J Appl Biol Chem (2020) 63(3), 189–195
https://doi.org/10.3839/jabc.2020.026

Online ISSN 2234-7941
Print ISSN 1976-0442

Article: Microbiology

De novo genome assembly and single nucleotide variations for Soybean yellow common mosaic virus using soybean flower bud transcriptome data

Yeonhwa Jo[1] · Hoseong Choi[1] · Sang-Min Kim[2] · Bong Choon Lee[2] · Won Kyong Cho[1]

Received: 23 April 2020 / Accepted: 21 July 2020 / Published Online: 30 September 2020

Abstract The soybean (*Glycine max* L.), also known as the soya bean, is an economically important legume species. Pathogens are always major threats for soybean cultivation. Several pathogens negatively affect soybean production. The soybean is also known as a susceptible host to many viruses. Recently, we carried out systematic analyses to identify viruses infecting soybeans using soybean transcriptome data. Our screening results showed that only few soybean transcriptomes contained virus-associated sequences. In this study, we further carried out bioinformatics analyses using a soybean flower bud transcriptome for virus identification, genome assembly, and single nucleotide variations (SNVs). We assembled the genome of *Soybean yellow common mosaic virus* (SYCMV) isolate China and revealed two SNVs. Phylogenetic analyses using three viral proteins suggested that SYCMV isolate China is closely related to SYCMV isolates from South Korea. Furthermore, we found that replication and mutation of SYCMV is relatively low, which might be associated with flower bud tissue. The most interesting finding was that SYCMV was not detected in the cytoplasmic male sterility (CMS) line derived from the non-CMS line that was severely infected by SYCMV. In summary, *in silico* analyses identified SYCMV from the soybean flower bud transcriptome, and a nearly complete genome of SYCMV was successfully assembled. Our results suggest that the low level of virus replication and mutation for SYCMV might be associated with plant tissues. Moreover, we provide the first evidence that male sterility might be used to eliminate viruses in crop plants.

Keywords Genome · Flower bud · Single nucleotide variation · Soybean · *Soybean yellow common mosaic virus* · Transcriptome

Won Kyong Cho (✉)
E-mail: wonkyong@gmail.com

[1]Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Republic of Korea

[2]Crop Foundation Division, National Institute of Crop Science, RDA, Wanju 55365, Republic of Korea

Introduction

The soybean (Glycine max), also known as the soya bean, is an economically important legume species belonging to the genus *Glycine* [1]. It is a rich source of protein and oil for humans as well as animals [2,3]. In Asian countries, many fermented foods based on the soybean are widely consumed as health foods [2]. In addition, the soybean has an ability to fix atmospheric nitrogen via symbioses together with soil-borne microorganisms [4]. Furthermore, the complete genome sequence of the soybean enables us to identify several genetic traits that can be further usefully applied in the development of improved soybean cultivars [5].

Pathogens are always major threats for soybean cultivation [6]. Several pathogens—including bacteria, fungi, nematodes, and viruses—negatively affect soybean production. The soybean is also known as a susceptible host to many viruses [7]. Of the known viruses infecting the soybean, only a few viruses seriously hinder soybean production. For example, *Soybean mosaic virus* (SMV) in the family *Potyviridae* is a pathogenic virus causing stunted growth and crinkled leaves [8]. Another virus, *Bean pod mottle virus* (BPMV) in the family *Secoviridae,* causes mottled and crinkled leaves [9]. Co-infection of two viruses that can interact with each other leads to severe disease symptoms in the soybean [10]. SMV is transmitted by aphids, while BPMV is transmitted by the bean leaf beetle, *Cerotoma trifurcata* [11,12]. Furthermore, seed transmission of the two viruses has been reported [8,13].

