Opinion piece

Endogenous pararetroviruses in rice genomes as a fossil record useful for the emerging field of palaeovirology

SUNLU CHEN* AND YUJI KISHIMA

Laboratory of Plant Breeding, Research Faculty of Agriculture, Hokkaido University, Sapporo 060-8589, Japan

Pararetroviruses are double-stranded DNA viruses that use reverse transcription for replication, but do not integrate into host genomes. They include plant viruses of the family *Caulimoviridae*, such as *Rice tungro bacilliform virus* (RTBV), which is part of the virus complex causing rice tungro disease in South and South-East Asia (Jones *et al.*, 1991). There is no experimental evidence of obligatory integration processes or integrase mechanisms in pararetroviruses. However, studies of cases of pararetrovirus vertical transmission have discovered integrated pararetrovirus sequences in plant genomes (Harper *et al.*, 1999; Ndowora *et al.*, 1999; Richert-Poggeler *et al.*, 1996). These pararetrovirus-derived sequences were subsequently termed endogenous pararetroviruses (EPRVs). With the rapid development of sequencing technologies, increasing numbers of different types of EPRV have been discovered in various plant genomes.

EPRVs have been classified into two groups: those that can be activated to cause episomal infections (i.e. infective EPRVs) and those that are possibly incapable of inducing episomal infection (i.e. non-infective EPRVs) (Harper *et al.*, 2002). Three cases of infective EPRV have been confirmed in the respective interspecific hybrids of tobacco, petunia and banana (reviewed by Chabannes and Iskra-Caruana, 2013), which suggests that some EPRVs are a reservoir for viral infections. Most EPRVs are non-infective because of their mutated, deficient structures or epigenetically repressed states in plant genomes (Staginnus and Richert-Poggeler, 2006; Staginnus *et al.*, 2009), but researchers have focused mainly on infective EPRVs. However, there is value in a more thorough characterization of non-infective EPRVs, particularly in terms of virus as well as plant evolution. Endogenous RTBV-like (eRTBVL) sequences in the rice (*Oryza sativa* L.) genome were the first cereal EPRVs to be identified, and are amongst the most comprehensively studied non-infective EPRVs (Chen *et al.*, 2014; Kunii *et al.*, 2004; Liu *et al.*, 2012).

Using eRTBVLs as models, we discuss how EPRVs in plant genomes can serve as a genomic fossil record of past viral sequences and invasion events. This will help to broaden the understanding of viruses, particularly with regard to their evolution and interactions with their plant hosts. We also discuss the perspectives and challenges for future studies, especially those

involving the possible effects of EPRVs on plant biological processes, and justify why a greater focus on EPRVs is warranted.

HOW DO PARARETROVIRUSES COME TO BE PRESENT IN PLANT GENOMES AND HOW CAN THEY SERVE AS A FOSSIL RECORD?

EPRVs are frequently located in retrotransposon-rich regions of plant genomes (Staginnus and Richert-Poggeler, 2006), and therefore it was assumed that pararetrovirus integration was mediated by retrotransposons (i.e. pararetrovirus RNA hitchhikes on retrotransposon integrase activity). However, to date, there is no proof of genetic hitchhiking associated with retrotransposon activities or other active integration events. Detailed sequence analyses of EPRVs and their flanking regions generally support incidental integrations of pararetroviruses as a result of illegitimate recombinations between pararetrovirus and plant genomes (Liu *et al.*, 2012). However, these integration events were not completely random. A typical case involves eRTBVLs that were preferentially integrated into AT-rich regions in rice genomes (Kunii *et al.*, 2004). This preferential integration has also been observed for some other EPRVs in a broad range of plant genomes (Geering *et al.*, 2014). Comprehensive genomic analyses of rice subspecies have revealed that AT-rich regions, especially AT repeats, are hot spots for DNA double-strand breaks (DSBs) (Liu *et al.*, 2012). When pararetrovirus DNA sequences appeared in the rice nucleus, they were incidentally used for DSB repair of AT-rich regions through a non-homologous end-joining pathway (Liu *et al.*, 2012). This resulted in illegitimate recombinations between pararetrovirus and host genomes, suggesting that these incidental integrations might have initially been of benefit to the stability of the host genome. The AT-rich regions were also recognized as possible scaffold/matrix attachment regions (Liu *et al.*, 2012) involved in the organization of chromatin and the regulation of gene expression. The cleavage of topoisomerase IIA cleavage sites in scaffold/matrix attachment regions may promote DSBs that can be exploited for DNA integration (Liu *et al.*, 2012). However, the role of the nuclear scaffold/matrix in pararetrovirus integration has not been characterized in detail.

Integration by DSB repair of AT-rich regions is probably not the only mechanism enabling pararetrovirus integration. In some plant

*Correspondence: Email: sunlu@abs.agr.hokudai.ac.jp

genomes (e.g. *petunia* and *Fritillaria imperialis*), EPRVs are partly or even mostly located in centromeric or pericentromeric regions (Becher *et al.*, 2014; Richert-Poggeler *et al.*, 2003). However, the mechanism responsible for this type of integration, and its effects on the evolution and function of the plant centromere, are largely unknown. In some plant genomes, several copies of EPRVs, including eRTBVL sequences, cluster to form tandem repeats (Liu *et al.*, 2012; Richert-Poggeler *et al.*, 2003). This clustering may have resulted from the linking of the termini of linear pararetrovirus sequences prior to integration, or from recombinations between EPRVs and episomal pararetrovirus genomes in the plant nucleus. With regard to recombinations between EPRVs and pararetroviruses, it is interesting to consider whether horizontal gene transfers from EPRVs to pararetroviruses can or in fact did occur by recombination during infection.

Pararetrovirus DNA sequences were integrated into plant somatic or germline genomes by DSB repair or other unknown mechanisms. However, only integrations that occurred in germline cells could be inherited by subsequent generations. Some integrated pararetrovirus sequences in individual plants were fixed within the plant population by genetic drift or natural selection, finally becoming EPRVs in these plants (the process of fixation of an integration event is called endogenization). During and after the endogenization process, EPRVs may have accumulated mutations and undergone epigenetic regulation. Consequently, EPRVs are a record of ancient pararetrovirus sequences, and mirror past pararetrovirus invasions.

According to the known distribution patterns of EPRVs within and across plant taxa, a specific tendency of integrations appears to have occurred. Some plant genomes harbour multiple diverse EPRVs, whereas others harbour no known EPRVs; some types of pararetrovirus have been integrated into the genomes of widespread, but distinct, plant species, whereas others have been integrated into a very limited variety of plant genomes (Geering *et al.*, 2014; Staginnus and Richert-Poggeler, 2006; Staginnus *et al.*, 2009). Plant proteins that influence the success of integrations of pararetrovirus sequences have not yet been identified. It is also unclear whether some pararetrovirus proteins can affect the integration tendency. Despite the abundance of characterized eRTBVL sequences, no extant RTBV sequences have been detected in sequenced plant genomes. A protein encoded by RTBV open reading frame 2 participates in capsid assembly to complete virion packing, but the counterpart of this protein was not encoded by any of the eRTBVL sequences, and was therefore absent from the virus of eRTBVLs (Liu and Kishima, 2014). Thus, this protein was assumed to suppress the integration of RTBV genomes into rice genomes by decreasing the number of unpacked viral genomes (Liu and Kishima, 2014). Further research is required to clarify the specific interactions that influence pararetrovirus integration.

WHAT CAN EPRVS TELL US ABOUT THE VIRAL COMMUNITY?

Most of the viruses of EPRVs are ancient but novel pararetrovirus species with diverse genomic structures and genes (Geering *et al.*, 2010, 2014), and the analysis of EPRVs in plant genomes may be an effective additional approach to the study of pararetrovirus biodiversity and ecology. Extant plant pararetroviruses currently fall into eight genera (International Committee on Taxonomy of Viruses: <http://ictvonline.org/virusTaxonomy.asp>), and three new genera have been suggested for the viruses of some EPRVs (Geering *et al.*, 2010, 2014). RTBV is the only recognized species in the genus *Tungrovirus*, and the only extant pararetrovirus known to infect rice. The virus of eRTBVLs is an ancient *Tungrovirus* species, and its genomic structure differs from that of RTBV (Chen *et al.*, 2014; Kunii *et al.*, 2004). Screening for orthologous eRTBVL sequences in *Oryza* species revealed some endogenization events of eRTBVLs that occurred in Africa millions of years ago (Chen, S. and Kishima, Y., unpublished data). Several distinct EPRVs similar to *Petunia vein clearing virus* (PVCV; genus *Petuvirus*) have also been identified in the genomes of rice and related plants (Geering *et al.*, 2014). Based on structural and phylogenetic analyses, a new genus (*Florendovirus*) has been proposed for the viruses of these PVCV-like EPRVs (Geering *et al.*, 2014). Interestingly, some PVCV-like EPRVs, including those in rice genomes, have been found to be derived from a virus with a bipartite genome (Geering *et al.*, 2014). All known pararetroviruses have monopartite genomes.

The viral species of EPRVs may now be extinct, but the direct (or close) progenies of some of these viruses might still be circulating in certain plant species. Some EPRV sequences may have helped to protect plants from related viruses (Staginnus and Richert-Poggeler, 2006; discussed in detail below), resulting in only mild or weak disease symptoms during viral infections. This may have caused these viruses to be overlooked. However, these viruses may be a source for the emergence of viral diseases in the future.

WHAT EVOLUTIONARY STORIES CAN EPRVS TELL US?

Natural fossils of viral nucleotides and particles are lacking, but EPRVs are genomic fossils. The study of EPRVs, such as eRTBVLs, can unravel multiple aspects of virus evolution and co-evolution between viruses and plants over long periods.

First, EPRVs can further our understanding of the origins of current pararetroviruses. Rice tungro disease has been responsible for severe epidemics that have adversely affected rice production in many parts of Asia (Jefferson and Chancellor, 2002). However, the evolutionary origin of RTBV is unknown. Our analysis of eRTBVL sequences has revealed that the divergence between RTBV and the virus of eRTBVLs occurred earlier than the speciation of *Oryza rufipogon*, which is the direct progenitor of

rice (Chen *et al.*, 2014). In addition, the two viruses shared a common progenitor at least 160 000 years ago (Chen *et al.*, 2014). This finding suggests that the ancestral lineages of RTBV emerged very early.

Second, EPRVs can provide novel insights into viral macroevolution. The experimental and phylogenetic evidence from current pararetroviruses has revealed a high recombination rate for pararetrovirus genomes. However, the detection of recombination signals using the reconstructed genomes of the virus lineages of eRTBVLs has indicated that a few commonly occurring recombination events determined the main long-term virus genealogy (Chen *et al.*, 2014). This result suggests that the long-term recombination rate may be much lower than the short-term rate. Integrated sequences of an ancient animal virus (hepadnavirus) in bird genomes have revealed that the long-term mutation rate for the virus was 10^{-8} substitutions per site per year (Gilbert and Feschotte, 2010). Long-term plant virus mutation rates have not been reported, but they could be calculated using the multiple sequences of an EPRV that were endogenized into plant genomes at different time points (Gilbert and Feschotte, 2010). The long-term mutation rates of plant viruses may differ from those of animal viruses because of differences in host evolutionary rates. In addition, a comparison of the different viral genomic structures for EPRVs enables the study of the macroevolution of viral genome organizations, open reading frames and functional domains.

Third, EPRVs can help to clarify the long-term evolutionary dynamics of host–virus interactions. The analysis of eRTBVL endogenization events has revealed that the evolution of the different virus lineages of eRTBVLs was spatio-temporally coupled with the divergence and spread of rice populations (Chen *et al.*, 2014). This analysis provides evidence for the co-evolution between pararetroviruses and host plants. Furthermore, orthologous sequences of a given EPRV in different plant genomes can indicate the host range of a virus and allow the identification of ancient host switch events, which represent an evolutionary pressure for plant hosts and viruses. By combining data from various fields, including palaeobotany, geography and climatology, the complex host–virus interactions underlying host switch events can be investigated.

Finally, plant phylogeny can be used to elucidate the timelines of the endogenization of EPRVs in plant genomes, and vice versa. Specifically, EPRVs can serve as special markers to clarify the evolutionary histories and phylogenetic relationships of plant species. The presence/absence patterns and sequence polymorphisms in a given EPRV locus (i.e. orthologous EPRV sequences) in different plant populations or species can indicate the relationships among related plants. Based on the known evolutionary history of rice, most eRTBVL sequences have been confirmed to have been endogenized in rice genomes before domestication, and the genetic flow of eRTBVL sequences is in accordance with the history of rice domestication (Chen *et al.*, 2014). The loci of the sequences from

the eRTBVL-X family, which is the youngest of six eRTBVL families, have been detected only in specific accessions of the *japonica* subspecies of rice and in *O. rufipogon* from China (Chen *et al.*, 2014). This finding indicates that these eRTBVL loci could be used to characterize the late stages of the domestication of *japonica* rice.

DO EPRVS CONFER RESISTANCE TO PARARETROVIRUS INFECTIONS?

Abundant numbers of EPRVs have been fixed in various plant species during evolution. Consequently, the potential role of EPRVs in the interplay between viruses and hosts has been debated for a long time (Staginnus and Richert-Pöggeler, 2006). EPRV-derived short-interfering RNAs (siRNAs) and host epigenetic silencing of EPRVs have frequently been reported in plants (Becher *et al.*, 2014; Staginnus and Richert-Pöggeler, 2006). Transgenes with an EPRV-derived promoter have been found to be methylated and silenced in tobacco lines that possess the corresponding EPRVs, but not in *Arabidopsis thaliana*, which lacks these EPRVs (Mette *et al.*, 2002). These observations led researchers to suggest the existence of EPRV-derived antiviral resistance that relies on homology-dependent gene silencing. The EPRV-derived siRNAs target homologous pararetrovirus genomes *in trans* to repress their transcription through promoter methylation, or target pararetrovirus RNA at the post-transcriptional level (Staginnus and Richert-Pöggeler, 2006). The histone modification pathway also regulates EPRVs (Noreen *et al.*, 2007). Therefore, EPRV-derived siRNAs may induce negative histone modifications on the minichromosomes of related pararetroviruses in plant nuclei. However, direct evidence for EPRV-derived resistance has not been found to date.

Several studies have reported that some endogenous retroviruses in animal genomes have been co-opted as cellular genes that function as inhibitors of viral infections at the protein level (reviewed by Aswad and Katourakis, 2012). It is still unknown whether some EPRVs have been domesticated by host plants to serve as antiviral protein-coding genes. One way to identify a co-opted EPRV is to screen conserved EPRVs that have retained their coding ability after long-term evolution, and to examine their transcriptional and translational activities in response to infection by a related virus. Such EPRVs might produce a protein with competitive binding activity for the viral protein targets. Such competitive inhibition may disrupt viral genome replication or particle assembly, or interfere with the intercellular movement of viruses.

In addition to being a record of ancient viral sequences, EPRVs represent ancient invasion events. Invasions by ancient pararetroviruses probably drove the evolution of host antiviral factors. Host antiviral proteins, especially those that interact directly with viral proteins, need to continually and rapidly evolve to counteract the constantly evolving viruses. Consequently, these antiviral proteins

exhibit positive selection signatures in certain domains. Positive selection, possibly driven by ancient retroviruses, has been observed in animal genomes (e.g. positive selection of the primate TRIM5 antiviral protein; Sawyer *et al.*, 2005). To assess the possible impact from ancient pararetrovirus invasions, orthologous anti-pararetroviral factors need to be identified and analysed in multiple plant species with or without EPRVs.

It is interesting to consider whether ancient rice pararetroviruses or their eRTBVL sequences have been involved in conferring resistance to their extant sister RTBV. Kunii *et al.* (2004) suggested that there was an association between eRTBVL profiles and the degree of RTBV susceptibility (i.e. species with low eRTBVL copy numbers tend to be vulnerable to RTBV infection). An abundance of microhomologous regions between eRTBVL and RTBV sequences enables base pairing between eRTBVL-derived siRNAs and RTBV sequences, which may guide plant epigenetic machinery to induce DNA methylation in RTBV genomes or histone methylation of the viral minichromosomes. In addition, multiple copies of members of the eRTBVL-D family (the oldest eRTBVL family) have been conserved in all examined rice accessions (Chen *et al.*, 2014). Additional research is required to comprehensively characterize the evolution of the coding abilities of these eRTBVL sequences. Furthermore, ancient viral invasion events revealed by eRTBVLs might impose natural selection on certain anti-pararetroviral factors, which could also result in some resistance against related viruses.

CONCLUSION

Three extreme cases of episomal infections caused by infective EPRVs have been used as models for EPRV studies; however, most EPRVs are non-infective. Using non-infective EPRVs, such as eRTBVLs, as models has enabled research on viral genomic fossils, which has contributed to the emerging field of palaeovirology, namely the study of viruses on evolutionary timescales. The eRTBVLs in rice genomes represent valuable examples of EPRVs as a fossil record for investigations into the origin and evolution of pararetroviruses, as well as the long-term interaction and co-evolution between pararetroviruses and plants. EPRVs may also play important roles in host plant biological activities: in particular, they may confer resistance to pararetrovirus infections. However, direct evidence of any such roles is still lacking. More focus on EPRV research is warranted, and will provide further insights into host–virus systems.

ACKNOWLEDGEMENTS

We thank the three anonymous reviewers for their helpful comments and suggestions. We thank the Sekisui Chemical Grant Program (Tokyo, Japan) for financial support (YK).

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

- Aswad, A. and Katourakis, A. (2012) Paleovirology and virally derived immunity. *Trends Ecol. Evol.* **27**, 627–636.
- