Plant Pathol. J. 33(5) : 478-487 (2017)
https://doi.org/10.5423/PPJ.OA.03.2017.0060
pISSN 1598-2254 eISSN 2093-9280

Research Article Open Access

The Plant Pathology Journal
©The Korean Society of Plant Pathology

De novo Genome Assembly and Single Nucleotide Variations for Soybean Mosaic Virus Using Soybean Seed Transcriptome Data

Yeonhwa Jo[1], Hoseong Choi[1], Miah Bae[1], Sang-Min Kim[2], Sun-Lim Kim[2], Bong Choon Lee[2], Won Kyong Cho[1]*, and Kook-Hyung Kim[1]*

[1]Department of Agricultural Biotechnology, Research Institute of Agriculture and Life Sciences, and Plant Genomics and Breeding Institute, College of Agriculture and Life Sciences, Seoul National University, Seoul 08826, Korea
[2]Crop Foundation Division, National Institute of Crop Science, RDA, Wanju 55365, Korea

(Received on March 20, 2017; Revised on June 7, 2017; Accepted on June 27, 2017)

Soybean is the most important legume crop in the world. Several diseases in soybean lead to serious yield losses in major soybean-producing countries. Moreover, soybean can be infected by diverse viruses. Recently, we carried out a large-scale screening to identify viruses infecting soybean using available soybean transcriptome data. Of the screened transcriptomes, a soybean transcriptome for soybean seed development analysis contains several virus-associated sequences. In this study, we identified five viruses, including soybean mosaic virus (SMV), infecting soybean by *de novo* transcriptome assembly followed by blast search. We assembled a nearly complete consensus genome sequence of SMV China using transcriptome data. Based on phylogenetic analysis, the consensus genome sequence of SMV China was closely related to SMV isolates from South Korea. We examined single nucleotide variations (SNVs) for SMVs in the soybean seed transcriptome revealing 780 SNVs, which were evenly distributed on the SMV genome. Four SNVs, C-U, U-C, A-G, and G-A, were frequently identified. This result demonstrated the quasispecies variation of the SMV genome. Taken together, this study carried out bioinformatics analyses to identify viruses using soybean transcriptome data. In addition, we demonstrated the application of soybean transcriptome data for virus genome assembly and SNV analysis.

Keywords **:** *de novo* genome assembly, single nucleotide variation, soybean mosaic virus

Handling Associate Editor **:** Jeon, Junhyun

Soybean (*Glycine max* (L.) Merr.) is the most important legume crop, representing 50% of the global legume crop area and 68% of global legume production (Herridge et al., 2008). Soybean is consumed as health food, providing a rich source of proteins, and as well as vegetative oil production (Messina, 1999; Pimentel and Patzek, 2005). Moreover, soybean plays an important role for dinitrogen (N_2) fixation, which is an important natural process (Herridge et al., 2008).

Several diseases in soybean, such as cyst, brown spot, charcoal rot, and sclerotinia stem rot, lead to yield losses in major soybean-producing countries (Wrather et al., 2001). In addition, soybean can be infected by diverse viruses. Although a small numbers of viruses infecting soybean cause serious economic problems in soybean production, it is always important to control and to manage viral diseases in soybeans (Hill and Whitham, 2014). The best known soybean virus is *Soybean mosaic virus* (SMV), a member of the family *Potyviridae*, causing soybean mosaic disease. In addition, bean pod mottle virus (BPMV), soybean vein necrosis virus, tobacco ringspot virus, soybean dwarf virus, peanut mottle virus, peanut stunt virus, and alfalfa mosaic virus are important viruses infecting soybeans (Hill and

*Co-corresponding authors.
WK Cho
Phone) +82-2-880-4687, FAX) +82-2-873-2317
E-mail) wonkyong@gmail.com
K-H Kim
Phone) +82-2-880-4677, Fax) +82-2-873-2317
E-mail) kookkim@snu.ac.kr

🅭 This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Articles can be freely viewed online at www.ppjonline.org.

Whitham, 2014).

Many plant viruses have been identified based on viral disease symptoms and several detection methods. However, virus infection in plants does not always cause disease symptoms, and many plants showing viral disease symptoms are very often co-infected by different viruses. Recent advances in next generation sequencing (NGS) technology lead to identification of numerous known as well as novel viruses by means of metagenomics (Barba et al., 2014; Massart et al., 2014). Not only NGS data for virus detection but also many plant transcriptome data contain virus sequences, which might be amplified along with infected host transcripts (Burger and Maree, 2015; Jo et al., 2016). The identification of virus sequences in the plant transcriptome is no longer surprising, because most plant viruses are RNA viruses and many of them carry poly(A) tail, which is easily amplified by oligo d(T) primers for cDNA synthesis.

Recently, we carried out a large-scale screening to identify viruses infecting soybean in the world using available soybean transcriptome data. Of them, we found that a soybean transcriptome for soybean seed development analysis contains many virus sequences. In this study, we conducted a bioinformatics analyses for virus identification, virus genome assembly, phylogenetic analysis, and single nucleotide variations of the SMV.

Materials and Methods

Plant materials, library preparation, and next generation sequencing. The plant material used for RNA-Seq was soybean cultivar Heinong44. Plants were grown in the experimental station in Beijing from May to August according to the previous study (Song et al., 2013). Total RNAs were extracted from seeds at six different developmental stages, which were classified according to the seed weight. The cDNA was synthesized using poly(A)-containing RNAs. A single RNA-Seq library was constructed and sequenced by single-end sequencing using the Illumina HiSeq 2000 system. The raw data is available in the SRA database (http://www.ncbi.nlm.nih.gov/sra/SRR1777405).

Raw data processing and *de novo* transcriptome assembly. All bioinformatics analyses were performed in the Linux (Linux Mint version 17)-installed workstation (four 16-core CPUs and 256 GB ram). We downloaded the raw data from the SRA database using the SRA toolkit (Leinonen et al., 2011). The raw SRA data were converted to FASTQ files using the SRA toolkit. For the *de novo* assembly of transcriptomes, we used Trinity version 2.0.6 (Haas et al., 2013). *De novo* transcriptome assembly was performed according to the manuals provided by developers with default parameters.

