, 7:1044.

23. Yuan S, Li R, Chen S, Chen H, Zhang C, Chen L, Hao Q, Shan Z, Yang Z, Qiu D et al: RNA-Seq Analysis of Differential Gene Expression Responding to Different Rhizobium Strains in Soybean (Glycine max) Roots. *Frontiers in plant science* 2016, 7:721.

24. Lambirth KC, Whaley AM, Blakley IC, Schlueter JA, Bost KL, Loraine AE, Piller KJ: A Comparison of transgenic and wild type soybean seeds: analysis of transcriptome profiles using RNA-Seq. *BMC biotechnology* 2015, 15:89.

25. Wang HK, Ng YK, Koh E, Yao L, Chien AS, Lin HX, Lee YK: RNA-Seq reveals transcriptomic interactions of Bacillus subtilis natto and Bifidobacterium animalis subsp. lactis in whole soybean solid-state co-fermentation. *Food microbiology* 2015, 51:25-32.

26. Whaley A, Sheridan J, Safari S, Burton A, Burkey K, Schlueter J: RNA-seq analysis reveals genetic response and tolerance mechanisms to ozone exposure in soybean. *BMC genomics* 2015, 16:426.

27. Belamkar V, Weeks NT, Bharti AK, Farmer AD, Graham MA, Cannon SB: Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor family in soybean (Glycine max) during dehydration and salt stress. *BMC genomics* 2014, 15:950.

28. Shi G, Huang F, Gong Y, Xu G, Yu J, Hu Z, Cai Q, Yu D: RNA-Seq analysis reveals that multiple phytohormone biosynthesis and signal transduction pathways are reprogrammed in curled-cotyledons mutant of soybean [Glycine max (L.) Merr]. *BMC genomics* 2014, 15:510.

29. Zhao M, Zhang B, Lisch D, Ma J: Patterns and Consequences of Subgenome Differentiation Provide Insights into the Nature of Paleopolyploidy in Plants. *The Plant Cell* 2017, 29(12):2974-2994.

30. Wang X, Hu L, Wang X, Li N, Xu C, Gong L, Liu B: DNA Methylation Affects Gene Alternative Splicing in Plants: An Example from Rice. *Molecular plant* 2016, 9(2):305-307.

31. Schnable JC, Springer NM, Freeling M: Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A* 2011, 108(10):4069-4074.

32. Li Q, Qiao X, Yin H, Zhou Y, Dong H, Qi K, Li L, Zhang S: Unbiased subgenome evolution following a recent whole-genome duplication in pear (Pyrus bretschneideri Rehd.). *Hortic Res* 2019, 6:34.

33. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C et al: A reference genome for common bean and genome-wide analysis of dual domestications. *Nature genetics* 2014, 46(7):707-713.

34. Wang L, Cao C, Ma Q, Zeng Q, Wang H, Cheng Z, Zhu G, Qi J, Ma H, Nian H et al: RNA-seq analyses of multiple meristems of soybean: novel and alternative transcripts, evolutionary and functional implications. *BMC Plant Biol* 2014, 14:169.
35. Langmead B, Salzberg SL: Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012, 9(4):357-359.

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009, 25(16):2078-2079.

37. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 2010, 28(5):511-515.

38. Krueger F, Andrews SR: Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011, 27(11):1571-1572.

39. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G: Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* 2014, 11(8):817-820.

40. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD et al: Global epigenomic reconfiguration during mammalian brain development. *Science* 2013, 341(6146):1237905.

41. Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J: Statistical inference of chromosomal homology based on gene colinearity and applications to Arabidopsis and rice. *BMC bioinformatics* 2006, 7:447.

42. Wang X, Wang J, Jin D, Guo H, Lee TH, Liu T, Paterson AH: Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events. *Mol Plant* 2015, 8(6):885-898.

43. Yaish MW, Al-Lawati A, Al-Harrasi I, Patankar HV: Genome-wide DNA Methylation analysis in response to salinity in the model plant caliph medic (Medicago truncatula). *BMC genomics* 2018, 19(1):78.

44. Bouyer D, Kramdi A, Kassam M, Heese M, Schnittger A, Roudier F, Colot V: DNA methylation dynamics during early plant life. *Genome biology* 2017, 18(1):179.

