Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

ENDOGENOUS PARARETROVIRUSES IN THE GENOME OF CUCUMBER ARE SILENT AND INVOLVED WITHIN PLANT GENES

HAWRAA ALYASIRY
Plant Protection Department - Faculty of Agriculture - University of Kufa - Najaf – Iraq.

OSAMAH NADHIM ALISAWI
Plant Protection Department - Faculty of Agriculture - University of Kufa - Najaf – Iraq.

FADHL ALFADHL
Plant Protection Department - Faculty of Agriculture - University of Kufa - Najaf – Iraq.

Abstract

Cucumber (Cucumis sativus) is widely planted and hosted many plant viruses in Iraq. However, endogenous papraretroviruses (EPRVs) have not been studied in details in the genome of cucumber. In this study, next generation sequencing (NGS) and bioinformatics techniques were used to diagnose endogenous Pararetroviruses in the cucumber genome. The illumine platform produced 89.893.674 and 53.458.772 short clean reads with a length of 151 and 101 bases for total DNA and RNA data respectively. The results from Repeat Explorer and mapping to reference revealed that two viral elements belong to Florendovirus genus found in the examined genome, named CsatAV-Iraq and CsatBV -Iraq. The full length of CsatAV-Iraq and CsatBV-Iraq were **7147 and** 7390 bp respectively. The CsatAV-Iraq encodes seven coding domains; MP, PTZ00440, antiphage_ZorA_4, two RT_LTRs, RT_RNaseH_2, RNase_HI. The CsatBV-Iraq has seven domains of DUF5864, reverse transcriptase (RT_LTR and RVT_1), three domains of RNaseH (RH), and RVT_2. Some plant genes were found to be involved in the two Florendo viruses, which predicted their integration and position in the genome. Genome proportion and copy number values showed that CsatBV-Iraq is more abundant than CsatAV-Iraq. The sequences were existed in RNA transcripts with low expression level. The phylogeny revealed the close relationship between the two EPRVs with their related ones.

Keywords: CsatAV-Iraq, CsatBV-Iraq, Next generation sequencing, Bioinformatics

1. INTRODUCTION

Cucumber (Cucumis sativus L.) is an edible vegetable plant that belongs to the Cucurbitaceae family (Eifediyi and Remison, 2010). With 7 chromosome pairs, cucumber has a genome size of 367 Mbp (Arumanagathan and Earle 1991). It is also used as a primary model system for determining sex in addition to displaying a variety of sexual expressions. Moreover, cucurbits have a highly conserved genome, making them excellent plants to study vascular biology, since xylem and phloem saps can be collected for long-distance signaling studies. Despite cucurbits' agricultural and biological importance, very little is known about their genetics and genome. Therefore, the genome was sequenced and assembled especially for the domestic cucumber (Lough and Lucas 2006; Xoconostle-Cázares et al., 1999; Huang et al., 2009). About 20 plant viruses have been registered to infect cucumber such as Cucumber mosaic virus (CMV), Zucchini yellows mosaic virus (ZYMV), and Watermelon mosaic virus 2 (WMV2) (Bos 2000). The Next Generation Sequencing (NGS) method can identify the sequences of millions of DNA fragments, and it can also analyze several genes at once.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

The use of NGS has evolved as any new technology has over the years (Yohe et al., 2017). With the advent of next-generation sequencing (NGS), it became possible to know the full nucleotide sequence of the host plant, and viruses arose on the host plant under stress conditions without external infections, leading to the discovery of endogenous Pararetroviruses (EPRVs) in plants (Alisawi, 2019). Among the 1.6 billion year old viruses, endogenous Pararetroviruses (EPRVs) belong to the Caulimoviridae family, which are retroviral viruses incorporated into the genomes of plants (Schmidt et al., 2021, Richert-Pöggeler et al.,2021). Some of these viruses are controlled by the plant genome and are not able to cause infection in plants. Hence, they remain dormant within the host genome and are passed through generations, while the process is known as epigenetic inheritance (Schmidt et al. 2021). The high pressure of stresses breaks the dormancy of such viruses and reactivate these sequences through reverse transcription and turn from DNA to mRNA and then to DNA again and form complete viral particles are often associated with devastating diseases (Geering et al. 2014; Alisawi, 2019). Recently, two endogenous elements belong to the genus of Florendovirus have been detected in the cucumber genome (Geering et al., 2014). However, endogenous papraretroviruses (EPRVs) have not been studied in details in the genome of cucumber. In order to extend the search of such units in the genome of cucumber, we conducted a study to identify and measure the effectiveness of endogenous viruses within cucumber genome.