Identification of viruses and sequence alignment. To identify virus-associated contigs, we conducted blast search using standalone BLAST version 2.1.19 installed in the Linux system (Madden, 2013). All assembled contigs were subjected to MEGABLAST search, which is optimized for highly similar sequences, against complete reference sequences for viruses and viroids (http://www.ncbi.nlm.nih.gov/genome/viruses/) with E value 1e-5 as a cutoff. In addition, all raw data were converted to FASTA files using the SRA toolkit and subjected to a MEGABLAST search against the viral reference database with E value 1e-5 as a cutoff. We used the Burrows–Wheeler Aligner (BWA) software for sequence alignment on the reference virus genome with default parameters (Li and Durbin, 2009).

***De novo* assembly of SMV genomes.** The 79 SMV-associated contigs identified by the BLAST were retrieved by the BLASTCMD program in the standalone BLAST system. To assemble SMV genomes, the identified viral contigs were aligned against the SMV reference genome (NC_002634.1) using ClustalW implemented in the MEGA6 program (Tamura et al., 2013) The nearly complete consensus genome of SMV was manually obtained. Raw data were again aligned on the assembled consensus SMV genome to confirm sequences by BWA. The poly(A) tail at the 3’ end of the assembled SMV genome was removed. We obtained a nearly complete consensus genome for SMV China (accession number NC_002634.1) from soybean transcriptome.

Identification of SNVs in soybean transcriptome. In order to analyze SNVs of SMV China in the soybean transcriptome, the raw data were aligned on the consensus genome of SMV China using the BWA program with default parameters. The aligned SAM files by BWA were converted into BAM files by SAMtools (Li et al., 2009). For SNV calling, we sorted the BAM files and then generated the VCF file format using mpileup (Danecek et al., 2011). BCFtools implemented in SAMtools was finally used to call SNVs. The positions of identified SNVs on the SMV genome were visualized by the Tablet program (Milne et al., 2010).

Construction of phylogenetic trees. In order to reveal phylogenetic relationships of the obtained consensus genome for SMV China with known SMV isolates, we generated three phylogenetic trees. The complete SMV isolate China genome sequence as well as two polyprotein sequences were blasted against NCBI nucleotide and non-redundant protein databases. Best-matched sequences were retrieved for the construction of phylogenetic tree. The obtained sequences were aligned by the ClustalW program with default parameters. After alignment, we deleted unnecessary sequences. The manually edited aligned sequences were subjected to construction of a phylogenetic tree using the MEGA6 program. The phylogenetic tree was constructed by the neighbor-joining method, with 1,000 bootstrap replicates.

Results

De novo soybean transcriptome assembly and identification of viruses in the soybean seeds. We screened available soybean transcriptome data deposited in NCBI's Sequence Read Archive (SRA) database in order to identify viruses infecting soybean. Of screened soybean transcriptomes, a transcriptome conducting a gene expression profile during soybean seed development contains several virus-associated sequences (accession number SRR1777405) (Song et al., 2013). In order to identify virus-associated contigs, we *de novo* assembled the transcriptome of soybean using Trinity program, resulting in 116,108 transcripts (contigs) with 710 bp for contig N50 (Table 1). Next, we blasted 116,108 transcripts against the viral reference database. After removing redundant sequences and endogenous viral sequences, we identified 83 contigs-associated with viruses (Table 2). Most contigs (79 contigs) were associated with SMV. The lengths of SMV-associated contigs ranged from 224 to 3,636 nt (Fig. 1A). Four contigs were associated with BPMV, lettuce infectious yellow virus (LICV), lettuce chlorosis virus (LCV), and cucumber mosaic virus (CMV), respectively. The lengths of contigs associated with the four viruses ranged from 232 nt (LCV RNA2) to 1,015 nt (bean common mosaic virus) (Fig. 1A). Other than a contig-associated with LICV (1E-08), virus-associated contigs display reliable E values indicating significance of blast results (Table 2).

Table 1. Summary of *de novo* soybean transcriptome assembly using Trinity

Accession number	SRR1777405[a]
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	67363642 bp

[a] We assembled raw data from two different libraries using Trinity program. The statistics of assembled contigs were calculated by TrinityStats.pl in the Trinity program.

De novo genome assembly of SMV from a soybean transcriptome. Of identified viruses, SMV was severely infected in the soybean seeds. Fortunately, 79 contigs associated with SMV mostly covered the SMV reference genome (Table 2). A total of 79 contigs associated with SMV were mapped on the SMV reference genome (accession number NC_002634.1) (Eggenberger et al., 1989) (Fig. 1B). After sequence alignment followed by manual modification, we assembled a nearly complete consensus genome of SMV referred as SMV China (Fig. 1C). The SMV China is composed of 9,507 nucleotides (nt) encoding two proteins such as GP1 and GP2. GP1 encodes a polyprotein (nt 54 to 9,254) which is further cleaved into ten mature proteins such as P1 (P1 proteinase), HC-Pro (helper component proteinase), P3 (P3 protein), 6K1 (6K1 protein), CI (cylindrical inclusion), 6K2, NIa-VPg (Nuclear inclusion protein a-genome linked viral protein), NIa-Pro, NIb (nuclear inclusion protein b), and coat protein (CP) while GP2 encodes PIPO (pretty interesting potyviridae ORF) protein (nt 2,804 to 3,031)

