Detection, discrimination and discovery of a new *Tobacco streak virus* strain.

M. Dutta[a], A. Ali[b], U. Melcher[a,c]

[a] Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, Oklahoma, 74078-3035 USA

[b] Department of Biological Science, The University of Tulsa, Oklahoma, 74104, USA.

[c]Corresponding author: Ulrich Melcher

Tel.: 405-744-6210

FAX: 405-744-7799

ulrich.melcher@okstate.edu

Department of Biochemistry and Molecular Biology, Oklahoma State University, 246 NRC, Stillwater, Oklahoma, 74078-3035 USA

Abstract:

Soybean plants that exhibited symptoms of virus infection were sampled from different counties of Oklahoma. These plants were tested serologically for 15 viruses known to infect soybean plants. Fifty-seven samples that exhibited typical virus-like symptoms did not test positive for any of the 15 viruses used in a dot-immunobinding assay (DIBA). Four samples were pooled and used for next generation sequencing using the 454- Roche protocol. Sequence and phylogenetic analysis of the sequences obtained revealed infection with a distinct strain of *Tobacco streak virus* (TSV). TSV was one of the 15 viruses initially tested for using DIBA and had tested negative. TSV belongs to the genus *Ilarvirus* and has been reported as a causal agent of bud blight in soybean crops in Brazil and the United States. Out of 10 reported primer pairs for TSV reverse transcription-polymerase chain reaction (RT-PCR), only two had the potential, based on sequence similarity, to amplify part of the genome of the distinct strain of TSV found in Oklahoma and only one was actually able to amplify the region. In this study, a new primer pair, specific to all known TSV and capable of amplifying the Oklahoma strain (TSV-OK), was designed from a highly conserved region of coat protein (CP) sequences and end-point PCR and quantitative RT-PCR detection methods were developed and their sensitivity assayed. This is the first report of primers designed from this specific highly conserved region in the CP of TSV for detection of TSV. Twenty-three of the 57 DIBA soybean samples that initially tested negative were retested with the new specific end-point PCR method and found positive for TSV infection.

Keywords: Microbial forensic methods; 454 sequencing; RT-qPCR; SYBR green

Footnote:

Abbreviations: *Tobacco streak virus* (TSV), *Soybean mosaic virus* (SMV), *Bean pod mottle virus* (BPMV), dot-immunobinding assay (DIBA), *Tobacco ringspot virus* (TRSV), *Tomato ringspot virus* (TomRSV), movement protein (MP), coat protein (CP), *Strawberry necrotic shock virus* (SNSV), *Parietaria mottle virus* (PMoV), *Blackberry chlorotic ringspot virus* (BCRV), electron microscopy (EM), Next-generation sequencing (NGS), RNA-dependent RNA polymerase (RdRp), untranslated region (UTR)

1. Introduction:

Study of diseases that may affect soybean is important to ensure good commercial crop yields. Mosaic, mottling, rugosity of leaves, mottling of seed coats, discoloration of seeds of soybean can be caused by strains of several viruses [*Tobacco streak virus* (TSV), *Soybean mosaic virus* (SMV), and *Bean pod mottle virus* (BPMV)] [1-3]. During a survey of soybean fields in Oklahoma a number of soybean plant samples were tested serologically by dot-immunobinding assay (DIBA) against the antisera of 15 different viruses including TSV (Ali et al. 2014 In preparation). More than 50% of the collected samples were negative to all 15 tested viruses. In this work, some of the previously collected DIBA negative samples (Ali et al. 2014, in preparation) were tested by NGS, which identified a novel strain of TSV that infects soybean plants in Oklahoma. TSV has been reported to cause bud blight in Brazil and the United States [4]. Bud blight causes pod necrosis and dieback of terminal portions of stems of a plant and has been observed in soybean in the past and in recent years. In the U.S., TSV was previously isolated from plants expressing symptoms of bud blight in commercial soybean fields in Wisconsin and Oklahoma [4]. Bud blight has several additional known causal agents, including *Tobacco ringspot virus* (TRSV), *Tomato ringspot virus* (TomRSV), and SMV [4]. Almost all soybean varieties tested have been susceptible [5,3]. Not much is known about the impact of TSV on soybean health and productivity and hence TSV infection merits further investigation. [4]

Viruses within genus *Ilarvirus* family *Bromoviridae*, of which TSV is the type member, have positive-sense, single-stranded RNA tripartite genomes (RNA 1 to 3). RNA 1 and

RNA 2 encode for the viral replication protein and RNA-dependent RNA polymerase (RdRp), respectively [6]. RNA 3 encodes the movement protein (MP) and a subgenomic RNA 4 that is translated into the coat protein (CP) [6].

Currently, there is no specified level of sequence similarity to distinguish among species, strains or subgroups of ilarvirus members [7]. Serological relatedness has been used to assign species to subgroups, and serological differences, host specificity or geographical location have been used to distinguish different species belonging to the same genus [8,9]. RNA 2 of members of subgroups 1 and 2 have a unique second open reading frame absent in other ilarvirus subgroups [6]. The type isolate of TSV was originally isolated from white clover (TSV-WC). Species within subgroup 1 include TSV, *Strawberry necrotic shock virus* (SNSV), *Parietaria mottle virus* (PMoV), *Blackberry chlorotic ringspot virus* (BCRV) and the proposed species "Bacopa chlorosis virus" [6].

TSV has a broad host range that includes members of *Solanaceae* and *Leguminosae*. TSV has been detected frequently amongst weeds bordering agricultural fields from where it serves as a source of inoculum that could infect crops. Thrips *Microcephalothrips abdominalis* and *Thrips tabaci* are thought to play roles in the transmission of TSV [10]. The mechanism of transmission is described mainly as mechanical in nature, where TSV-contaminated pollen grains are introduced into the plant through wounds [11-13].

In the diagnosis of novel, unidentified or unusual viral plant diseases, causative agents may go undetected when using methods such as PCR or ELISA, because these tests are very specific to a particular species or even strain of a virus. Also, given the fast rate of evolution of many RNA viruses, methods to be developed need to detect and identify

viral variants in a broader sense [14]. Methods such as electron microscopy (EM) or sap inoculation of plant virus indicator species do not allow accurate species level diagnosis. This research explores Next-generation sequencing (NGS) which offers an alternative solution where sequences are generated in a non-specific fashion and identification is made based on similarity alignment against GenBank data [13,15].

TSV is well known to be genetically diverse [9]. This diversity could contribute to failure of TSV identification by conventional methods like ELISA. In such cases, detection of genetically diverse viruses can be facilitated by NGS. Rapid diagnosis can be difficult in cases where a virus is not well recognized as infecting a particular plant host or geographical region and is further complicated where the virus differs from known strains. Currently, most methods in use for virus detection are species-specific [16] (e.g. PCR or ELISA). A few detection techniques have been developed for genera and families [11]. Also, sap mechanical inoculation and PCR with degenerate primers, can detect viruses in a species-specific manner but are limited by the fact that they can be applied to a relatively small range of potential viruses. NGS technology has proved to be a powerful, sensitive technology [13,17,18] especially suited to detecting unknown or unsuspected viruses.