So far, several techniques have been developed for the detection

and identification of viruses. In general, the identification of viruses is based on the observation of disease symptoms that might be caused by known viruses followed by the identification of viral nucleic acids and proteins [14,15]. Thus, information, such as disease symptoms and diagnostic tools for the target viruses, is required. However, this approach has many limitations for virus identification. For example, non-target viruses as well as novel viruses cannot be identified. To overcome the limits of the traditional methods for virus detection and identification, many research groups have adopted next-generation sequencing (NGS)-based approaches [16]. As a result, it is currently possible to identify several viruses infecting a host simultaneously. Furthermore, not only known viruses but also novel viruses are frequently identified by NGS techniques. In addition, NGS along with several bioinformatics tools facilitates the assembly of viral genomes [17]. Recently, studies found that many plant transcriptomes are infected by viruses, suggesting the utility of plant transcriptomes for virus identification and viral genome assembly [18,19].

Recently, we carried out systematic analyses to identify viruses infecting the soybean using soybean transcriptome data. Our screening results showed that only few soybean transcriptomes contained virus-associated sequences. In this study, we further carried out bioinformatics analyses for virus identification, genome assembly, and single nucleotide variations (SNVs) of an identified viral genome using a soybean transcriptome.

Materials and Methods

Plant materials, library preparation, and NGS

In this study, two different soybean cultivars, NJCMS1A and NJCMS1B, were used for transcriptome analyses. NJCMS1A is a cytoplasmic male-sterile line, while NJCMS1B is a recurrent parent line. Detailed information can be found in the previous study [20]. Soybean plants were grown in the Experimental Station, National Center for Soybean Improvement, Nanjing Agricultural University, Nanjing, Jiangsu Province, China, 2012. Flower bud tissues with different sizes were harvested before the abortion stage. Collected samples were immediately frozen in liquid nitrogen and subjected to total RNA extraction using a TRIzol kit (Invitrogen, Carlsbad, CA, USA). Two different libraries for the two cultivars were prepared using a TruSeq RNA Sample Prep Kit (Illumina, San Diego, CA, USA) following the manufacturer's instructions. The prepared libraries were pair-end sequenced by Illumina's HiSeq 2000 system. The raw data can be accessed through the SRA (Sequence Read Archive) database (http://www.ncbi.nlm.nih.gov/sra/) with the accession numbers SRR1752764 and SRR1752076.

Raw data processing, *de novo* transcriptome assembly, and virus identification

Raw data were downloaded from the SRA database and converted to a FASTQ file using the SRA toolkit [21]. All bioinformatics analyses were conducted in a workstation (four 16-core CPUs and 256 GB ram) with a Linux operating system (Linux Mint version 17). We first screened virus-associated reads in the two transcriptomes by a blast search. To do so, the FASTQ files were converted to FASTA files using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). FASTA files were blasted against the viral reference database (http://www.ncbi.nlm.nih.gov/genome/viruses/) using standalone blast with an E-value of 1e-5 as a cutoff [22]. We found that only NJCMS1B contained several virus-associated reads. We *de novo* assembled the transcriptome of NJCMS1B using Trinity version 2.0.6 with default parameters [23]. Again, the assembled transcriptome was subjected to the MegaBLAST search against the viral database with an E-value of 1e-5 as a cutoff.

De novo assembly of SYCMV genomes and sequence alignment

In order to assemble the *Soybean yellow common mosaic virus* (SYCMV) genome using transcriptome data, nine contigs associated with SYCMV were retrieved from the soybean transcriptome using the BLASTCMD program. We aligned the nine contigs on the SYCMV reference genome (NC_016033.1) using ClustalW implemented in the MEGA6 program [24]. After manual modification, a nearly complete genome of SYCMV isolate China was obtained. To confirm the assembled SYCMV genome sequence, raw data were again mapped on the assembled SYCMV genome by Burrows-Wheeler Aligner (BWA) with default parameters [25].