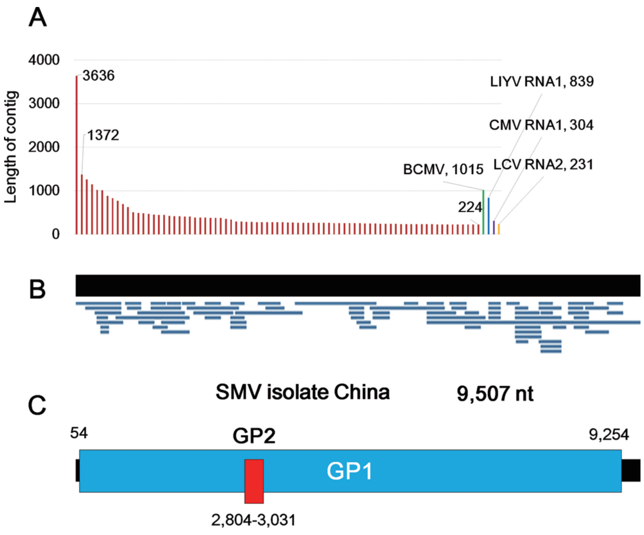

Fig. 1. *De novo* assembly of SMV isolate in China using transcriptome data. (A) Size distribution of virus-associated contigs. Red-colored bar indicates SMV-associated contigs. Four viruses with respective contig length were indicated. (B) Alignment of 79 SMV-associated contigs on the assembled genome of SMV isolate in China using BWA program. Black bar indicates the reference SMV genome. Sequence alignment was visualized by Tablet program. (C) Genome organization of SMV isolate in China. The nucleotide positions of two proteins, GP1 and GP2, were indicated.

Table 2. Summary of blast results to identify virus-associated contigs

Query id	Subject id	Name of virus	Identity (%)	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	E value	Bit score
TR2274|c0_g1_i1	NC_002634.1	Soybean mosaic virus	93.13	233	16	0	2	234	8571	8803	3.00E-93	342
TR3618|c0_g1_i1	NC_002634.1	Soybean mosaic virus	91.02	256	23	0	1	256	1342	1597	2.00E-94	346
TR3618|c0_g1_i2	NC_002634.1	Soybean mosaic virus	90.58	276	26	0	1	276	1342	1617	2.00E-100	366
TR3858|c0_g1_i1	NC_002634.1	Soybean mosaic virus	97.35	264	7	0	1	264	910	1173	2.00E-125	449
TR3858|c0_g1_i2	NC_002634.1	Soybean mosaic virus	96.6	235	8	0	1	235	939	1173	1.00E-107	390
TR4672|c0_g1_i1	NC_002634.1	Soybean mosaic virus	96.55	261	9	0	1	261	9036	9296	2.00E-120	433
TR4672|c0_g1_i2	NC_002634.1	Soybean mosaic virus	97.7	261	6	0	1	261	9036	9296	2.00E-125	449
TR5077|c1_g1_i1	NC_002634.1	Soybean mosaic virus	94.19	258	15	0	3	260	4680	4937	9.00E-109	394
TR5077|c1_g1_i2	NC_002634.1	Soybean mosaic virus	91.47	258	22	0	3	260	4680	4937	4.00E-97	355
TR5102|c0_g1_i1	NC_002634.1	Soybean mosaic virus	91.96	224	18	0	1	224	7552	7329	6.00E-85	315
TR5869|c0_g1_i1	NC_002634.1	Soybean mosaic virus	91.98	212	17	0	5	216	7243	7032	6.00E-80	298
TR5869|c0_g2_i1	NC_002634.1	Soybean mosaic virus	92.45	212	16	0	5	216	7243	7032	1.00E-81	303
TR5869|c0_g3_i1	NC_002634.1	Soybean mosaic virus	92.92	212	15	0	5	216	7243	7032	3.00E-83	309
TR5869|c0_g4_i1	NC_002634.1	Soybean mosaic virus	92.92	212	15	0	5	216	7243	7032	3.00E-83	309
TR7406|c0_g1_i1	NC_002634.1	Soybean mosaic virus	94.64	280	15	0	1	280	2677	2956	6.00E-121	435
TR7406|c0_g1_i2	NC_002634.1	Soybean mosaic virus	92.12	241	19	0	1	241	2677	2917	1.00E-92	340
TR7406|c0_g1_i3	NC_002634.1	Soybean mosaic virus	94.16	274	16	0	1	274	2677	2950	5.00E-116	418
TR7406|c0_g1_i4	NC_002634.1	Soybean mosaic virus	93.36	241	16	0	1	241	2677	2917	1.00E-97	357
TR8100|c0_g1_i1	NC_002634.1	Soybean mosaic virus	97.86	234	5	0	12	245	6060	6293	4.00E-112	405
TR9520|c0_g1_i1	NC_002634.1	Soybean mosaic virus	95.06	385	19	0	1	385	8268	7884	2.00E-172	606
TR9520|c0_g1_i2	NC_002634.1	Soybean mosaic virus	96.65	239	8	0	4	242	8122	7884	6.00E-110	398
TR9520|c0_g1_i3	NC_002634.1	Soybean mosaic virus	94.66	356	19	0	1	356	8268	7913	2.00E-156	553
TR9520|c0_g1_i4	NC_002634.1	Soybean mosaic virus	94.38	356	20	0	1	356	8268	7913	9.00E-155	547
TR9520|c0_g1_i5	NC_002634.1	Soybean mosaic virus	95.06	385	19	0	1	385	8268	7884	2.00E-172	606
TR9520|c0_g1_i6	NC_002634.1	Soybean mosaic virus	96.19	210	8	0	4	213	8122	7913	8.00E-94	344
TR9520|c0_g1_i7	NC_002634.1	Soybean mosaic virus	96.88	385	12	0	1	385	8268	7884	0	645
TR13605|c0_g1_i1	NC_002634.1	Soybean mosaic virus	92.25	400	31	0	10	409	8665	9064	8.00E-161	568
TR13605|c0_g1_i2	NC_002634.1	Soybean mosaic virus	94.75	400	21	0	10	409	8665	9064	2.00E-177	623
TR15892|c0_g1_i1	NC_002634.1	Soybean mosaic virus	92.64	231	17	0	2	232	5845	5615	2.00E-90	333

Table 2. Continued

Query id	Subject id	Name of virus	Identity (%)	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	E value	Bit score
TR20496|c0_g1_i1	NC_002634.1	Soybean mosaic virus	96.88	224	7	0	1	224	2087	1864	3.00E-103	375
TR22770|c0_g1_i1	NC_002634.1	Soybean mosaic virus	91.67	240	20	0	1	240	6413	6652	2.00E-90	333
TR22770|c0_g1_i2	NC_002634.1	Soybean mosaic virus	92.53	281	21	0	2	282	6372	6652	2.00E-111	403
TR25078|c0_g1_i1	NC_002634.1	Soybean mosaic virus	88.54	253	29	0	1	253	8730	8478	1.00E-82	307
TR25078|c0_g2_i1	NC_002634.1	Soybean mosaic virus	94.72	246	13	0	16	261	8627	8382	2.00E-105	383
TR25078|c0_g2_i2	NC_002634.1	Soybean mosaic virus	93.7	349	22	0	1	349	8730	8382	2.00E-147	523
TR25078|c0_g2_i3	NC_002634.1	Soybean mosaic virus	95.72	187	8	0	43	229	8568	8382	5.00E-81	302
TR25078|c0_g2_i4	NC_002634.1	Soybean mosaic virus	90.91	253	23	0	1	253	8730	8478	1.00E-92	340
TR32819|c0_g1_i1	NC_002634.1	Soybean mosaic virus	91.7	265	22	0	2	266	2515	2251	6.00E-101	368
TR32819|c0_g2_i1	NC_002634.1	Soybean mosaic virus	92.08	265	21	0	2	266	2515	2251	1.00E-102	374
TR34507|c0_g1_i1	NC_002634.1	Soybean mosaic virus	87.27	377	44	4	4	378	3523	3149	1.00E-118	427
TR37651|c0_g1_i1	NC_002634.1	Soybean mosaic virus	87.61	218	24	3	2	218	410	625	2.00E-65	250
TR37651|c0_g3_i1	NC_002634.1	Soybean mosaic virus	87.27	487	57	4	2	487	410	892	1.00E-155	551
TR37706|c0_g2_i1	NC_002634.1	Soybean mosaic virus	90.51	274	24	2	1	273	1128	1400	9.00E-99	361
TR41793|c1_g1_i1	NC_002634.1	Soybean mosaic virus	92.89	394	28	0	1	394	7483	7876	2.00E-162	573
TR41793|c1_g1_i2	NC_002634.1	Soybean mosaic virus	93.15	438	29	1	1	437	7483	7920	0	641
TR41793|c1_g1_i3	NC_002634.1	Soybean mosaic virus	91.55	213	18	0	23	235	7486	7698	8.00E-79	294
TR41793|c1_g1_i4	NC_002634.1	Soybean mosaic virus	93.93	445	27	0	1	445	7483	7927	0	673
TR41793|c1_g1_i5	NC_002634.1	Soybean mosaic virus	91.59	226	19	0	1	226	7473	7698	2.00E-84	313
TR41793|c1_g1_i6	NC_002634.1	Soybean mosaic virus	93.03	445	31	0	1	445	7483	7927	0	651
TR41793|c1_g1_i7	NC_002634.1	Soybean mosaic virus	91.17	419	37	0	1	419	7473	7891	2.00E-161	569
TR44246|c0_g1_i2	NC_002634.1	Soybean mosaic virus	87.9	157	18	1	87	242	477	633	2.00E-45	183
TR44822|c4_g1_i1	NC_002634.1	Soybean mosaic virus	97.83	460	10	0	2	461	843	384	0	795
TR44822|c4_g1_i2	NC_002634.1	Soybean mosaic virus	97.65	765	18	0	2	766	843	79	0	1314
TR44822|c4_g1_i3	NC_002634.1	Soybean mosaic virus	97.27	622	14	1	2	623	843	225	0	1051
TR44822|c4_g2_i1	NC_002634.1	Soybean mosaic virus	90.13	1256	122	2	1	1255	1991	737	0	1631
TR44822|c4_g2_i2	NC_002634.1	Soybean mosaic virus	91.46	820	70	0	1	820	1918	1099	0	1127
TR44822|c4_g2_i3	NC_002634.1	Soybean mosaic virus	92.75	483	35	0	1	483	1991	1509	0	699
TR44822|c4_g2_i4	NC_002634.1	Soybean mosaic virus	88.67	256	29	0	1	256	1617	1362	2.00E-84	313