1. Methodology:

2.1 Sample collection

Symptomatic leaf samples from 57 soybean plants, used in this study, were collected in a soybean survey in July 2012 made in Noble and Kay counties of Oklahoma USA and were negative by dot-immunobinding assay (DIBA) to 15 viruses known to infect

soybeans (Ali et al. 2014, in preparation). Soybean plants were at the young stage before flowering. Field symptoms of a typical virus-infected plant included mild mottling, rugosity and chlorotic lesions on the third tri-foliate compound leaf. However, all symptomatic plants tested negative for TSV by DIBA. To further investigate the cause of the symptoms and whether an unknown virus was causing the disease symptoms, NGS was performed.

2.2 454-sequencing

Total RNA was extracted from leaf samples using Tri Reagent (Molecular Research Center, Cincinnati, OH). The RNA concentrations of the samples were estimated by Nanodrop (Thermo scientific, Waltham, MA). Nanodrop concentrations ranged from 600ng/µl to 1700ng/µl. Four samples of RNA concentrations greater than 1 µg/µl were pooled for sequencing. These samples were chosen because of their relatively high RNA yields (> 1 µg/µl) and purity as evidenced by bright bands on gel electrophoretic separation on a 1.5% agarose gel.

The pooled RNA extract mixtures were used to synthesize double-stranded cDNA using the cDNA Synthesis System kit (Roche, Indianapolis, IN) following the manufacturer's protocols. A library was then produced from the RNA extract, using kits supplied by 454-Roche. Sequencing was performed following the manufacturer's protocols.

2.3 Sequence analysis

The sequence reads obtained were assembled using NEWBLER v. 2.7 (Roche). The resulting contigs and unassembled sequences were aligned to the NCBI nr GenBank

viruses database (taxid: 10239) using BLAST 2.7. Three contigs (00539, 00542 and 00545) were chosen for further sequence and phylogenetic analyses.

Open reading frame prediction and protein translation analyses were performed with ExPASY, a translating bioinformatics tool [19]. Sequence alignments and phylogenetic trees were constructed (using a maximum-likelihood method with 500 bootstrap replicates) using Clustal W and MEGA (v. 6) [20]. Phylogenetic trees were constructed using the amino acid sequences encoded by RNA 1, RNA 2 and RNA 3 of TSV isolates (Table 1) and by contigs 00539, 00542 and 00545. Methods of phylogenetic analyses used were neighbor-joining, maximum likelihood and maximum parsimony. Evolutionary distances (number of base substitutions per site) were computed using the Maximum Composite Likelihood method and were used to determine branch lengths on trees. Codon positions included were 1st + 2nd + 3rd + noncoding. Initial tree(s) for the heuristic searches were obtained automatically by maximum parsimony when the number of common sites was below 100 or less than one-fourth of the total number of sites, or by the BIONJ method with a MCL distance matrix.

Estimates of evolutionary divergence between sequences was carried out based on the pair-wise amino acid distances per site between contigs 00539 and isolates of TSV (Table1) that had produced BLAST 2.7 hits with high statistical significance (E-value cutoff 10^{-3}). Standard error estimate(s) were obtained by a bootstrap procedure (500 replicates). Analyses were conducted using the Dayhoff matrix based model [21,20].

Recombination analysis was made by RDP [22] with nucleotide sequence alignments of contigs 00539, 00542 and 00545 separately with the corresponding segments of other

TSV isolates (Table 1). When a recombination event was supported by at least three different algorithms out of the seven incorporated in the RDP software the recombination event was considered to be significant. These analyses were performed using the default settings from the different detection programs.

2.4 Primer design

Because of inadequacies of published primers, specific PCR primers were designed targeting the TSV coat protein gene: the sense primer TSV1789Fnd 5'-GCTATCGTCTGCAGCCTCGA-3' (1758-1777) and the antisense primer TSV1982Rnd, 5'-CCACATCGCACACAGGAATT-3' (1932-1951). The primers were designed based on the relatively conserved region (nt 1700-1970) of ilarvirus coat protein genes. Published degenerate primers TSV1789F and TSV1982R [11] were used as reference to design strain-specific primers. Web-interface applications Primer3, mFold, and BLASTn were used to predict thermodynamic parameters. Internal secondary structures and self-dimers were determined using mFold. Specificity was confirmed in silico by aligning the primer sequences using ClustalW to contig 00545. Primers were synthesized by IDT (Integrated DNA Technologies, Inc., Coralville, IA).

2.5 cDNA synthesis

First-strand cDNA synthesis was performed, using random hexamer primers on RNA extracted from symptomatic soybean leaves. SuperscriptIII reverse transcriptase (Invitrogen, Carlsbad, CA) was used according to the manufacturer's instructions for reverse transcription of 1 µg of RNA template to synthesize the first-strand cDNA.

2.6 Endpoint RT-PCR

Endpoint RT-PCR assays were carried out with reaction mixtures that contained 10 µl of GoTaq Green Master Mix (Promega), 1 µl (5 µM) each of primers TSV1789Fnd and TSV1982Rnd, 2 µl cDNA template, and 6 µl of nuclease-free water (Ambion, Austin,TX). Reactions were performed in an MJ Research thermal cycler (PTC-100, Ramsey, MN). The cycling parameters consisted of an initial denaturation at 95°C for 1 min, followed by 35 cycles of denaturation at 95°C for 20 s, annealing at 62°C for 20 s and 56°C for 10 s, extension at 72°C for 45 s and a final extension for 3 min at 72°C. Negative (non-template control; water) controls were included in each round of PCR amplification. Five microliters of the reaction volume was electrophoresed in a 1.5% agarose gel in 1X TAE buffer. The product size (193 bp) was determined using 1 Kb Plus DNA Ladder (Invitrogen). Bands were excised from the gel and the Qiaquick Gel Extraction Kit was used to recover the PCR product. The purified DNA fragments were sequenced from both directions by automated DNA sequencing using "BigDye™"-terminated reactions and Applied Biosystems 3730 Genetic Analyzer at the Recombinant DNA/Protein Resource Facility at Oklahoma State University.

2.7 Inclusivity and Exclusivity panel

Reference positive controls for *Soybean mosaic virus* (SMV), *Bean pod mottle virus* (BPMV), *Cucumber mosaic virus* (CMV) and *Tobacco mosaic virus* (TMV), (Agdia, Elkhart, IN) constituted the exclusivity panel for specificity assays. In the inclusivity panel, TSV samples from Agdia (Elkhart, IN) were used.

2.8 Plasmid positive control pMD-01

A plasmid positive control pMD-01 was generated for the new TSV-OK isolate by cloning a partial gene sequence (193 bp) of TSV coat protein amplified from first strand cDNA using primer set TSV1789Fnd and TSV1982Rnd. Resulting amplicons were excised and eluted from the agarose gel using a Qiagen gel extraction kit and cloned into a plasmid vector (pCR2.1-TOPO) using a TOPO-TA Cloning Kit (Invitrogen, Grand Island, NY). Plasmid DNA pMD-01 carrying target gene sequence was purified from overnight grown bacterial cultures using a QIAprep Spin Miniprep Kit (Qiagen) and was sequenced at the Oklahoma State University Recombinant DNA/Protein Resource Facility using primers TSV1789Fnd and TSV1982Rnd. The resultant sequence aligned perfectly (100% nucleotide identity) with the target gene and plasmid sequences.