Identification of SNVs in soybean transcriptome

To examine the SNVs of SYCMV in the soybean flower bud transcriptome, we mapped raw data on the assembled SYCMV genome using the BWA program with default parameters. The obtained SAM file was converted into a BAM file using SAMtools [26]. The BAM file was sorted, and a VCF file was generated using mpileup [27]. SNVs were called using BCFtools implemented in SAMtools. To visualize the positions of identified SNVs, the mapped SAM file was imported into the Tablet program [28].

Generation of phylogenetic trees

To reveal the phylogenetic position of SYCMV isolate China, we generated three different phylogenetic trees. To do so, we used three protein sequences for coat protein, movement protein, and polyprotein P2ab. By BLASTP search, we identified other viral sequences that showed sequence similarity to the corresponding SYCMV proteins. A total of 12 coat proteins, 8 movement proteins, and 9 polyprotein P2ab protein sequences were aligned by the ClustalW program with default parameters. We manually edited aligned sequences by deleting unnecessary sequences. We constructed three different phylogenetic trees using the MEGA6 program. The neighbor-joining method and the Poisson model

with 1,000 bootstrap replicates were used for construction of the phylogenetic trees.

Results

Transcriptome assembly and virus identification

Recently, we screened soybean transcriptome data deposited in NCBI's SRA database in order to identify viruses infecting soybeans. Of the screened soybean transcriptomes, the soybean transcriptome SRR1752764 contained several virus-associated sequences. The soybean transcriptome was derived from flower bud tissues in the previous study [20]. The previous study examined two different soybean bud transcriptomes that were further divided into a cytoplasmic male sterility (CMS) line (SRR1752076) and a normal parent line (SRR1752764). The CMS transcriptome did not possess any virus-associated sequences. We *de novo* assembled the non-CMS soybean transcriptome by the Trinity program. The assembled transcriptome had 116,108 transcripts (contigs) with 710 bp of contig N50 value (Table 1). To identify the viruses infecting soybeans, the assembled contigs were blasted against the viral reference database. We found that nine contigs were associated with SYCMV (Table 2).

Table 1 Summary of *de novo* transcriptome assembly by Trinity. We assembled raw data from two different libraries using the Trinity program. The statistics of assembled contigs were calculated by TrinityStats.pl in the Trinity program

	SRR1752764
Total trinity genes	69183
Total trinity transcripts	116108
Percent GC	43.97
Contig N50	710 bp
Median contig length	428 bp
Average contig	580.18 bp
Total assembled bases	67,363,642 bp

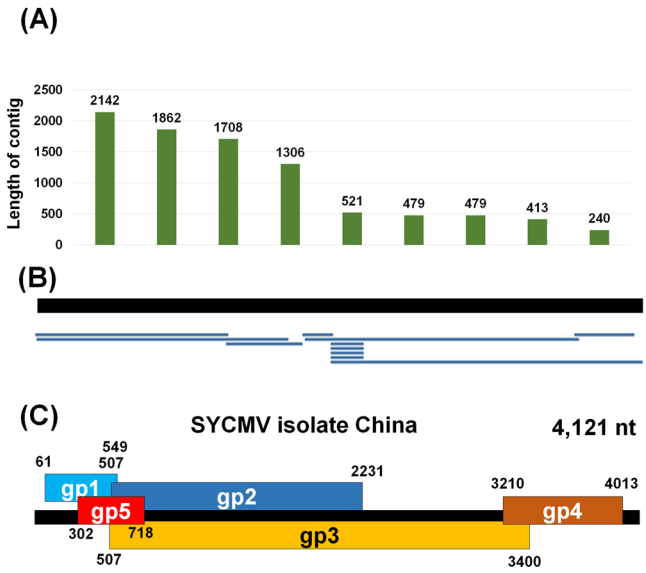

Fig. 1 *De novo* assembly of SYCMV isolate China using transcriptome data. (A) Size distribution of SYCMV-associated contigs. Green-colored bar indicates SYCMV-associated contigs with respective sequence lengths. (B) Alignment of nine SYCMV-associated contigs on the assembled genome of SYCMV isolate China using BWA program and visualized by Tablet program. The reference SYCMV genome is indicated by a black bar, and the blue-colored bars indicate SYCMV-associated contigs. (C) Genome organization of SYCMV isolate China. The nucleotide positions of five genes, gp1 to gp5, are indicated