Table 2. Continued

Query id	Subject id	Name of virus	Identity (%)	Alignment length	Mismatches	Gap opens	Query start	Query end	Subject start	Subject end	E value	Bit score
TR44822|c4_g2_i5	NC_002634.1	Soybean mosaic virus	93.9	246	15	0	19	264	1853	1608	4.00E-102	372
TR44822|c4_g2_i6	NC_002634.1	Soybean mosaic virus	94.81	231	12	0	19	249	1853	1623	9.00E-99	361
TR44822|c4_g2_i7	NC_002634.1	Soybean mosaic virus	94.15	410	24	0	1	410	1918	1509	5.00E-178	625
TR44822|c5_g1_i1	NC_002634.1	Soybean mosaic virus	95.98	994	40	0	2	995	5991	6984	0	1615
TR44822|c5_g1_i2	NC_002634.1	Soybean mosaic virus	94.11	3599	207	4	2	3596	5991	9588	0	5467
TR44822|c5_g2_i1	NC_002634.1	Soybean mosaic virus	93.3	224	15	0	4	227	8124	8347	6.00E-90	331
TR44822|c5_g1_i3	NC_002634.1	Soybean mosaic virus	96.21	501	19	0	2	502	5991	6491	0	821
TR44822|c5_g1_i4	NC_002634.1	Soybean mosaic virus	92.81	292	21	0	2	293	5991	6282	1.00E-117	424
TR44822|c6_g1_i1	NC_002634.1	Soybean mosaic virus	95.07	1015	50	0	1	1015	6049	5035	0	1598
TR44822|c6_g2_i1	NC_002634.1	Soybean mosaic virus	97.64	212	5	0	10	221	4930	4719	6.00E-100	364
TR44822|c6_g2_i2	NC_002634.1	Soybean mosaic virus	95.83	240	10	0	1	240	5051	4812	4.00E-107	388
TR44822|c6_g2_i3	NC_002634.1	Soybean mosaic virus	97.52	1372	34	0	1	1372	5146	3775	0	2346
TR44822|c6_g2_i4	NC_002634.1	Soybean mosaic virus	96.59	293	10	0	1	293	5146	4854	2.00E-136	486
TR44822|c6_g3_i1	NC_002634.1	Soybean mosaic virus	95.8	691	27	2	5	694	2746	2057	0	1114
TR44822|c6_g3_i2	NC_002634.1	Soybean mosaic virus	96.69	877	29	0	2	878	2822	1946	0	1459
TR44822|c6_g4_i1	NC_002634.1	Soybean mosaic virus	95.39	1149	53	0	1	1149	3889	2741	0	1829
TR44822|c6_g4_i2	NC_002634.1	Soybean mosaic virus	94.89	333	17	0	1	333	3598	3266	5.00E-147	521
TR45256|c0_g1_i1	NC_002634.1	Soybean mosaic virus	93.49	261	17	0	4	264	6897	6637	4.00E-107	388
TR45256|c0_g1_i2	NC_002634.1	Soybean mosaic virus	94.32	229	13	0	2	230	6865	6637	5.00E-96	351
TR47685|c0_g1_i1	NC_002634.1	Soybean mosaic virus	92.53	281	21	0	1	281	5082	5362	2.00E-111	403
TR47685|c0_g2_i1	NC_002634.1	Soybean mosaic virus	92.53	281	21	0	1	281	5082	5362	2.00E-111	403
TR44246|c0_g1_i1	NC_003397.1	Bean common mosaic virus	81.86	408	68	6	490	894	458	862	2.00E-91	339
TR19277|c0_g2_i1	NC_003617.1	Lettuce infectious yellows virus RNA1	75.34	146	29	6	467	607	6837	6694	1.00E-08	63.9
TR45572|c0_g2_i1	NC_012910.1	Lettuce chlorosis virus RNA2	87.96	191	22	1	15	205	8555	8366	1.00E-57	224
TR29303|c0_g1_i1	NC_002034.1	Cucumber mosaic virus RNA1	91.28	298	26	0	4	301	1334	1631	1.00E-112	407

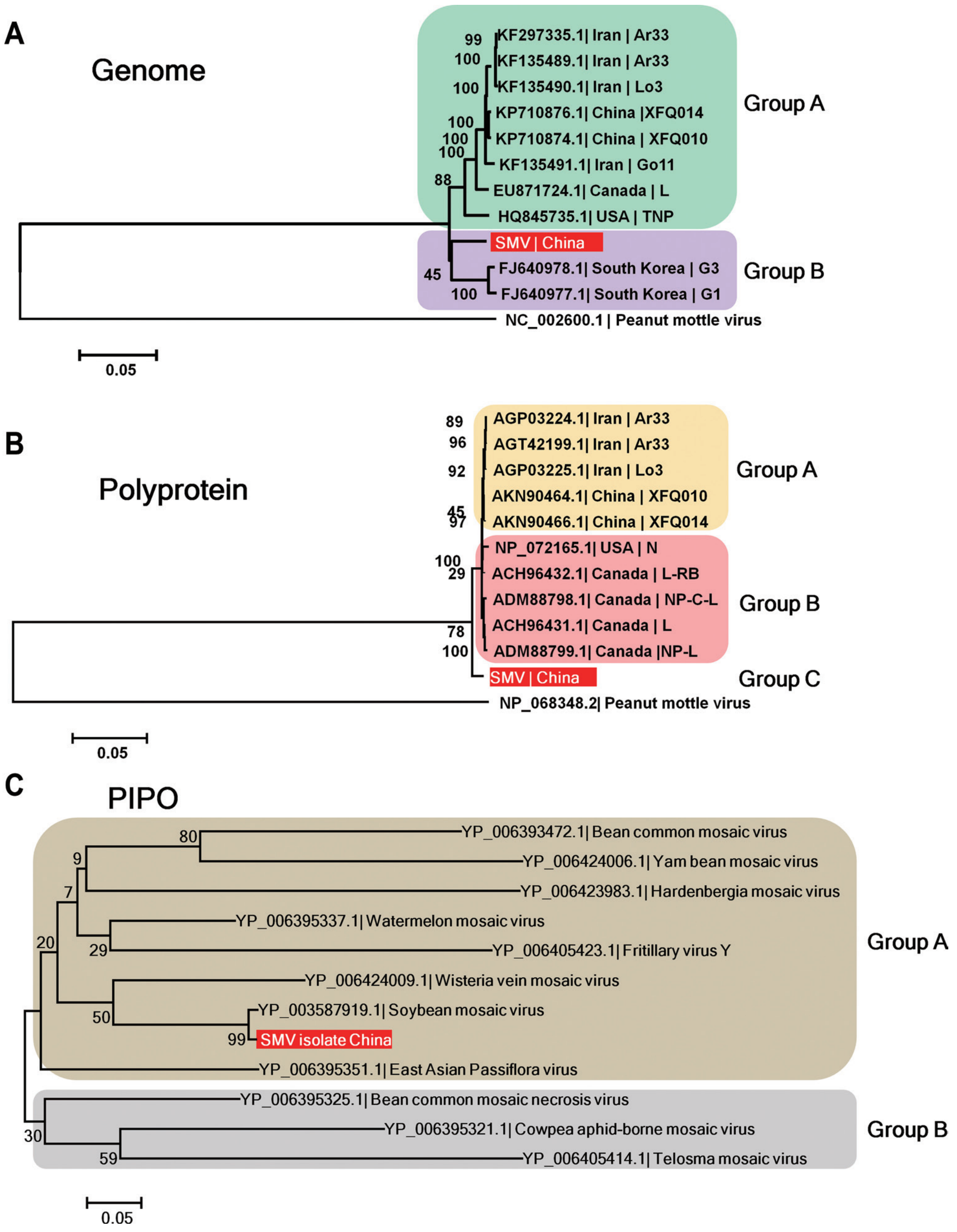
A
Genome
99
100
100
100
100
100
88
45
100
KF297335.1| Iran | Ar33
KF135489.1| Iran | Ar33
KF135490.1| Iran | Lo3
KP710876.1| China |XFQ014
KP710874.1| China | XFQ010
KF135491.1| Iran | Go11
EU871724.1| Canada | L
HQ845735.1| USA | TNP
SMV | China
FJ640978.1| South Korea | G3
FJ640977.1| South Korea | G1
NC_002600.1 | Peanut mottle virus
Group A
Group B
0.05
B
Polyprotein
89
96
92
45
97
100
29
78
100
AGP03224.1| Iran | Ar33
AGT42199.1| Iran | Ar33
AGP03225.1| Iran | Lo3
AKN90464.1| China | XFQ010
AKN90466.1| China | XFQ014
NP_072165.1| USA | N
ACH96432.1| Canada | L-RB
ADM88798.1| Canada | NP-C-L
ACH96431.1| Canada | L
ADM88799.1| Canada |NP-L
SMV | China
NP_068348.2| Peanut mottle virus
Group A
Group B
Group C
0.05
C
PIPO
80
9
7
20
29
50
99
30
59
YP_006393472.1| Bean common mosaic virus
YP_006424006.1| Yam bean mosaic virus
YP_006423983.1| Hardenbergia mosaic virus
YP_006395337.1| Watermelon mosaic virus
YP_006405423.1| Fritillary virus Y
YP_006424009.1| Wisteria vein mosaic virus
YP_003587919.1| Soybean mosaic virus
SMV isolate China
YP_006395351.1| East Asian Passiflora virus
YP_006395325.1| Bean common mosaic necrosis virus
YP_006395321.1| Cowpea aphid-borne mosaic virus
YP_006405414.1| Telosma mosaic virus
Group A
Group B
0.05