2.9 Quantification of target cDNA with SYBRgreen RT-qPCR

The concentrations of plasmid pMD-01 were measured using a NanoDrop v.2000 spectrophotometer (Thermo Fisher Scientific, Inc., Worcester, MA). One ng/µl of pMD-01 was used to quantify the target pathogen genomic nucleic acid with RT-qPCR, using cDNA synthesized from infected plant RNA. The plasmid equivalent concentration obtained from the C_T value was multiplied by a factor of 2 to yield concentration of TSV viral genomes in the sample. RT-qPCR amplification was carried out in 20 µl reaction mixtures containing 10 µl of Platinum SYBR Green, qPCR SuperMix-UDG (Invitrogen), 0.5 µl (10 µM) each of primers TSV1789Fnd and TSV1982Rnd, 2 µl of template cDNA, and 6.4 µl of nuclease-free water. Positive and negative controls were included in each round of PCR amplification, and each reaction was performed in three replicates. The cycling parameters included two initial holds each for 2 min at 50 and 95°C, followed by 40 cycles of 95°C for 15 s and 60°C for 45 s.

2.10 Sensitivity assays

To determine the assay detection limits using primer set TSV1789Fnd and TSV1982Rnd, two sensitivity assays were performed. Target cDNA from TSV-infected soybean leaves made using TSV1982Rnd was quantified as described above and serially diluted in 10-fold increments, from 1 ng to 1 fg per reaction. A standard curve was generated using target TSV-OK cDNA and plasmid pMD-01 (plasmid pMD-01 also serially diluted in 10-fold increments and used at 1 ng to 1 fg per reaction). Each RT-qPCR was performed in three replicates. Assay sensitivity was also tested using primers and methods described in Pappu et al. [23] since these primers were successful in detecting TSV-OK.

3. Results:

3.1 Analysis of total nucleic acids and 454 sequencing

Pooled RNA from symptomatic, but DIBA negative, samples were sequenced by 454-sequencing and 238,946 reads were generated. The average length of reads was 480bp. The reads were assembled into 600 contigs and the raw reads and the contigs were aligned with the NCBI GenBank nr database using BLAST 2.7. Contigs 00539 (length=3431), 00542 (length=2857) and 00545 (length=2162) were found to have significant similarity to sequences of TSV RNA 1, RNA 2 and RNA 3, respectively. Hits to isolates of TSV provided 99% coverage, 89% identity and an E value of 0.0. Strawberry necrotic shock virus and other ilarvirus member sequences were among the hits detected but these had low query coverage (approximately 40%) and identity (approximately 60%). Other hits with viral genomes had poor statistical significance

based on E value, % identity and query coverage. Using the three contigs as queries in a BLASTn search of sequences accumulated during the Plant Virus Biodiversity and Ecology project [24] revealed, from a single plant specimen of *Coreopsis grandiflora* collected in 2005 from The Tallgrass Prairie Preserve (TPP) of Osage Co. Oklahoma, adjacent to Kay County, four sequence reads covering residues 459-579 of contig 00545 with 98.2% identity (Supplemental file), whereas identity values compared with RNA 3 sequences from other TSV strains were between 81 and 83%.

3.2 Phylogenetic analyses

The resulting hits of complete genomic segments from the BLAST analysis (E value cut off 10^{-3}) of contigs 00539, 00542 and 00545, were further analyzed. An analysis of the pair-wise amino acid distances between replicase sequences of pairs of TSV isolates (Table 2) showed that contig 00539 and TSV 1973 are more distant from other isolates than those isolates are from one another. Similar results were obtained with RdRp and CP sequences encoded by contigs 00542 and 00545 respectively (Tables S2 and S3). MEGA 6 analyses were based on the amino acid sequences of the polypeptides encoded by RNA 1 and RNA 2 and the CP ORF of RNA 3. In the case of RNA 1 (replicase), it was found that all isolates from India clustered together (Fig. 1A). Isolates from the USA (Illinois and Henry) clustered with TSV 2334 from eastern Australia. Contig 00539 and another eastern Australian isolate, TSV isolate 1973, formed a distinct clade, setting these two isolates apart from the rest of the TSV isolates. A similar pattern was seen in the trees built with the RdRp sequence of contig 00542 polypeptide (Figure S1) and with the CP of contig 00545 (Figure 1B).

3.3 Recombination analysis

For RNA 1, two separate recombination events were detected in contig 00539 (Table 3). One of these events was seen in the 5'UTR region. The major and minor parents were TSV Illinois and TSV 1973 from eastern Australia. Two other events were also detected in the RNA 2. One event was in TSV pumpkin from India (RdRp, 387-528) and the parental isolates involved were TSV 2334 (major parent) and contig 00542 (minor parent). The second recombination event was detected in TSV 2334 (RdRp, 2914-2978) and the isolates involved were TSV pumpkin and TSV Henry (Kentucky). Three other separate recombination events were detected in RNA 3. One event is located in the CP coding region of RNA 3 (nt 1513-2012). This recombination was found in isolates TSV Henry and TSV Illinois. The major and minor parents involved were TSV India and TSV (N), respectively. The second recombination event is located in the MP coding region of RNA 3 (nt 369-604). This recombination event was present in four Indian isolates: TSV Bangalore, TSV Okra, TSVAPGA and TSV India. The major and minor parents are TSV Henry and TSV (N). The third recombination was observed in the CP coding region of RNA 3 (nt 1294-1591). The major and minor parents involved were Contig 00545 and TSV 2334, respectively. All these instances of recombination are supported by at least 3 methods (Table 3) employed by RDP3 in detecting recombination and met the criteria to be considered true recombinants. The average P-values of each event for each detection program are shown in Table 3.

3.4 Detection using End point PCR and SYBR green qRT-PCR

Sequences of published TSV primer pairs previously used for detection [25,23,11,26,27] were aligned to contigs 00539, 00542 and 00545, respectively for RNA 1, RNA 2 and RNA 3. No primers matched perfectly. Each primer for contigs 00539 and 00542 had two or more mismatched positions with published primers. Hence, these contigs were not considered as targets for amplification. For contig 00545, there was one instance where there was just one mismatch and two instances where there were two mismatches (Table S1). Considering that alignments of published primers, with contig 00545 exhibited greater amplification potential than with contigs 00539 and 00542, contig 00545 was selected as the target gene for primer design.

Previously designed primers with less than three mismatches were used for amplification. Endpoint RT-PCR was carried out using primers and methods described in Ravi et al. [25], Sharman et al. [11] and Pappu et al. [23]. The degenerate primers TSV1789F and TSV1982R [11] had only one mismatch in the reverse primer (Table S1) but failed to amplify the product. Detection was successful only with primers and methods mentioned in Pappu et al. [23]. Primers designed by Pappu et al. [23] were able to amplify the target region in the cDNA and gave a 600 bp product but lacked sensitivity. Hence an endpoint RT-PCR method was developed with newly-designed strain-specific primers TSV1789Fnd and TSV1982Rnd.

The redesigned TSV specific primers TSV1789Fnd and TSV1982Rnd yielded sharp bright bands at 193 bp. Bands were excised and purified PCR products were sequenced. The resulting sequence had complete nucleotide identity to the target region of contig00545. Twenty-three of 57 DIBA TSV-negative samples were tested using primers TSV1789Fnd and TSV1982Rnd and were found to be positive for TSV by this test. The

samples were also tested with SYBR green qRT-PCR using primers TSV1789Fnd and TSV1982Rnd and found to be positive. All RT-PCR tests with the exclusivity panel tested negative and the RT-PCR test with the positive control (TSV from Agdia, Elkhart, IN) was positive (Figure S3).

3.5 Sensitivity assays

cDNAs synthesized from total RNA of virus-infected samples were quantified by generating a standard curve with known quantities of pMD-01. This step allows standardization of assay sensitivity when performed using plant samples in which the virus titer is unknown and/or possibly low. The target cDNA and plasmid pMD-01 were serially diluted in 10-fold increments from 1ng to 1 fg for each reaction. End point RT-PCR was able to detect virus titer as low as 1 fg in case of both cDNA and plasmid pMD-01 (Figure 2). A standard curve was generated in SYBR green qRT-PCR assay using target cDNA and plasmid pMD-01 serially diluted in 10 fold increments from 1ng to 1 fg. Primers TSV1789Fnd and TSV1982Rnd in SYBR green qRT-PCR assays were able to detect the target DNA up to 1 fg with both plasmid pMD-01 and cDNA (Figure 3). Each qRT-PCR assay was performed in three replicates. The R^2 values were 0.95 and 0.96 for plasmid pMD-01 and cDNA respectively (Table S4). Plasmid pMD-01 and cDNA had slopes of -2.2 and -2.1 respectively and they both had reaction efficiencies of 1.8. An end point PCR was also performed with primers and methods described in Pappu et al. [23] and these reactions were found to be sensitive up to 10 ng (Figure S2).

4. Discussion

Accurate diagnosis is necessary for effective management of disease. NGS can provide a useful adjunct to diagnosis of viral diseases of plants for several reasons. Multiple viruses transmitted by different vectors can cause similar symptoms and thus require different management strategies. Often, virus infections in plants can be asymptomatic or incipiently symptomatic. However, the potential to turn into severe outbreaks persist by developing obvious symptoms given the right factors and conditions (tolerance level of the plant, inability of the virus to adapt to the host, presence of vectors and other environmental factors). Virus evolution, especially of RNA viruses, is driven by mutations, recombination and reassortment [28] and can influence the molecular characteristics used as basis for their detection by methods like RT-PCR and ELISA. General antibodies can fail to detect some strains of a particular virus due to changes in antigenic determinants. In RT-PCR if the primers are not a good match to the template they may fail to amplify the target gene or, when the primers are not sensitive enough for diagnostic purposes when the viral titer is low. In these instances, NGS is an alternative rapid tool for accurate diagnosis.

Contigs 00539, 00542 and 00545 define a molecularly distinct TSV strain as shown by the pair-wise amino acid distances per site, phylogenetic and recombination analyses. In the phylogenetic tree of RNA 1 (Fig. 1), TSV 1973 and contig 00539 share a branch connecting them to the rest of the tree, even though a greater similarity between them than to other isolates is not evident in their pair-wise amino acid distance analysis (Table 2). This discrepancy could be the result of the presence of a recombination event between these two strains in the 5' UTR region and the RNA 1 helicase region (Table 3). We also detected the presence of recombination in the CP region between these two strains. In the

CP phylogenetic tree (Fig. 1), TSV 1973 and contig 00545 each form a long branch, and share a clade connecting them to the rest of the tree, even though their pair-wise amino acid distances per site do not indicate that TSV 1973 and contig 00545 have a greater similarity to one another than to other isolates (Table S3).

The fact that the TSV strain discovered in Oklahoma's soybean crop (or a close relative) was also found in a nature preserve, the Tallgrass Prairie Preserve, on which soybeans had never been grown suggests that the Oklahoma disease may arise from movement of the virus from wild plants into cultivated soybean fields or from the soybean plants to the Preserve plants [24]. Further biological characterization of TSV-OK is in progress.

While NGS allows detection of most organisms present in a sample, it suffers from time-consuming limitations during assembly and analysis of sequence data suggesting that NGS based virus discovery will only become more complicated with time. To circumvent these limitations, it has been shown that unique pathogen-specific sequences can be used in searches of unassembled, non-quality checked, sequence data to accurately detect pathogens in plant samples [29]. This process, termed E-probe Diagnostic Nucleic acid Analysis, treats the NGS sample sequences as the "reference" database that are queried with pathogen specific sequences termed E-probes. This process was very efficient at detecting known RNA and DNA viruses from infected plants [30]. However, the process only detects viruses for which the user directly queries using virus specific e-probes. We are testing the use of broad detection E-probes developed from probes designed for a virus microarray study (manuscript in preparation). These are general probes that allow detection of viruses at the family level. Preliminary results show efficient and accurate first line diagnostic testing can be done following this procedure.

Acknowledgements

This study was funded by the USDA-CSREES Plant Biosecurity Program (grant number 2010-85605-20542) and the Oklahoma Agricultural Experiment Station. H. Hwang and the Bioinformatics Core Facility are acknowledged for providing exceptional consistency and skill while preparing libraries and performing pyrosequencing. The authors thank the National Science Foundation (NSF-DEB 842703) and the Robert J. Sirny professorship for additional support. The authors also thank Kay Scheets and Francisco Ochoa-Corona for internal review.

References

1. Weiden MD, Ginsberg HS (1994) Deletion of the E4 region of the genome produces adenovirus DNA concatemers. Proc Natl Acad Sci Usa 91 (1):153-157

2. Fagbenle HH, Ford RE (1970) Tobacco-streak-virus isolated from soybeans, Glycine-max. Phytopathology 60 (5):814-&

3. Ghanekar AM, Schwenk FW (1974) Seed transmission and distribution of tobacco streak virus in 6 cultivars of soybeans. Phytopathology 64 (1):112-114

4. Rabedeaux PF, Gaska JM, Kurtzweil NC, Grau CR (2005) Seasonal progression and agronomic impact of Tobacco streak virus on soybean in Wisconsin. Plant Disease 89 (4):391-396. doi:10.1094/pd-89-0391

5. Wang Y, Hobbs HA, Hill CB, Domier LL, Hartman GL, Nelson RL (2005) Evaluation of ancestral lines of US soybean cultivars for resistance to four soybean viruses. Crop Science 45 (2):639-644

6. Pallas V, Aparicio F, Herranz MC, Sanchez-Navarro JA, Scott SW (2013) The Molecular Biology of Ilarviruses. In: Maramorosch K, Murphy FA (eds) Advances in Virus Research, Vol 87, vol 87. Advances in Virus Research. Elsevier Academic Press Inc, San Diego, pp 139-181. doi:10.1016/b978-0-12-407698-3.00005-3

7. Bujarski J, Figlerowicz M, Gallitlli D, Roossinck MJ, Scott SW (2012) Bromoviridae. In: King AMQ, Adams MJ, Carstens EB, Lefkowtiz EJ (eds) Virus Taxonomy, Ninth Report of the International Committe on Taxonomy of Viruses. Elsevier, Amsterdam, pp 965-976

8. Halk EL, Franke J (1983) Identification of serological types of Apple mosaic, Prunus necrotic ringspot and Tobacco streak viruses with monoclonal-antibodies. Phytopathology 73 (5):789-789

9. Clement DL, Converse RH (1986) Serological relationships among 4 tobacco streak virus isolates. Phytopathology 76 (8):842-842

10. Jones DR (2005) Plant viruses transmitted by thrips. European Journal of Plant Pathology 113 (2):119-157. doi:10.1007/s10658-005-2334-1

11. Sharman M, Thomas JE (2013) Genetic diversity of subgroup 1 ilarviruses from eastern Australia. Archives of Virology 158 (8):1637-1647. doi:10.1007/s00705-013-1628-4

12. Klose MJ, Sdoodee R, Teakle DS, Milne JR, Greber RS, Walter GH (1996) Transmission of three strains of tobacco streak ilarvirus by different thrips species using virus-infected pollen. Journal of Phytopathology-Phytopathologische Zeitschrift 144 (6):281-284. doi:10.1111/j.1439-0434.1996.tb01530.x

13. Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M, Boonham N (2009) Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. Molecular Plant Pathology 10 (4):537-545. doi:10.1111/j.1364-3703.2009.00545.x

14. Zheng L, Wayper PJ, Gibbs AJ, Fourment M, Rodoni BC, Gibbs MJ (2008) Accumulating variation at conserved sites in potyvirus genomes is driven by species discovery and affects degenerate primer design. PLoS ONE 3 (2):e1586. doi:10.1371/journal.pone.0001586

15. Ho T, Tzanetakis IE (2014) Development of a virus detection and discovery pipeline using next generation sequencing. Virology 471–473 (0):54-60. doi:http://dx.doi.org/10.1016/j.virol.2014.09.019

16. Arif M, Aguilar-Moreno GS, Wayadande A, Fletcher J, Ochoa-Corona FM (2014) Primer Modification Improves Rapid and Sensitive In Vitro and Field-Deployable Assays for Detection of High Plains Virus Variants. Applied and Environmental Microbiology 80 (1):320-327. doi:10.1128/aem.02340-13

17. Al Rwahnih M, Daubert S, Golino D, Rowhani A (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus

infection that includes a novel virus. Virology 387 (2):395-401. doi:10.1016/j.virol.2009.02.028

18. Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I, Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: a generic method for diagnosis, discovery and sequencing of viruses. Virology 388 (1):1-7

19. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31 (13):3784-3788

20. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution 30 (12):2725-2729. doi:10.1093/molbev/mst197

21. Schwarz R. DM (1979) Matrices for detecting distant relationships. Atlas of protein sequences: 353-358

22. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P (2010) RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26 (19):2462-2463. doi:10.1093/bioinformatics/btq467

23. Pappu HR, Hammett KRW, Druffel KL (2008) Dahlia mosaic virus and Tobacco streak virus in dahlia (Dahlia variabilis) in New Zealand. Plant Disease 92 (7):1138-1138. doi:10.1094/pdis-92-7-1138b

24. Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U (2006) Plant virus biodiversity and ecology. PLoS Biol 4 (3):e80. doi:10.1371/journal.pbio.0040080

25. Kumar NA, Narasu ML, Zehr UB, Ravi KS (2008) Molecular characterization of Tobacco streak virus causing soybean necrosis in India. Indian Journal of Biotechnology 7 (2):214-217

26. Bhat AI, Jain RK, Ramiah M (2002) Detection of Tobacco streak virus from sunflower and other crops by reverse transcription polymerase chain reaction. Indian Phytopathology 55 (2):216-218

27. Hosseini S, Habibi MK, Mosahebi G, Motamedi M, Winter S (2012) First report on the occurrence of Tobacco streak virus in sunflower in Iran. Journal of Plant Pathology 94 (3):585-589

28. Ali A, Roossinck MJ (2010) Genetic bottlenecks during systemic movement of Cucumber mosaic virus vary in different host plants. Virology 404 (2):279-283. doi:http://dx.doi.org/10.1016/j.virol.2010.05.017

29. Stobbe AH, Schneider WL, Hoyt PR, Melcher U (2014) Screening metagenomic data for viruses using the e-probe diagnostic nucleic acid Assay (EDNA). *Phytopathology*, **104**, 1125-1129

30. Stobbe AH, Daniels J, Espindola AS, Verma R, Melcher U, Ochoa-Corona F, Garzon C, Fletcher J, Schneider W (2013) E-probe Diagnostic Nucleic acid Analysis (EDNA): A

1 theoretical approach for handling of next generation sequencing data for diagnostics.
2 Journal of Microbiological Methods 94 (3):356-366. doi:10.1016/j.mimet.2013.07.002

3

Table 1: Isolate identity, geographical locations and accession numbers of the RNAs used for analyses in this article.

TSV isolate	Geographical location	Accession numbers		
		RNA 1	**RNA2**	**RNA3**
Tobacco streak virus OK	Oklahoma, USA	KP256522	KP256521	KP256520
Tobacco streak virus isolate 1973	Eastern Australia	JX463334.1	JX463335.1	JX463336.1
Tobacco streak virus isolate pumpkin	Karnataka, India	FJ561299.1	FJ561300.1	
Tobacco streak virus isolate okra	Karnataka, India	FJ561302.1		FJ561304.1
Tobacco streak virus isolate REPDHP	Tamil Nadu, India	KF264473.1		
Tobacco streak virus isolate REPKAR	Tamil Nadu, South India	KF264474.1		
Tobacco streak virus isolate REPCBE2	Tamil Nadu, South India	KF264472.1		
Tobacco streak virus isolate Henry	Kentucky, USA	JX073656.1	JX073657.1	JX073658.1
Tobacco streak virus isolate 2334	Eastern Australia	JX463337.1	JX463338.1	JX463339.1
Tobacco streak virus isolate Illinois	Illinois, USA	FJ403375.1	FJ403376.1	FJ403377.1
Tobacco streak virus from India	Karnataka, India			DQ067449.1
Tobacco streak virus isolate TSVAPGA	Andhra Pradesh, India			FJ417082.1
Tobacco streak virus	Karnataka, India			FJ655171.1

Tobacco streak virus	Karnataka, India			FJ655170.1
Tobacco streak virus (N)	Netherlands			X00435.1
Tobacco streak virus isolate 1025	Eastern Australia			JX463347.1

1

2

Table 2: Pair-wise amino acid distances in RNA 1 of selected TSV strains. The amino acid distances per site between sequences is shown in the lower left triangle with standard error estimates in the upper right triangle. Analyses were conducted using the Dayhoff matrix based model with 500 bootstrap replicates.

	Contig 00539	**TSV Illinois**	**TSV 2334**	**TSV Henry**	**TSV REPCBE 2**	**TSV REPKAR**	**TSV REPDHP**	**TSV Okra**	**TSV Pumpkin**	**TSV 1973**
Contig00539		0.008	0.009	0.009	0.009	0.009	0.009	0.009	0.010	0.011
TSV Illinois	**0.082**		0.006	0.006	0.007	0.007	0.007	0.007	0.008	0.010
TSV 2334	**0.083**	0.042		0.006	0.007	0.007	0.007	0.007	0.007	0.010
TSV Henry	**0.083**	0.043	0.045		0.008	0.008	0.008	0.008	0.008	0.011
TSV REPCBE2	**0.092**	0.068	0.061	0.069		0.001	0.001	0.002	0.003	0.011
TSV REPKAR	**0.093**	0.069	0.062	0.070	0.001		0.002	0.002	0.003	0.011
TSV REPDHP	**0.093**	0.068	0.062	0.070	0.003	0.003		0.002	0.003	0.011
TSV Okra	**0.094**	0.069	0.063	0.071	0.005	0.006	0.006		0.004	0.011
TSV Pumpkin	**0.099**	0.074	0.068	0.075	0.010	0.010	0.010	0.013		0.011
TSV 1973	**0.126**	**0.120**	**0.119**	**0.123**	**0.139**	**0.140**	**0.140**	**0.141**	**0.146**	

Table 3: Analysis of recombination of RNA sequences of TSV isolates

Virus isolate	**Genomic region**	**Event**	**Major/minor parents**	**RDP analysis**	**Average P- value**
Consensus 00539	RNA 1	1	TSV Illinois/TSV 1973	RDP Geneconv Bootscan Maxchi Chimaera	$3.27x10^{-7}$ $5.56x10^{-5}$ $6.02x10^{-6}$ $1.75x10^{-3}$ $7.00x10^{-5}$
Consensus 00539	RNA 1 (3037-3168)	1	TSV Illinois/TSV 1973	RDP Maxchi Chimaera	$5.43x10^{-3}$ $2.22x10^{-3}$ $6.99x10^{-3}$
TSV pumpkin	RNA 2 (387-528)	1	TSV 2334/ Consensus 00542	RDP Maxchi Chimaera	$5.19x10^{-3}$ $2.33x10^{-3}$ $6.01x10^{-3}$
TSV 2334	RNA 2 (2914-2978)	1	TSV pumpkin/TSV Henry	RDP Geneconv Bootscan Maxchi Chimaera	$3.70x10^{-5}$ $8.89x10^{-8}$ $1.59x10^{-8}$ $1.03x10^{-4}$ $2.01x10^{-2}$
TSV Henry TSV Illinois	RNA 3 (1513-2012)	2	TSV India/TSV (N)	RDP Bootscan Maxchi	$1.20x10^{-2}$ $4.43x10^{-2}$ $6.25x10^{-3}$
TSV Bangalore TSV Okra TSVAPGA TSV India	RNA 3 (369-604)	4	TSV Henry/ TSV (N)	Maxchi Chimaera SiScan	$8.78x10^{-3}$ $2.24x10^{-2}$ $2.17x10^{-2}$
TSV 1973	RNA 3 (1294-1591)	1	Consensus 00545/TSV 2334	RDP Bootscan Maxchi Chimaera SiScan	$2.86x10^{-4}$ $2.28x10^{-2}$ $7.32x10^{-7}$ $2.38x10^{-5}$ $4.28x10^{-11}$

Figure Legends

Fig 1: Phylogenetic relationships among TSV strains: (A) Amino acid sequences encoded by contig 00539 and different isolates of TSV with statistically significant relatedness; (B) Predicted CP sequences encoded by contig 00545 and other isolates of TSV with statistically significant relatedness. Trees shown were inferred by using the maximum-likelihood method of MEGA6 based on the Tamura-Nei model. Trees with the highest log likelihood (-16662.1875) are shown. Branches corresponding to partitions reproduced in more than 50 % of the 500 bootstrap replicates are numbered. Abbreviations used are as follows: TSV (Tobacco streak virus) followed by the name of the isolate.

Fig 2: Endpoint RT-PCR sensitivity assays. Target cDNA (A) from TSV infected soybean leaves and plasmid pMD-01 (B) carrying the target gene fragment were serially diluted in 10-fold increments and used at 1 ng to 1 fg per reaction. Lane L, 1-kb ladder (Invitrogen). Lanes 1 to 7 contain serially diluted (10-fold increments) cDNA or plasmid pMD-01 at concentrations from 1 ng to 1 fg.

Fig 3: Sensitivity assays using real-time SYBR green RT-qPCR for TSV. Target cDNA (A) from TSV-infected soybean leaves and plasmid pMD-01 (B) carrying the target gene fragment was serially diluted in 10-fold increments and used at 1 ng to 1 fg per reaction. R^2 is linear correlation, M is the slope, and Ex is reaction efficiency. Each reaction was performed in three replicates. The x axes show the concentrations used, and the y axes show the observed cycle thresholds.

Figure 1A

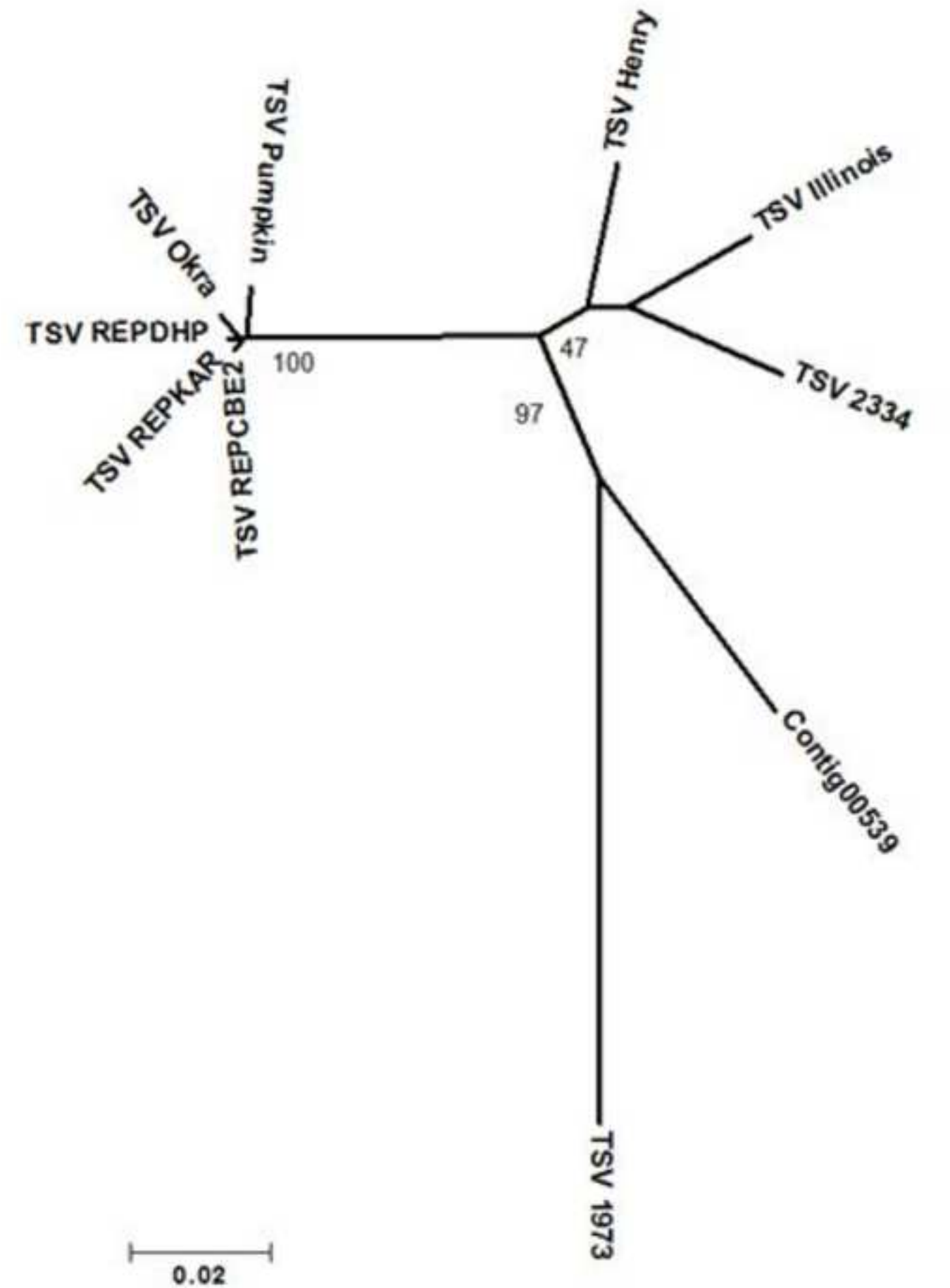

TSV Henry
TSV Illinois
TSV 2334
Contig00539
TSV 1973
TSV Pumpkin
TSV Okra
TSV REPDHP
TSV REPKAR
TSV REPCBE2
100
47
97
0.02