De novo assembly of SYCMV genome and its genome organization

The sizes of the nine contigs associated with SYCMV were relatively long ranging from 240 nt (nucleotides) to 2,142 nt (Fig. 1A). After alignment of the nine contigs to the SYCMV reference genome sequence (accession number NC_016033.1), we found that the nine contigs mostly covered the SYCMV genome (Fig. 1B). Through sequence alignment with manual modification, we assembled a nearly complete genome of SYCMV referred to as

Table 2 Summary of MegaBLAST result to identify virus-associated contigs

Query id	Subject ids	Identity (%)	Alignment length	Mis-matches	Gap opens	Query start	Query end	Subject start	Subject end	Evalue	Bit score
TR91382|c0_g2_i1	gi|347762083|ref|NC_016033.1|	94.67	2120	113	0	1	2120	4138	2019	0	3290
TR93259|c0_g1_i1	gi|347762083|ref|NC_016033.1|	95.1	1859	91	0	4	1862	1852	3710	0	2929
TR55604|c0_g1_i1	gi|347762083|ref|NC_016033.1|	89.7	1709	174	2	1	1708	1730	23	0	2180
TR48049|c0_g1_i1	gi|347762083|ref|NC_016033.1|	88.51	1306	150	0	1	1306	18	1323	0	1581
TR47810|c0_g1_i1	gi|347762083|ref|NC_016033.1|	94.44	522	27	2	1	521	1311	1831	0	802
TR83344|c0_g1_i1	gi|347762083|ref|NC_016033.1|	93.22	413	28	0	1	413	3677	4089	5.00E-173	608
TR91382|c0_g1_i1	gi|347762083|ref|NC_016033.1|	97.86	234	4	1	225	457	2252	2019	3.00E-111	403
TR91382|c0_g1_i1	gi|347762083|ref|NC_016033.1|	98.25	229	4	0	1	229	2247	2019	1.00E-110	401
TR46265|c0_g1_i1	gi|347762083|ref|NC_016033.1|	95.59	204	9	0	37	240	2036	1833	8.00E-89	327