Fig. 2. Phylogenetic relationship of the assembled SMV isolate China with known SMV isolates. Phylogenetic trees of SMV isolates using complete genomes (A), polyproteins (B), and PIPO sequences (C). The respective genome and protein sequences were blasted against NCBI database and highly matched sequences were used for construction of phylogenetic trees using MEGA6 program using neighbor-joining method with 1000 bootstrap replications. Kimura 2-parameter and Poisson substitution model were used for nucleotide and protein sequences, respectively.

(Fig. 1C).

Phylogenetic relationships of the SMV isolate China. In order to find genetic relationships of the assembled SMV China with known SMV isolates, we constructed phylogenetic trees. The phylogenetic tree using SMV complete genome sequences showed two groups of SMV isolates (Fig. 2A). The SMV China belongs to group B along with two SMV isolates from South Korea. Using polyprotein sequences, the SMV China in group C was distantly related with other SMV isolates (Fig. 2B). The phylogenetic tree using PIPO protein sequences confirmed that SMV China is a member of SMV belonging to group A, which contains seven viruses including BPMV (Fig. 2C). Based on phylogenetic analyses, it seems that the consensus genome of SMV China is genetically close to the SMV isolates from South Korea.

Single nucleotide variations of SMV in the soybean seeds. It is well known that RNA viruses exhibit quasispecies nature, exhibiting several variants in the infected host. Therefore, we examined single nucleotide variations (SNVs) for SMV in the soybean seeds. The identified SMV China was used as a reference. After BWA align-

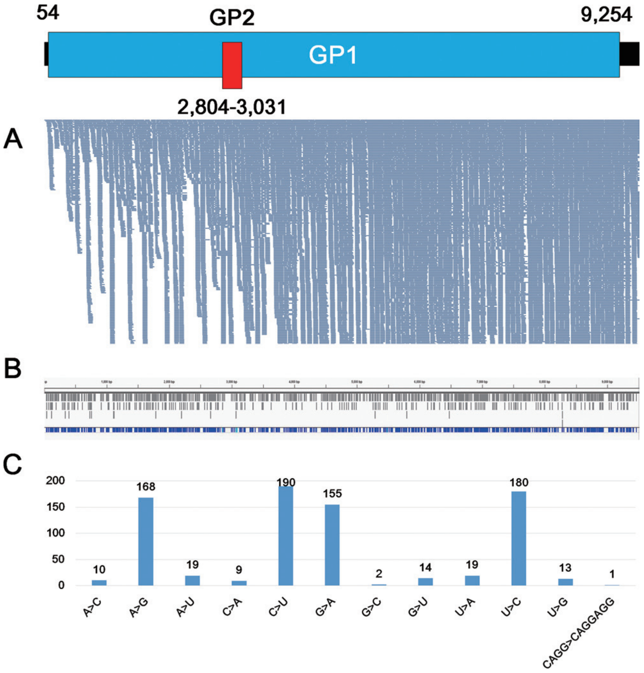

Fig. 3. SNVs of SMV in the soybean seed transcriptome. (A) Raw data were mapped on the genome sequence of SMV isolate China using BWA and visualized by Tablet program. (B) The positions of identified single nucleotide variations on the SMV were visualized by Tablet program. Detailed information for SNVs can be found in Supplementary Table 1. (C) The numbers of identified SNVs of SMV in the soybean seed transcriptome.

ment of raw data against SMV China, SNVs were identified using SAMtools (Fig. 3A). The SNVs in this study was derived from a population of different isolates. As a result, we identified 780 SNVs (Supplementary Table 1). SNVs were evenly distributed along the SMV genome (Fig. 3B). Most SNVs were Single nucleotide polymorphisms (SNPs) except one InDel (CAGG to CAGGAGG) at nt 640 of SMV China (Table S1). Four SNVs, C-U (190 SNVs), U-C (180 SNVs), A-G (168 SNVs), and G-A (155 SNVs), were frequently identified (Fig. 3C). Based on SNV results, the mutation rate for SMV in the soybean seeds was 8.2045%, indicating a high level of mutations for the SMV RNA genome. In addition, we calculated the ratio of Ts/Tv (Transition versus Transversion). The Ts/Tv ratio for SMV China was 8.06 (693/86).

The amount of viral RNA in the soybean transcriptome. It might be of interest to examine viral RNAs in the analyzed soybean transcriptome. Of 116,108 contigs, virus-associated contigs account for 0.068% (79 contigs). The length of total assembled contigs was 67,363,642 bp and the total length of virus-associated contigs 36,022 bp, accounting for 0.0535%. The amount of virus-associated reads accounts for 0.0529% (39,403/74,431,152) of reads. Moreover, we calculated SMV copy numbers within the soybean transcriptome resulting in 414 SMV virus copies, which is highly correlated with sequence coverage of SMV genome. This result indicates high variability of SMV genome.

Discussion

Development of NGS provides various DNA as well as RNA sequencing data (Metzker, 2010). The main purposes of DNA and RNA sequencing is elucidation of the genome and transcriptome of target eukaryotic and prokaryotic organisms (Morozova and Marra, 2008). In case of bacteria, metagenomics using 16s rRNA sequences that are highly conserved in bacteria species is intensively performed to study bacterial communities under specific conditions (Wang and Qian, 2009). However, viruses do not have any conserved sequences like bacteria, and genomes of viruses are mostly very small (Edwards and Rohwer, 2005). Therefore, virus-specific sequencing usually requires a purification step for NGS. For example, extraction of double-stranded RNAs from virus-infected organisms followed by NGS is one of the efficient approaches to identify viruses (Yanagisawa et al., 2016). Moreover, sequencing of small RNAs is an alternative technique for virus identification and genome assembly (Vodovar et al., 2011). In addition, RNA-Seq is also a good technique to identify viruses that have a poly(A) tail. However, several recent studies demonstrated that viruses and viroids without a poly(A) tail can be detected by RNA-Seq (Burger and Maree, 2015; Jo et al., 2016).

In this study, we identified several viruses infecting soybean. This transcriptome was initially conducted for expression profiling of soybean seed development. Thus, this transcriptome is not derived from a single condition but from six developmental seed stages in which several seeds might be included for total RNA extraction. Although we identified five viruses that might infect soybean, four viruses other than SMV were identified based on only one single contig, and their presence should be validated by other methods. In many cases, the partial viral sequence or contig is homologous to a closely related virus, not the target virus. Thus, it is possible that the identified virus-associated contigs might be not from the infected viruses but from other viruses which share similar viral sequences.

SMV is seed-borne and transmitted by aphids (Domier et al., 2011). Soybean seeds infected by SMV often display a discolored and mottled seed. In addition, BCMV is known as a seed-borne virus (Refugee et al., 1987). Seed-borne viruses can be actually infected in embryo, such as BCMV, or carried on the seed coat (Jafarpour et al., 1979). In addition, seed transmission of CMV has been identified in several plants such as pepper, spinach, and lupin (Ali and Kobayashi, 2010; Wylie et al., 1993; Yang et al., 1997). Based on previous knowledge on seed-borne viruses, the identification of SMV, BCMV, and CMV in the soybean seed is not surprising. In addition, the infection of LCV in green bean (*Phaseolus vulgaris* L.) has been recently reported (Ruiz et al., 2014). However, the infection of LIYV and LCV, which are members in the genus *Crinivirus*, in the soybean seed should be validated.

The soybean transcriptome was derived not from a single soybean seed but from a mixture of soybeans which were further divided into six developmental stages of seeds. The lengths of assembled contigs-associated with SMV in this study might be shorter than virus-associated contigs from a single plant due to the transcriptome containing several variants of SMV. Therefore, the assembled genome of SMV China is a consensus sequence of several SMV variants. Although the portion of SMV-associated sequences accounted for about 0.05% in the total transcriptome, the coverage of SMV genome in this study was about 414, and its coverage was also visualized by the alignment of raw data on the genome of SMV China. As a result, we could *de novo* assemble SMV genome based on enough sequence data associated with SMV.

Based on the assembled SMV genome, we could also identify SNVs for SMV. As we expected, we found several SNVs that resulted from a mixture of SMV infected diverse seed samples. However, we could not reveal the exact number of variants. Furthermore, the identification of SNVs in SMV demonstrated that not a specific region of SMV but several regions of SMV genome were highly mutated. The presence of several SMV variants in the soybean seeds is a very interesting finding, indicating that SMV is highly replicated in the developing seeds; this might be correlated with some disease symptoms in the soybean seeds caused by SMV. It might be of interest to examine replication rates of SMV in different developmental stages and tissues; this could provide evidence of the quasispecies nature of SMV in the near future.

Phylogenetic analyses suggested that the identified SMV isolate China was very different from other known SMV isolates based on polypeptide sequences. However, SMV isolate China seems to be highly correlated with two SMV isolates from South Korea, suggesting the phylogenetic correlation between geographical regions and SMV isolates.

Our SNV analysis in the soybean seeds indicates a high level of quasispecies nature for SMV. Mutations were not in a specific region but in most regions of SMV genome. Furthermore, we found that A-G and C-U conversions and vice and versa were frequent.

Taken together, our bioinformatics analyses using soybean seed transcriptomes identified five viruses infecting the soybean seeds. Of these five viruses, we *de novo* assembled the genome of SMV isolate China and analyzed SNVs revealing quasispecies nature of SMV in the soybean seeds for the first time. Our approaches and analyses in this study are valuable for the virus-associated studies using NGS-based transcriptome data.

Acknowledgments

This work was partially supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government Ministry of Education (No. NRF-2016R1D1A1A02937216); the Agenda Program, the Rural Development Administration (RDA) (No. PJ01194803); and the Vegetable Breeding Research Center (No. 710001-05) through the Agriculture Research Center program from the Ministry for Food, Agriculture and Rural Affairs, Republic of Korea. WKC was supported by a research fellowship from the Brain Korea 21 Plus Project.

References

Ali, A. and Kobayashi, M. 2010. Seed transmission of *Cucumber mosaic virus* in pepper. *J. Virol. Methods* 163:234-237.

Barba, M., Czosnek, H. and Hadidi, A. 2014. Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses* 6:106-136.

Burger, J. T. and Maree, H. J. 2015. Metagenomic next-generation sequencing of viruses infecting grapevines. *Methods Mol. Biol.* 1302:315-330.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G. and Durbin, R. 2011. The variant call format and vcftools. *Bioinformatics* 27:2156-2158.

Domier, L. L., Hobbs, H. A., McCoppin, N. K., Bowen, C. R., Steinlage, T. A., Chang, S., Wang, Y. and Hartman, G. L. 2011. Multiple loci condition seed transmission of *Soybean mosaic virus* (SMV) and smv-induced seed coat mottling in soybean. *Phytopathology* 101:750-756.

Edwards, R. A. and Rohwer, F. 2005. Viral metagenomics. *Nat. Rev. Microbiol.* 3:504-510.

Eggenberger, A. L., Stark, D. M. and Beachy, R. N. 1989. The nucleotide sequence of a soybean mosaic virus coat protein-coding region and its expression in *Escherichia coli*, *Agrobacterium tumefaciens* and tobacco callus. *J. Gen. Virol.*

70:1853-1860.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N. and Regev, A. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc.* 8:1494-1512.

Herridge, D. F., Peoples, M. B. and Boddey, R. M. 2008. Global inputs of biological nitrogen fixation in agricultural systems. *Plant Soil* 311:1-18.

Hill, J. H. and Whitham, S. A. 2014. Control of virus diseases in soybeans. *Adv. Virus Res.* 90:355-390.

Jafarpour, B., Shepherd, R. and Grogan, R. 1979. Serologic detection of bean common mosaic and lettuce mosaic viruses in seed. *Phytopathology* 69:1125-1129.

Jo, Y., Choi, H., Yoon, J.-Y., Choi, S.-K. and Cho, W. K. 2016. *In silico* identification of *Bell pepper endornavirus* from pepper transcriptomes and their phylogenetic and recombination analyses. *Gene* 575:712-717.

Leinonen, R., Sugawara, H. and Shumway, M. 2011. The sequence read archive. *Nucleic Acids Res.* 39:D19-D21.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. 2009. The sequence alignment/map format and samtools. *Bioinformatics* 25:2078-2079.

Madden, T. 2013. The BLAST sequence analysis tool. In: *The NCBI handbook* (2nd ed.), ed. by National Center for Biotechnology Information. Bethesda, MD, USA.

Massart, S., Olmos, A., Jijakli, H. and Candresse, T. 2014. Current impact and future directions of high throughput sequencing in plant virus diagnostics. *Virus Res.* 188:90-96.

Messina, M. J. 1999. Legumes and soybeans: overview of their nutritional profiles and health effects. *Am. J. Clin. Nutr.* 70:439S-450S.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11:31-46.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. 2010. Tablet-next generation sequence assembly visualization. *Bioinformatics* 26:401-402.

Morales, F. J. and Castano, M. 1987. Seed transmission characteristics of selected bean common mosaic virus strains in differential bean cultivars. *Plant Dis.* 71:51-53.

Morozova, O. and Marra, M. A. 2008. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 92:255-264.

Pimentel, D. and Patzek, T. W. 2005. Ethanol production using corn, switchgrass, and wood; biodiesel production using soybean and sunflower. *Nat. Resour. Res.* 14:65-76.

Ruiz, M., Simón, A., García, M. and Janssen, D. 2014. First report of *Lettuce chlorosis virus* infecting bean in spain. *Plant Dis.* 98:857.1.

Song, Q.-X., Li, Q.-T., Liu, Y.-F., Zhang, F.-X., Ma, B., Zhang, W.-K., Man, W.-Q., Du, W.-G., Wang, G.-D., Chen, S.-Y. and Zhang, J. S. 2013. Soybean GmbZIP123 gene enhances lipid content in the seeds of transgenic Arabidopsis plants. *J. Exp. Bot.* 64:4329-4341.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. 2013. Mega6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30:2725-2729.

Vodovar, N., Goic, B., Blanc, H. and Saleh, M.-C. 2011. *In silico* reconstruction of viral genomes from small rnas improves virus-derived small interfering rna profiling. *J. Virol.* 85:11016-11021.

Wang, Y. and Qian, P.-Y. 2009. Conservative fragments in bacterial 16s rRNA genes and primer design for 16s ribosomal DNA amplicons in metagenomic studies. *PLoS One* 4:e7401.

Wrather, J., Anderson, T., Arsyad, D., Tan, Y., Ploper, L., Porta-Puglia, A., Ram, H. and Yorinori, J. 2001. Soybean disease loss estimates for the top ten soybean-producing counries in 1998. *Can. J. Plant Pathol.* 23:115-121.

Wylie, S., Wilson, C., Jones, R. and Jones, M. 1993. A polymerase chain reaction assay for cucumber mosaic virus in lupin seeds. *Aust. J. Agr. Res.* 44:41-51.

Yanagisawa, H., Tomita, R., Katsu, K., Uehara, T., Atsumi, G., Tateda, C., Kobayashi, K. and Sekine, K.-T. 2016. Combined DECS analysis and next-generation sequencing enable efficient detection of novel plant RNA viruses. *Viruses* 8:70.

Yang, Y., Kim, K. S. and Anderson, E. J. 1997. Seed transmission of cucumber mosaic virus in spinach. *Phytopathology* 87:924-931.