2. PROCEDURE

2.1. Plant material and DNA sequencing

Two samples were collected from one leaf of cucumber for DNA and RNA extraction. In an Eppendorf tube, samples were immersed in RNA-Later solution and sent to DNA-Link Company, Republic of Korea. The company's instructions were followed for extracting DNA and RNA. TruSeq DNA Library prep kit and TruSeq whole RNA library prep kit were used to prepare NGS libraries for DNA and RNA sequencing. DNA samples were examined according to the manufacturer's instructions using the Novaseq6000 2x150bp reads technique with the WGS application (PCR Free550). A NovaSeq6000, 2x101PE was used to sequence the total RNA after determining the quality of the RNA sample with an Agilent 2100 Expert Bio analyzer.

2.2. Graph-based read clustering with Repeat-Explorer

In last-generation sequencing data, the Repeat-Explorer pipeline was used to explore and characterize EPRV clusters and repetitive DNA sequences (Novák et al., 2013). Repeat-Explorer2 (Galaxy 2.3.8.1) was applied to cluster, and viridplantae version 3.0 was selected for taxonomic and protein domain identification. Afterwards, the clusters were imported into the Repbase dataset (Jurka et al., 2005) and the Basic Local Alignment Search Tool for further identification (Altschul et al., 1990). The sequences of these viruses were aligned to previously published EPRVs, and genus-level identification of the suggested EPRVs was achieved.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

2.3. Map to reference

We used two EPRVs as references that identified earlier by Geering et al., (2014) CsatAV and CsatBV, and belong to the genus Florendovirus. Mapping of raw Illumina reads against the identified viruses produced a report that showed the number of assembled reads out of total reads used. Based on the data, we calculated copy numbers and genome proportions as follows: 1- Copy number = number of assembled reads X read length / reference sequence length. The genome proportion is calculated by multiplying the number of aligned reads by the total number of NGS reads X 100 (Mustafa et al., 2018).

2.4. Phylogenetic analysis

In order to create a robust phylogeny model, the maximum likelihood (ML) method was applied to MEGA 11 (Tamura et al., 2013). Clustal W alignment was used to align the sequences (about 7000 bp for each). General Time Reversible (GTR) was used to reconstruct the tree. 35 EPRVs were applied to build the tree of the Florendovirus group.

3. RESULTS

As a result of the Illumina platform production, about 89.893.674 short clean reads with a length of 151 bases were generated from DNA data. In contrast, there were about 53.458.772 short clean reads with a length of 101 bases in the RNAseq data. Repeat-Explorer revealed that the cucumber genome contains one cluster of endogenous Pararetroviruses representing the Florendovirus genus. The contigs of this cluster represented two Florendo viruses that were identified earlier by Geering et al. (2014), CsatAV and CsatBV mapped against the whole DNA reads. According to the analysis, 23,476 reads were assembled against CsatAV (Fig 1), while 77,938 reads were assembled against CsatBV (Fig 2), to form the full sequence of each EPRV with possible coded domains. The two virus elements were named CsatAV –Iraq (https://www.girinst.org/2022/vol22/issue4/CsatAV-Iraq.html), and CsatBV –Iraq (https://www.girinst.org/2022/vol22/issue4/CsatBV-Iraq.html), following Geering et al. (2014), and the deposited in the Repbase dataset. The full length of CsatAV –Iraq and CsatBV –Iraq were 7147 and 7390 bp respectively. The CsatAV –Iraq encodes seven coding regions; viral movement protein (MP), antiphage_ZorA_4 (anti-phage defence ZorAB system protein ZorA) and PTZ00440 (reticulocyte binding protein 2-like protein), two domains of RT_LTR, and also two domains of RT_RNase_HI (RNase_HI_RT_Ty3/Gypsy family of RNase HI in long-term repeat retroelements) (Fig 3). The CsatBV –Iraq encodes DUF5864, RT_LTR, RVT_1, RT_RNaseH_2, RNase_H1, RNase_H1 and RVT_2 (Fig 4). The genome proportions for CsatAV-Iraq and CsatBV-Iraq were 0,02 and 0,086, while CsatAV-Iraq and CsatBV-Iraq had copy numbers of 495.9 and 1592, respectively.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