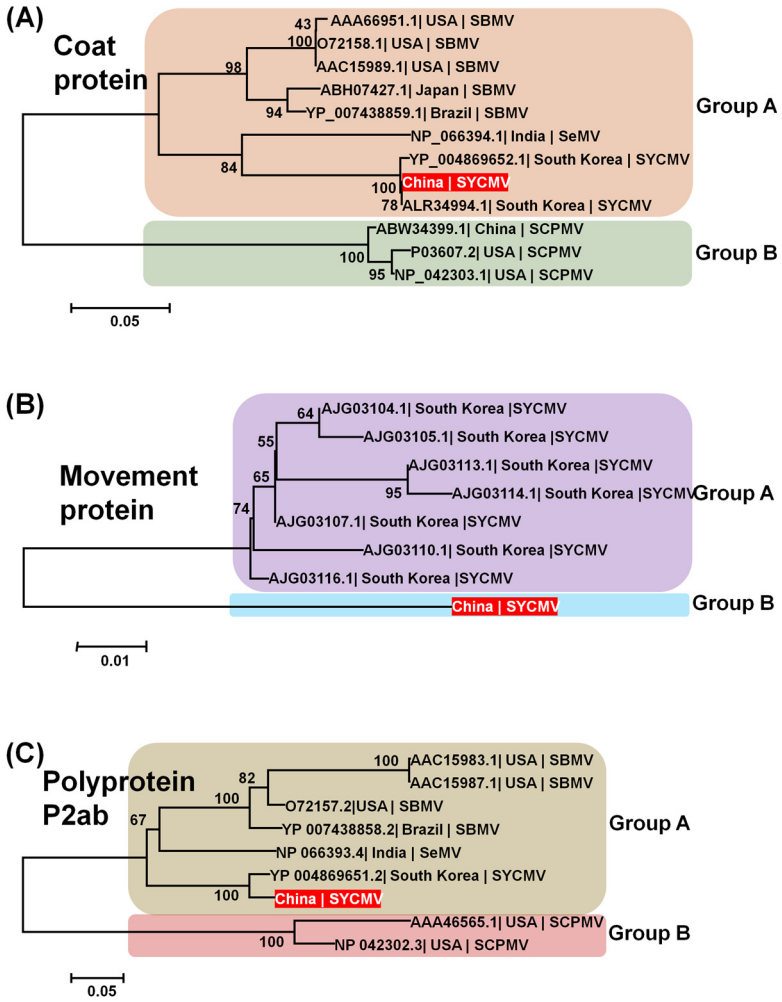

Fig. 2 Phylogenetic position of the assembled SYCMV isolate China based on three protein sequences. Phylogenetic trees of SYCMV isolates using coat protein (A), movement protein (B), and polyprotein P2ab sequences. The respective protein sequences were blasted against the NCBI database, and highly matched sequences were used for the construction of phylogenetic trees using the MEGA6 program. The neighbor-joining method and Poisson substitution model with 1,000 bootstrap replications were used

isolate China. The size of SYCMV isolate China (accession number KX096577) is 4,121 nt, and it contains five genes (Fig. 1C). As compared to the reference SYCMV genome, the genome of SYCMV isolate China is 17 nt and 14 nt shorter than the reference genome at the N-terminal and C-terminal, respectively. These regions are non-translated regions and are usually highly conserved among virus isolates.

The gp1 gene (nt 61 to 549) encodes a silencing suppressor protein with 162 amino acids (aa), and it might also encode a putative movement protein (Fig. 1C). The ORFx (nt 302-718) encodes the gp5 protein with 138 aa that possesses non-AUG initiation. The gp3 gene (nt 507-3,400) encodes a P2ab polyprotein with 964 aa, which is further cleaved into three mature proteins: putative polyprotein membrane anchor, protease, and VPg protein. The gp2 gene (nt 302-718) encodes a P2a polyprotein with 138 aa, which is further cleaved into five mature proteins: membrane anchor, protease, and VPg, P10, and P8 proteins. The gp4 gene (nt 3,210-4,013) encodes a capsid protein with 267 aa.

Phylogenetic position of SYCMV isolate China

We examined the phylogenetic position of the identified SYCMV isolate China. To our knowledge, the assembled SYCMV isolate China is the second SYCMV genome. Therefore, three protein sequences of SYCMV instead of genome sequences were subjected to phylogenetic analyses. Using coat protein sequences, SYCMV was closely related to two isolates of SYCMV from South Korea (Fig. 2A). In addition, SYCMV isolate China was found to belong to group A possessing five other *Southern bean mosaic virus* (SBMV) isolates and a *Sesbania mosaic virus* (SeMV). In the phylogenetic tree using movement proteins, SYCMV isolate China was distantly related with seven other SYCMV isolates from Korea (Fig. 2B). The phylogenetic tree using polyprotein

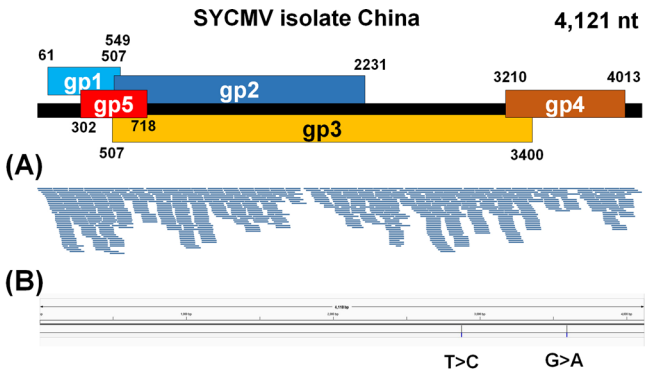

Fig. 3 SNVs of SYCMV in the soybean flower bud transcriptome. (A) The positions of identified SNVs on SYCMV were visualized by the Tablet program. (B) Positions of two identified SNVs on the SYCMV genome