45. Matthews SG, Phillips DI: Transgenerational inheritance of stress pathology. *Exp Neurol* 2012, 233(1):95-101.

**Additional File Legend**

Additional file 1: Figure S1. Homologous dotplot between Medicago (Medicago truncatula) and soybean (Glycine max) genomes. The best, secondarily best, and other matched homologous gene pairs output by BLAST were dotplotted using red, blue, and gray colors, respectively, in this figure. Colinear gene blocks
were inferred and circled. We used median synonymous substitution rates or Ks to help distinguish the blocks from different events. Based on chromosomal breakpoints and positional closeness, a set of soybean chromosomal regions making up one copy of a Medicago chromosome are circled in red, and another set making up another copy of the same Medicago chromosome are circled in blue. (PDF 402 kb)

Additional file 2: Figure S2. Soybean genomic regions mapped onto Medicago chromosomes. Genomic paralogy, orthology, and outparalogy information within and among soybean (G) and Medicago (M) genomes are displayed in six circles. The curved lines within the inner circle link duplicated genes produced by legume-common tetraploidy. The short lines forming the innermost Medicago chromosome circle (marked M) represent genes, which have one set of paralogous regions, forming another circle (the fourth one, also marked M). Each of the two sets of Medicago paralogous chromosomal regions has two orthologous copies in soybean. The resulting six circles are marked according to species by a capital letter (M or S). Each circle is formed by short vertical lines that denote homologous genes, colored according to chromosome number in their respective source plant as shown in the inset color scheme. (PDF 572 kb)

Additional file 3: Table S1. Gene retention and expression in homoeologous soybean regions. (DOCX 18 kb)

Additional file 4: Table S2. Gene expression between soybean homoeologous chromosome regions. (DOCX 17 kb)

Additional file 5: Figure S3. Soybean homoeologous gene expression along corresponding orthologous Medicago chromosomes. (a) Histogram of gene expression in 11 soybean sample; (b) root tip; (c) cotyledon; (d) hypocotyl; (e) callus; (f) SAM6D; (g) SAM17D; (h) SAM38D; (i) AM; (j) IBM; (k) IAM; (l) OF. Genes highly expressed in soybean homoeologous region group 1 (red), highly expressed in soybean homoeologous region group 2 (blue), without much difference in expression between regions (purple), and showing significant differences in expression between regions (gray) are counted, respectively. (PDF 555 kb)

Additional file 6: Figure S4. Methylation levels in soybean homoeologous regions from two additional soybean lines. (a) Histograms for WR016, genes highly methylated in soybean homoeologous region group 1 (red), highly methylated in soybean homoeologous region group 2 (blue), without much difference in methylation between regions (purple), and showing significant differences in methylation between regions (gray) are counted, respectively. (b) ZAYOU1, genes highly methylated in soybean homoeologous region group 1 (red), highly methylated in soybean homoeologous region group 2 (blue), without much difference in methylation between regions (purple), and showing significant differences in methylation between regions (gray) are counted, respectively. (c) curves along referring chromosomes for WR016, methylation levels in sliding windows of soybean homoeologous region group 1 (red) and homoeologous region group 2 (black), and the difference between the two groups (blue) are displayed. (d) ZAYOU1, methylation levels in sliding windows of soybean homoeologous region group 1 (red) and homoeologous region group 2 (black), and the difference between the two groups (blue) are displayed. (PDF 1303 kb)
Additional file 7: Figure S5. Integrated methylation levels in soybean homoeologous regions from three soybean lines. (PDF 269 kb)

Additional file 8: Table S3. Gene methylation between soybean homoeologous chromosome regions. (DOCX 15 kb)

Additional file 9: Table S4. DNA methylation in soybean homoeologous chromosome regions. (a) Histogram along corresponding orthologous Medicago chromosomes averaged for three soybean lines; (b) histogram for each individual soybean line. (DOCX 24 kb)

Figures
Soybean gene retention in duplicated regions. Gene retention level along duplicated regions in soybean is shown as to their corresponding orthologous Medicago truncatula chromosomes. Rates of retained genes in sliding windows of soybean homoeologous region group 1 (red) and homoeologous region group 2 (black), and the difference between the two groups (blue) are displayed. Note: different colors of gene retention are to distinguish the two homoeologous groups for each Medicago truncatula.
chromosome, but do not show that the groups of the same color among chromosomes form a subgenome.

**Figure 2**

Soybean gene expression in duplicated regions. Gene expression level along duplicated regions in soybean is shown as to their corresponding orthologous Medicago truncatula chromosomes. a: Histogram of gene expression. b: Gene expression curves along Medicago truncatula genomes. Gene expression levels in soybean homoelogous region group 1 (red) and homoelogous region group 2 (black), and the difference between group 1 and group 2 (blue) are displayed.

**Figure 3**

Soybean gene methylation in duplicated regions. Gene methylation levels along duplicated regions in soybean are shown, in reference to their corresponding orthologous Medicago truncatula chromosomes. a: Methylation levels in soybean homoeologous regions – histograms for soybean line W931A. b: Methylation levels in soybean homoeologous regions are displayed for soybean W931A. (b and d)
Soybean homoeologous region group 1 (red) and homoeologous region group 2 (black), and the difference between the two groups (blue) are displayed.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2.docx
- TableS3.docx
- TableS4.docx
- TableS1.docx
- fig.S1.pdf
- Formulas.docx
- fig.S3.pdf
- fig.S4.pdf
- fig.S5.pdf
- fig.S2.pdf