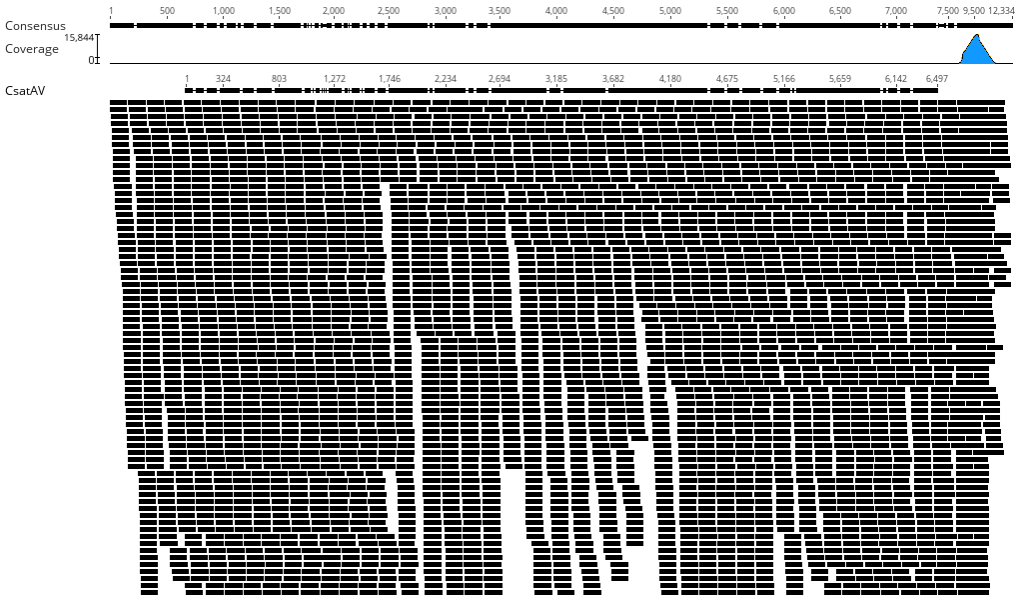

Fig1: The assembled reads of total DNA with complete coverage against CsatAV to produce the entire consensus sequence of CsatAV-Iraq