P2ab sequences exhibited two distinct clades (Fig. 2C). Group A contains two SYCMV isolates, four SBMV isolates, and a SeMV isolate. SYCMV is closely related with SYCMV isolate from South Korea. Based on the phylogenetic trees, SYCMV isolate China is a new member of SYCMV that has strong sequence similarity to the SYCMV isolates from South Korea.

Quasispecies of SYCMV in the soybean bud transcriptome

It is widely known that many RNA viruses are highly mutated in the infected host and several variants are exhibited in the same host, which is known for quasispecies. To examine SYCMV quasispecies, we analyzed SNVs for SYCMV in the soybean transcriptome. The assembled genome for SYCMV isolate China was used as a reference. Using the BWA program, we mapped raw data on the reference genome and subjected the aligned data to SNV calling using SAMtools (Fig. 3A). Many reads covering most regions of SYCMV were mapped on the SYCMV genome. Unexpectedly, we identified only two SNVs: T to C and G to A conversion at nt 2,875 and 3,594, respectively (Fig. 3B and Supplementary Table 1).

Number of viral RNAs in the soybean flower bud transcriptome

We thought it would be interesting to examine how many SYCMV viral RNAs were present in the soybean flower bud transcriptome. Thus, we calculated the number of SYCMV RNAs using assembled SYCMV-associated contigs and SYCMV-associated raw reads. Based on the assembled contigs, SYCMV accounted for 0.7751% (9/116,108 contigs), while SYCMV accounted for 0.001948% (487/25,004,615 reads) according to the number of reads. The copy number for SYCMV in the transcriptome was about 11.93, indicating sequence coverage of the SYCMV genome.

Discussion

Plants are frequently infected by diverse viruses. Viral infections in plants are generally identified based on disease symptoms followed by diagnostic methods, such as ELISA and PCR [14,15]. Viral infections in plants do not always cause disease symptoms, and many plants infected by different viruses are asymptomatic [29]. Recently, RNA-Seq using NGS techniques was widely applied in the transcriptome analyses of diverse plant species [30]. Plant transcriptome analyses in response to specific viruses reveal viral infections; however, plant transcriptomes that are not associated with viruses do not have any information associated with viruses even if the plant transcriptomes are infected by viruses. Of course, in many cases, NGS data are easily contaminated by unknown pathogens and vectors during sample and library preparation [31]. Therefore, the removal of non-plant sequences from the raw data is a necessary step. In contrast, plant transcriptomes infected by viruses might be of interest for plant virologists.

Based on the assumption that some plant transcriptomes might be infected by viruses, we performed a large-scale screening of soybean transcriptomes for virus infection. Based on our results, only few transcriptomes were infected by a specific virus, and most transcriptomes did not possess any viral sequences (unpublished data). Although the plant transcriptomes contained viral sequences, the number of viral sequences was very low. During our initial screening, we identified a soybean transcriptome that was solely infected by SYCMV [20]. Based on the previous study [20] in which we conducted soybean transcriptome analyses, no viral infections in soybean samples have been reported. This result suggests that plant samples were used for RNA-Seq without any knowledge on virus infection. *De novo* transcriptome assembly followed by sequence alignment on the reference SYCMV genome enabled us to assemble a nearly complete genome of SYCMV isolate China. To our knowledge, this is the second genome for SYCMV so far. In addition, phylogenetic analyses using three viral proteins confirmed that SYCMV isolate China is closely related with known SYCMV isolates from South Korea. Our results demonstrated that it is possible to assemble viral genomes using plant transcriptome data, as shown previously [18,19]. However, virus sequence coverage is important for virus assembly, as shown in the raw sequence alignment on the SYCMV genome.