Figure 1B

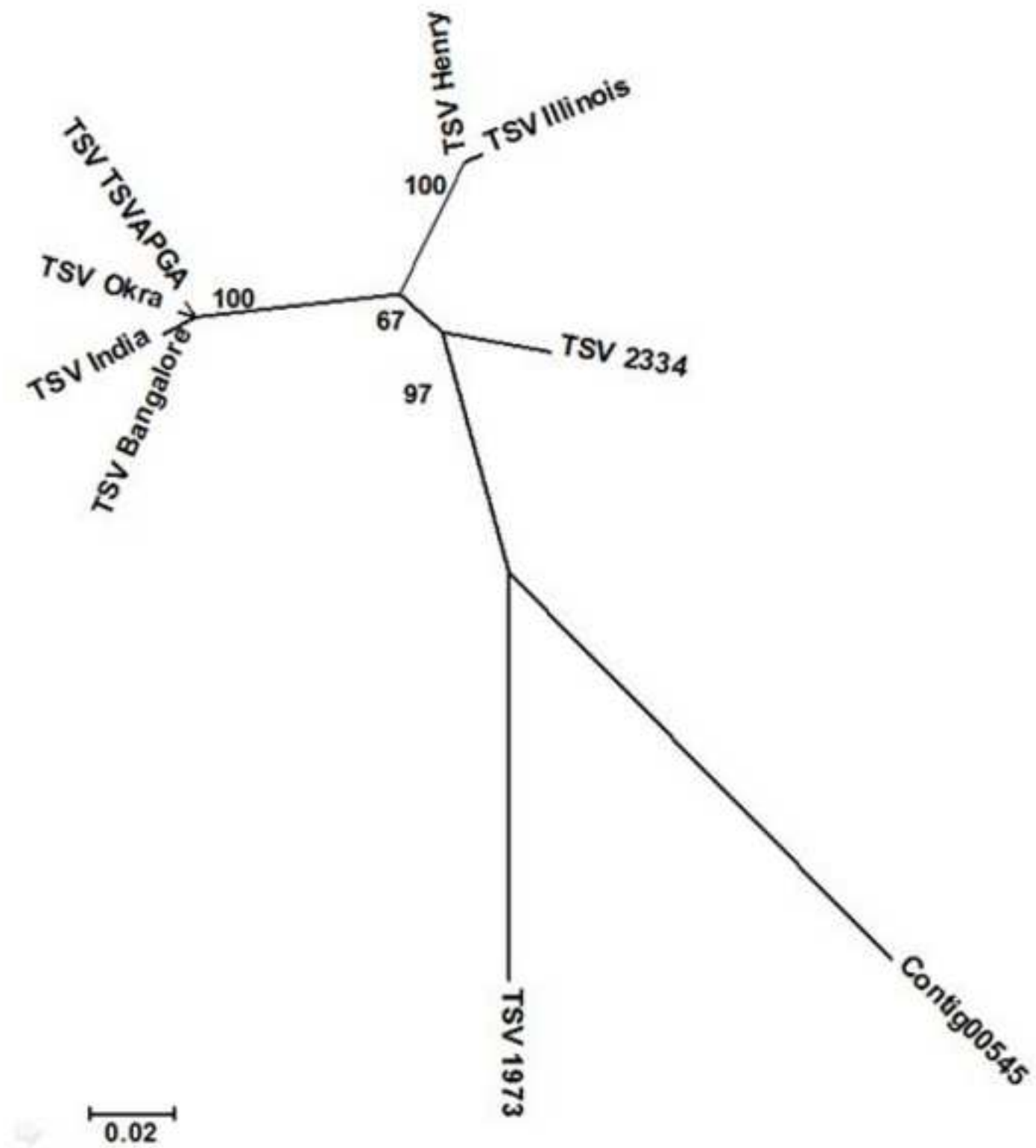

TSV Henry
TSV Illinois
TSV TSVAPGA
TSV Okra
TSV India
TSV Bangalore
100
100
67
97
TSV 2334
TSV 1973
Contig00545
0.02

Figure 2A

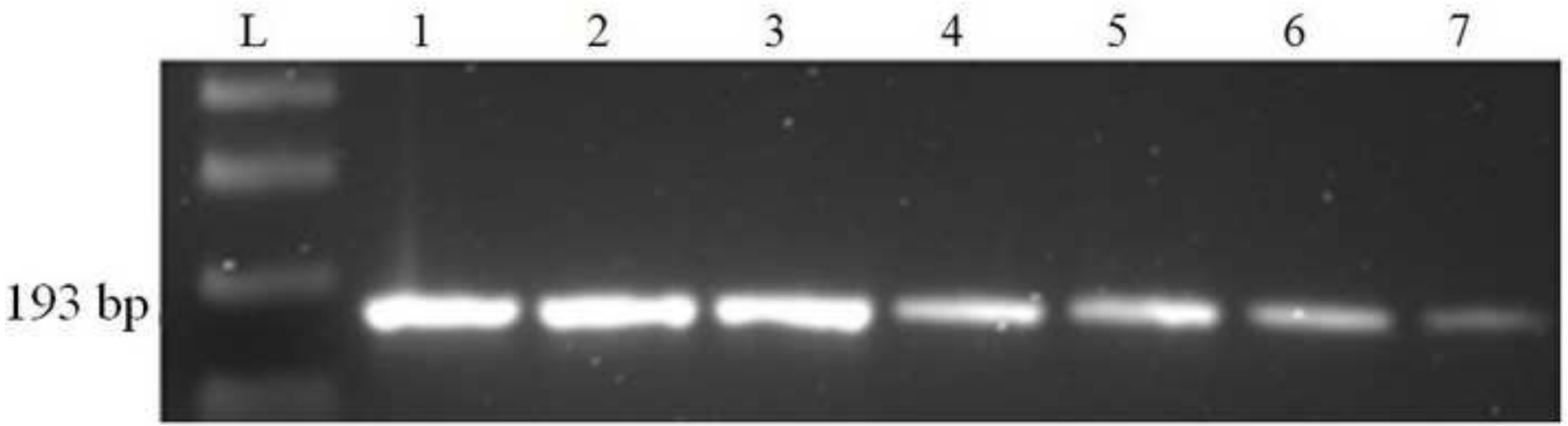


Figure 2B

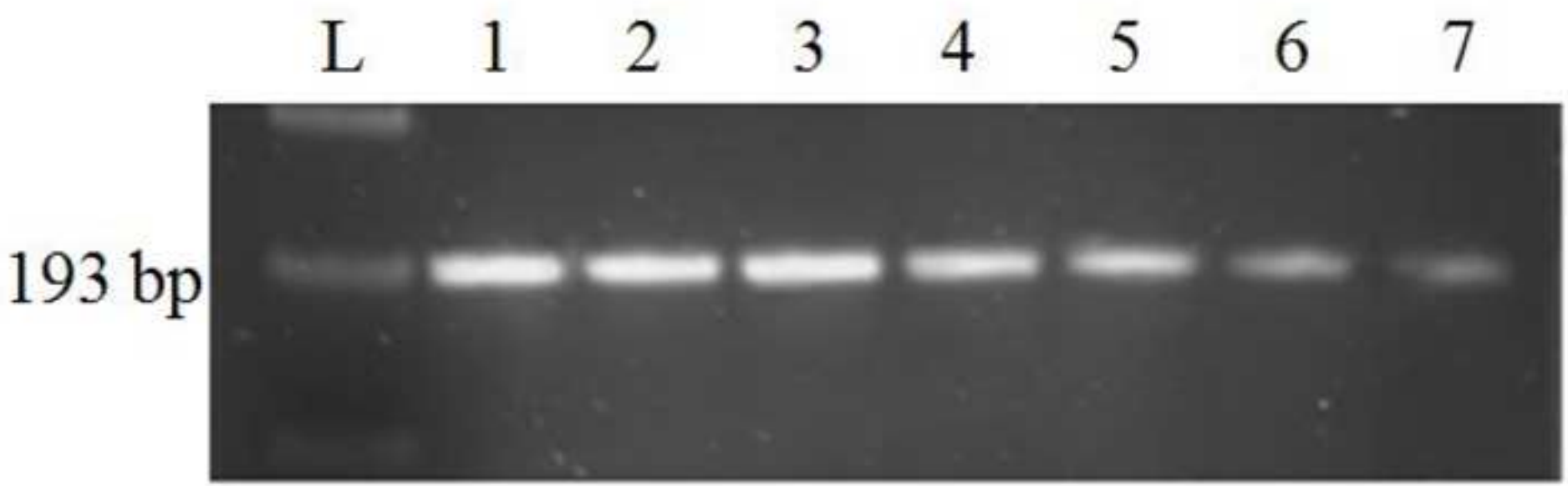


Figure 3A

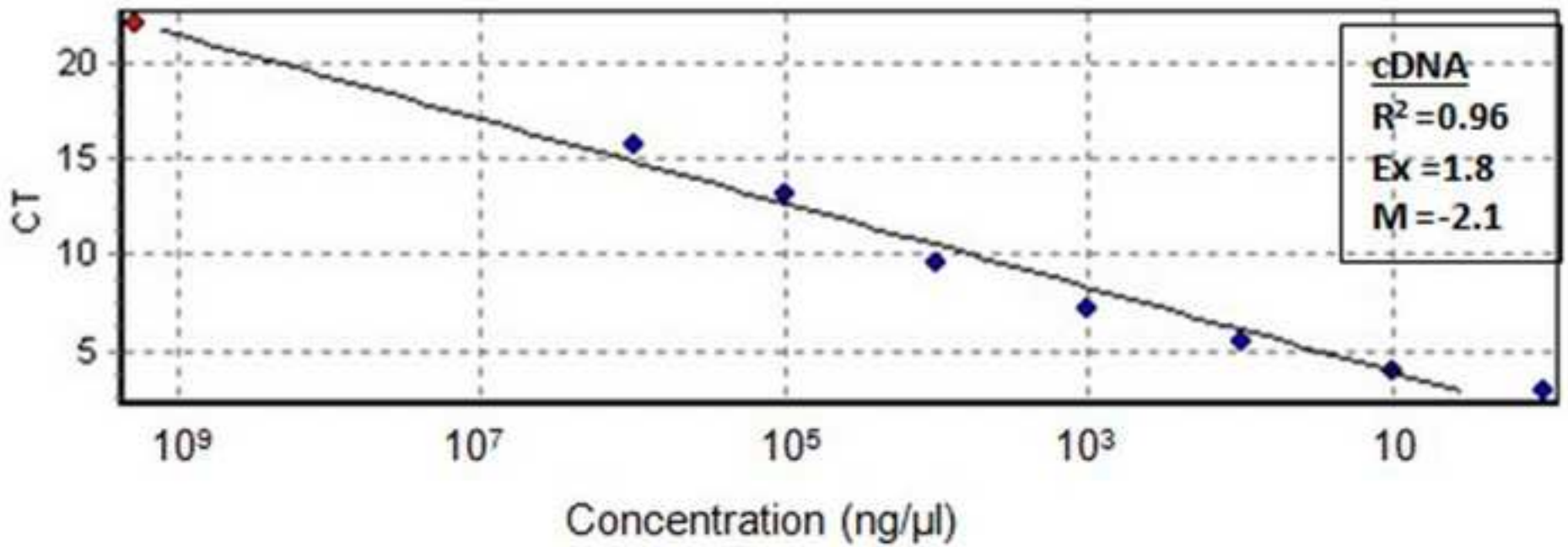
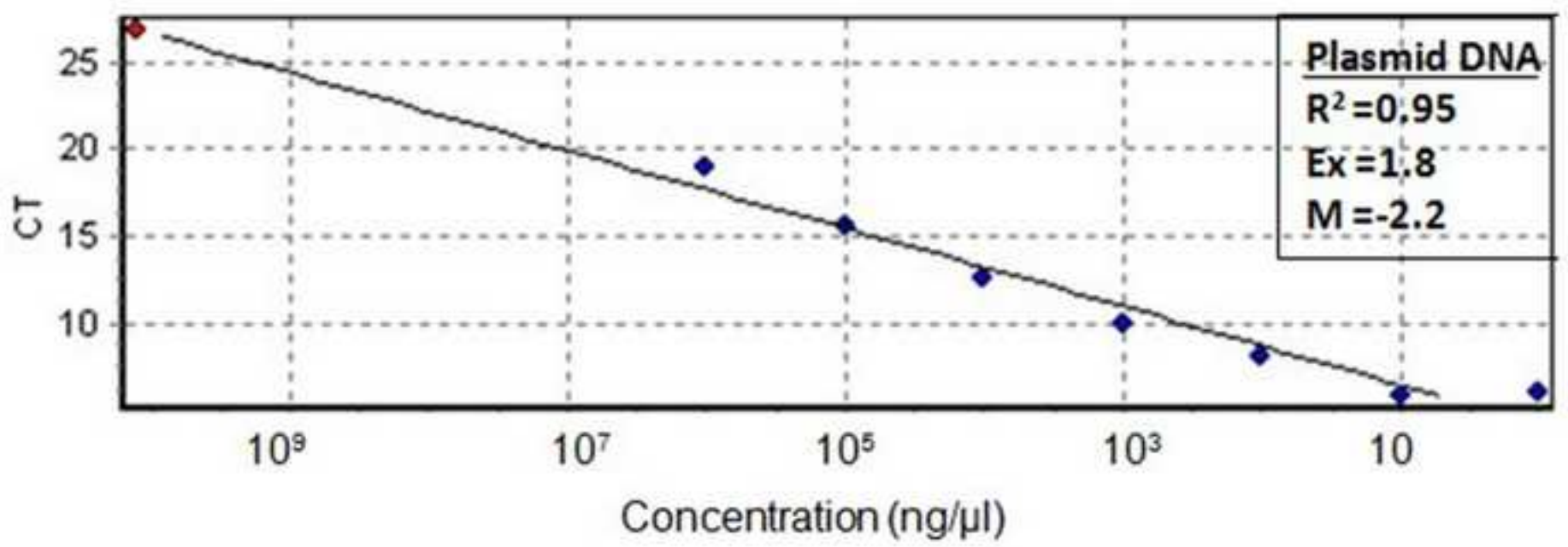

cDNA
R2 =0.96
Ex =1.8
M =-2.1
CT
20
15
10
5
10⁹
10⁷
10⁵
10³
10
Concentration (ng/µl)

Figure 3B


Plasmid DNA
R² =0.95
Ex =1.8
M =-2.2
CT
25
20
15
10
10⁹
10⁷
10⁵
10³
10
Concentration (ng/µl)