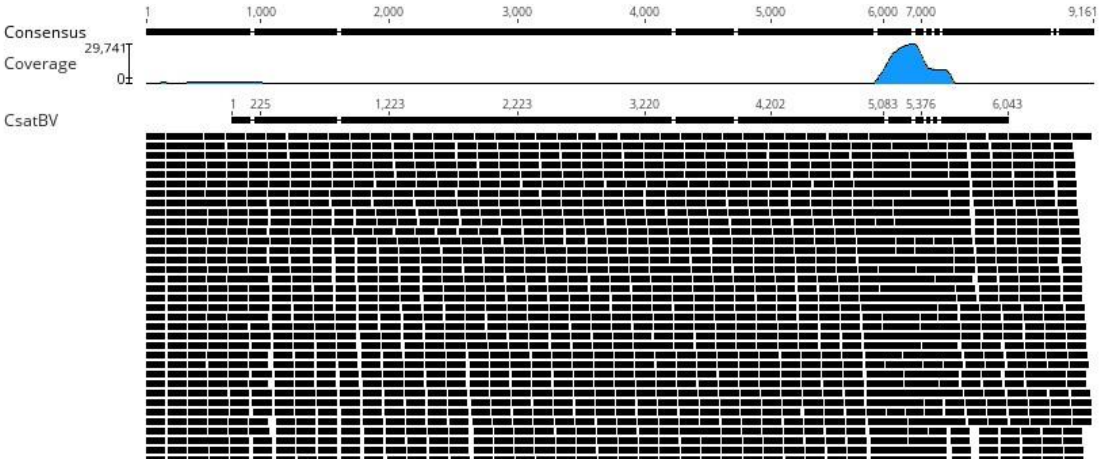

Fig 2: The assembled reads of total DNA with full coverage against CsatBV to produce the complete consensus sequence of CsatBV-Iraq

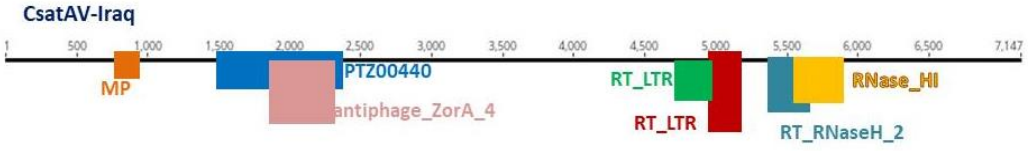

Fig 3: The entire sequence of CsatAV –Iraq shows seven protein domains, including two plant genes

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

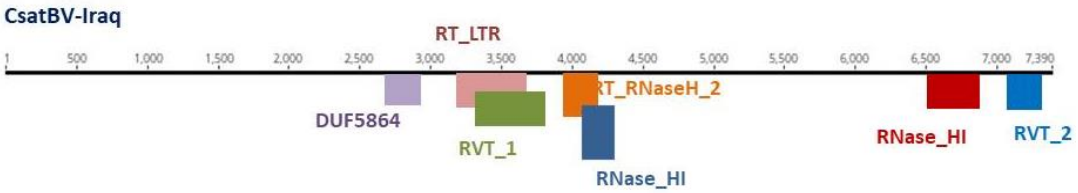

Fig 4: The entire sequence of CsatBV –Iraq shows seven protein domains with one plant gene involved

Both sequences were existed in RNA transcripts with very weak level as only one read assembled against of CsatAV–Iraq (Fig 5) and 1,337 reads mapped to CsatBV –Iraq (Fig 6). The phylogeny revealed the close relationship between the two EPRVs with their related ones (Fig 7).