Virus replication might be dependent on plant tissues. In this study, the transcriptome was derived from soybean flower buds. The identification of SYCMV in soybean flower buds indicates replication of SYCMV in this tissue. However, the number of virus-associated reads in the flower buds was very low. As mentioned in the Materials and Methods section, the soybean transcriptome was derived from mixed flower bud samples. Therefore, we expected strong sequence variations of SYCMV. However, our SNV analysis revealed only two SNVs, indicating a low level of SYCMV mutation in the flower bud. A previous study showed that the rate of virus replication and mutation might be dependent on tissues and developmental stages [19]. Therefore, the low level of viral replication and mutation might be associated with soybean tissue. Based on these results, we assume that flower

buds are not favorable tissues for virus replication, because flower buds are reproductive tissues rather than vegetative tissues.

In fact, the previous study examined two different soybean lines: CMS and non-CMS lines [20]. As we have shown in this study, interestingly, the non-CMS line was infected by SYCMV, while the CMS line did not contain any virus-associated reads. Based on the previous study, two different soybean plants were grown in the same growth conditions. This result indicates that the possibility of seed transmission for SYCMV in the non-CMS line might be very high. Otherwise, the CMS line should also have been infected by SYCMV through insect vectors. We carefully assume that SYCMV is a seed-transmitted virus like SMV and male sterility in the CMS line blocked the transmission of SYCMV from the non-CMS line to the CMS line during hybridization. In plants, male sterility is defined as the failure of functional pollens, resulting in the production of unviable male gametes [32]. Male sterility is regulated by mitochondrial genes along with nuclear genes. CMS is currently being used for hybrid breeding to increase the productivity of several crop plants [33]. A previous study examined the possible association of viruses or viroids with male sterility using petunia plants [34]. This study demonstrated that viruses were eliminated in a CMS line after heat and cold treatment followed by tissue culture using isolated apical meristem. Thus, we assume that CMS might be useful to generate virus-free plants.

Taken together, we identified SYCMV from the soybean flower bud transcriptome and successfully assembled a nearly complete genome of SYCMV by bioinformatics analyses. Furthermore, we assume that the low level of virus replication and mutation for SYCMV is associated with plant tissues. Moreover, we provide the first evidence that male sterility might be used to eliminate viruses in crop plants.

Acknowledgment This work was carried out with the support of the "Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ01498301)" conducted by the Rural Development Administration, Republic of Korea.

References

1. Singh R, Hymowitz T (1999) Soybean genetic resources and crop improvement. Genome 42: 605–616
2. Messina MJ (1999) Legumes and soybeans: overview of their nutritional profiles and health effects. Am J Clin Nutr 70: 439s–450s
3. Pimentel D, Patzek TW (2005) Ethanol production using corn, switchgrass, and wood; biodiesel production using soybean and sunflower. Nat Resour Res 14: 65–76
4. Herridge DF, Peoples MB, Boddey RM (2008) Global inputs of biological nitrogen fixation in agricultural systems. Plant Soil 311: 1–18
5. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463: 178–183
6. Wrather J, Stienstra W, Koenning S (2001) Soybean disease loss estimates for the United States from 1996 to 1998. Can J Plant Pathol 23: 122–131
7. Hill JH, Whitham SA (2014) Control of Virus Diseases in Soybeans. Control of Plant Virus Diseases: Seed-Propagated Crops 90: 355
8. Domier LL, Hobbs HA, McCoppin NK, Bowen CR, Steinlage TA, Chang S, Wang Y, Hartman GL (2011) Multiple loci condition seed transmission of Soybean mosaic virus (SMV) and SMV-induced seed coat mottling in soybean. Phytopathology 101: 750–756
9. Hobbs HA, Hill CB, Grau CR, Koval NC, Wang Y, Pedersen WL, Domier LL, Hartman GL (2006) Green stem disorder of soybean. Plant Dis 90: 513–518
10. Calvert L, Ghabrial S (1983) Enhancement by soybean mosaic virus of bean pod mottle virus titre in doubly infected soybeans. Phytopathology 73: 992–997
11. Domier LL, Steinlage TA, Hobbs HA, Wang Y, Herrera-Rodriguez G, Haudenshield JS, McCoppin NK, Hartman GL (2007) Similarities in seed and aphid transmission among Soybean mosaic virus isolates. Plant Dis 91: 546–550
12. Mabry TR, Hobbs HA, Steinlage TA, Johnson BB, Pedersen WL, Spencer JL, Levine E, Isard SA, Domier LL, Hartman GL (2003) Distribution of leaf-feeding beetles and Bean pod mottle virus (BPMV) in Illinois and transmission of BPMV in soybean. Plant Dis 87: 1221–1225
13. Fulton J, Cumberland D, Hodgsen O, Amsoy C, Vickery W (1983) Bean pod mottle virus: occurrence in Nebraska and seed transmission in soybeans. Plant Dis 67: 230-233
14. Mowat W, Dawson S (1987) Detection and identification of plant viruses by ELISA using crude sap extracts and unfractionated antisera. J Virol Methods 15: 233–247
15. Thomson D, Dietzgen RG (1995) Detection of DNA and RNA plant viruses by PCR and RT-PCR using a rapid virus release protocol without tissue homogenization. J Virol Methods 54: 85–95
16. Barba M, Czosnek H, Hadidi A (2014) Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses 6: 106–136
17. Li R, Gao S, Hernandez AG, Wechter WP, Fei Z, Ling K-S (2012) Deep sequencing of small RNAs in tomato for virus and viroid identification and strain differentiation. PLoS One 7: e37127
18. Jo Y, Choi H, Yoon J-Y, Choi S-K, Cho WK (2016) *In silico* identification of Bell pepper endornavirus from pepper transcriptomes and their phylogenetic and recombination analyses. Gene 575: 712–717
19. Jo Y, Choi H, Cho JK, Yoon J-Y, Choi S-K, Cho WK (2015) *In silico* approach to reveal viral populations in grapevine cultivar Tannat using transcriptome data. Scientific Rep 5: 15841
20. Li J, Han S, Ding X, He T, Dai J, Yang S, Gai J (2015) Comparative transcriptome analysis between the cytoplasmic male sterile line njcms1a and its maintainer njcms1b in soybean (Glycine max (L.) Merr.). PLoS One 10: e0126771
21. Leinonen R, Sugawara H, Shumway M (2010) The sequence read archive. Nucleic Acids Res 39: D19-D21
22. Madden T (2013) The BLAST sequence analysis tool. The NCBI Handbook, National Center for Biotechnology Information (US)
23. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N, Regev A (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Prot 8: 1494–1512
24. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:

2725–2729
25. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079
27. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST (2011) The variant call format and VCFtools. Bioinformatics 27: 2156–2158
28. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D (2010) Tablet-next generation sequence assembly visualization. Bioinformatics 26: 401–402
29. Roossinck MJ (2010) Lifestyles of plant viruses. Phil Trans R Soc B 365: 1899–1905
30. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63
31. Strong MJ, Xu G, Morici L, Bon-Durant SS, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK (2014) Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog 10: e1004437
32. Chase CD (2007) Cytoplasmic male sterility: a window to the world of plant mitochondrial–nuclear interactions. Trends in Genet 23: 81–90
33. Bohra A, Jha UC, Adhimoolam P, Bisht D, Singh NP (2016) Cytoplasmic male sterility (CMS) in hybrid breeding in field crops. Plant Cell Rep 35: 967–993
34. Evenor D, Joseph MB, Izhar S (1988) Attempts to detect extra genomial factors in cytoplasmic male-sterile petunia lines. Theor Appl Genet 76: 455–458