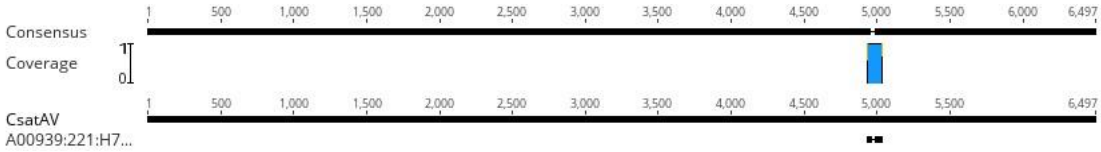

Fig 5: The mapping to reference tool shows only one read aligned to CsatAV-Iraq

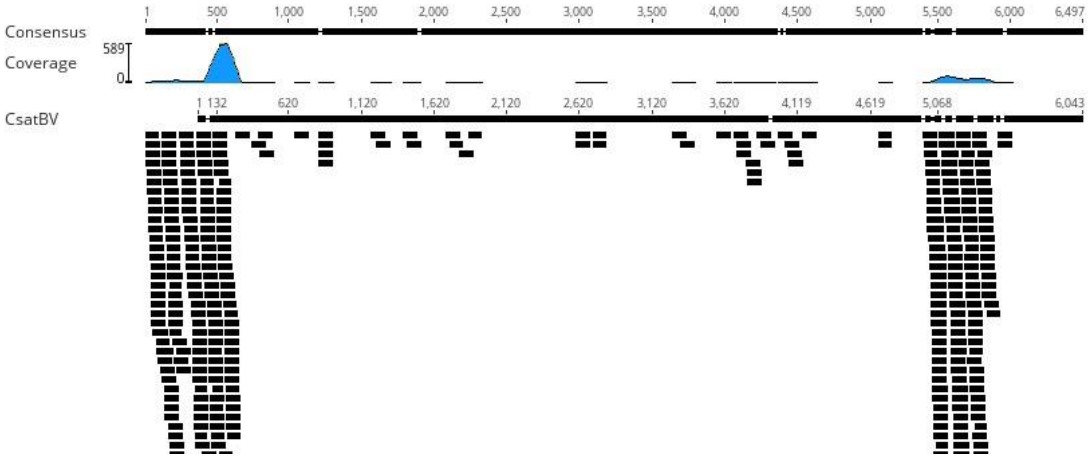

Fig 6: The mapping to reference tool shows dispersed RNA reads assembled to CsatBV-Iraq. The high coverage was at the start and nearly the end of the complete sequence

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

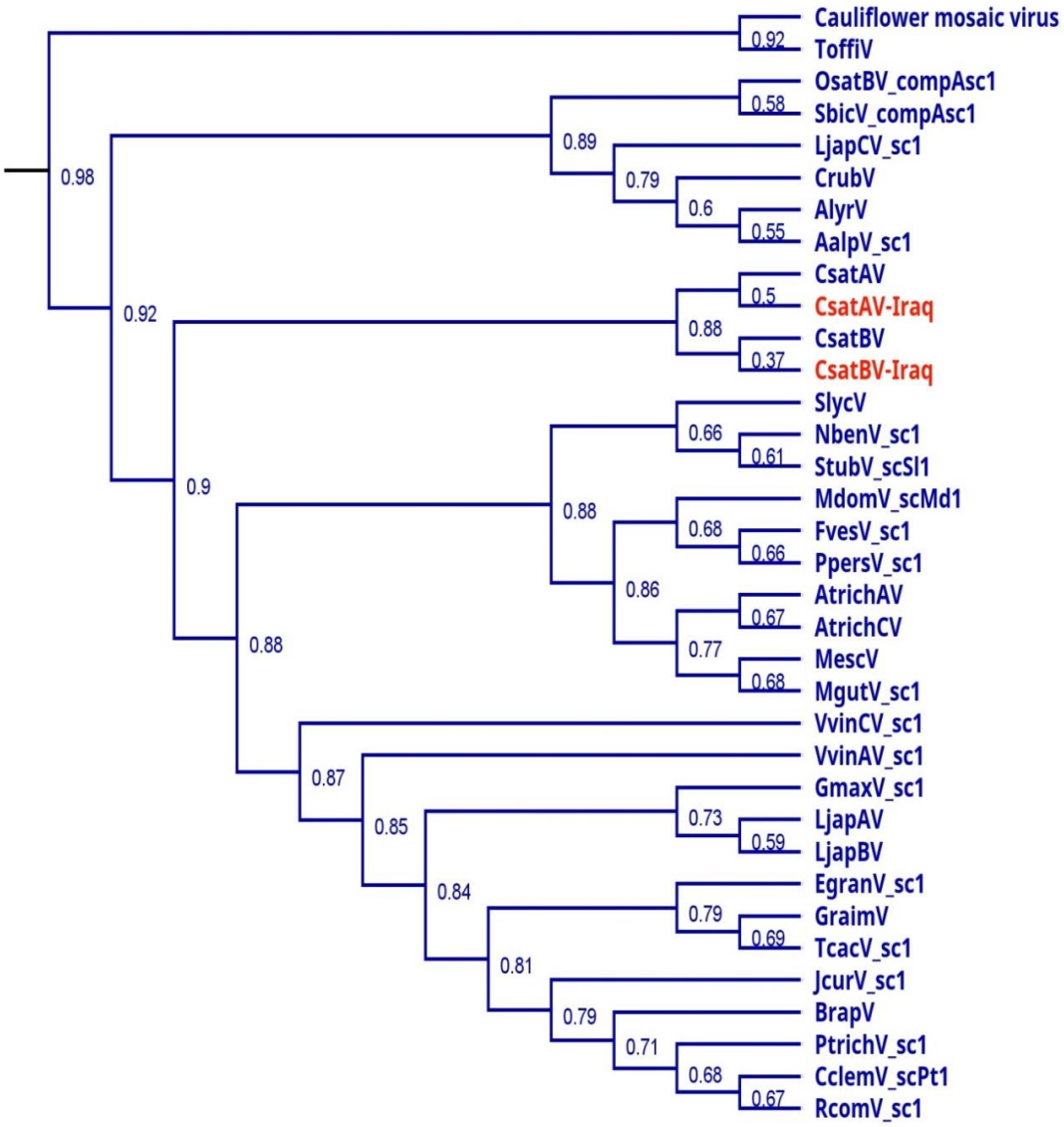

Fig 7: Maximum likelihood phylogenetic tree of a group of 34 Florendo viruses in addition to CaMV using entire sequences, showing high relationship between CsatAV-Iraq and CsatBV-Iraq (in red) with related Florendo viruses from Geering et al., (2014)

The sequences were aligned by ClustalW alignment and then tree was reconstructed through MrBayes inference. The number next each branch is for posterior probability (> 0.5 support).

4. DISCUSSION

In plants, endogenous Pararetroviruses (EPRVs) were detected through the sequencing protocols (Alisawi, 2019). Endogenous Pararetroviruses (EPRVs) are double-stranded nucleic acids incorporated into the genome of the host and are members of the

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

Caulimoviridae family. The elements are thought to have originated 1.6 billion years ago (Schmidt et al., 2021, Richert-Pöggeler et al., 2021). A Florendovirus, one of the Caulimoviridae genera, is the most prevalent virus in plant genomes (Geering et al., 2012; Diop et al., 2018). In most cases, these viruses are not infectious due to fragmentation and rearrangement. A gene silencing factor in the plant genome controls these viruses and prevents them from causing infection. Thus, they remain dormant in their host genomes and cannot escape. The integrants tend to hide in non- coding regions of the host genome (Schmidt et al. 2021). However, other integrated viruses may not subject to this factor as a result of its weak association and unstable nature in the host genome. There were two elements of Florendovirus identified earlier in cucumber genomes, CsatAV-Iraq and CsatBV -Iraq (Geering et al., 2014). More importantly, caulimoviruses have not been characterized in the genomes of cucumber yet. Further, Geering et al., (2014) did not explain the genome fragmentation of such viruses where our results show the clear integration of plant genes within the viral sequences, antiphage_ZorA_4 (anti-phage defence ZorAB system protein ZorA) and PTZ00440 (reticulocyte binding protein 2-like protein) in CsatAV -Iraq and DUF5864 in CsatBV-Iraq. Some RT and RNase H domains that belong to retrotransposon also incorporated in the two sequences. At least, one domain of RNase_HI originally belongs to Gypsy family and probably mapped in the sequence of CsatAV-Iraq and CsatBV-Iraq due to existence in the same region. The CsatBV-Iraq was more abundant than CsatAV-Iraq based on genome proportion and copy number values. Surprisingly, CsatAV-Iraq shows no expression so far, while CsatBV-Iraq shows dispersed and unequal level of coverage in non-conserved regions. The results showed high agreement with Geering et al. (2014) about 9% of Florendovirus positions within plant gene introns based on data from the Vitis vinifera genome suggest biological effects of such elements on gene expression. The Florendovirus motifs accumulated in fragile poly-TA repeat sites in heterochromatin regions, probably including secondary structures like hairpins, preventing chromosome fragility (Zlotorynski et al. 2003; Dillon et al. 2013). It is possible that these elements could act as fillers to repair DNA breaks either by microhomologymediated end joining or by non-homologous end joining. (Huertas 2010). The results disagreed with Geering et al. (2014) about Florendovirus sequences being represented well in EST databases, suggesting that florendoviruses are transcribed based on three out of 27 host genomes (Citrus clementia, Oryza sativa and Prunus persica), with a range of 47-57% alignment identities. It has been found that transcription of Florendovirus-like sequences varies depending on host effect and specificity. It would be helpful to further explore more facts from other host genomes.

5. CONCLUSION

NGS technology and bioinformatics used in this study and revealed that the genome of cucumber has two endogenous viral elements integrated within plant genes and belong to the genus Florendovirus, CsatAV –Iraq and CsatBV –Iraq. The entire sequences of both EPRVs have been registered. The analysis showed that the integrants were not well expressible in the RNA transcripts.

Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)
ISSN：1671-5497
E-Publication: Online Open Access
Vol: 42 Issue: 03-2023
DOI 10.17605/OSF.IO/9P4QD

REFERENCES

1. Alisawi, O. N. (2019). Virus integration and tandem repeats in the genomes of Petunia (Doctoral dissertation, University of Leicester).
2. Arumanagathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep. 9: 208-218. 10.1007/BF02672069.
3. Bos (2000). Plant viruses, unique and intriguing pathogens, Backhuys, Publishers, Leiden p: 358.
4. Dillon L.W., Pierce L.C., Ng M.C. & Wang Y.-H. (2013). Role of DNA secondary structures in fragile site breakage along human chromosome 10. Human Molecular Genetics 22, 1443-56.
5. Diop S.I.، Geering A.D.، Alfama-Depauw F.، Loaec M.، Teycheney P.-Y. & Maumus F. (2018). Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. Scientific reports 8، 572.
6. Eifediyi, E. K., & Remison, S. U. (2010). Growth and yield of cucumber (Cucumis sativus L.) as influenced by farmyard manure and inorganic fertilizer. Journal of Plant Breeding and Crop Science, 2(7), 216-220.
7. Geering, A. D. W., & Hull, R. (2012). Family caulimoviridae. Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses, 424-443.
8. Geering, A. D., Maumus, F., Copetti, D., Choisne, N., Zwickl, D. J., Zytnicki, M., & Teycheney, P. Y. (2014). Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. Nature Communications, 5(1), 1-11.
9. Huang, S., Li, R., Zhang, Z., Li, L. I., Gu, X., Fan, W., & Li, S. (2009). The genome of the cucumber, Cucumis sativus L. Nature genetics, 41(12), 1275-1281.
10. Huertas, P. (2010). DNA resection in eukaryotes: deciding how to fix the break. Nature structural & molecular biology, 17(1), 11-16.
11. Lough, T.J. & Lucas, W.J. (2006). Integrative plant biology: role of phloem long-distance macromolecular trafficking. Annu. Rev. Plant Biol. 57, 203–232.
12. Richert-Pöggeler, K. R., Vijverberg, K., Alisawi, O., Chofong, G. N., Schwarzacher, T., & Heslop-Harrison, J. S. (2021). Participation of multifunctional RNA in replication, recombination and regulation of endogenous plant Pararetroviruses (EPRVs). Frontiers in Plant Science, 12, 1148.
13. Schmidt, N., Seibt, K. M., Weber, B., Schwarzacher, T., Schmidt, T., & Heitkam, T. (2021). Broken, silent, and in hiding: Tamed endogenous Pararetroviruses escape elimination from the genome of sugar beet (Beta vulgaris). Annals of Botany, 128(3), 281-299.
14. Xoconostle-Cázares, B. et al. (1999) Plant paralog to viral movement protein that potentiates transport of mRNA into the phloem. Science 283, 94–98.
15. Yohe، S.، & Thyagarajan، B. (2017). Review of clinical next-generation sequencing. Archives of pathology & laboratory medicine، 141(11)، 1544-1557.
16. Zlotorynski E., Rahat A., Skaug J., Ben-Porat N., Ozeri E., Hershberg R., Levi A., Scherer S.W., Margalit H. & Kerem B. (2003) Molecular basis for expression of common and rare fragile sites. Molecular and Cellular Biology 23, 7143-51.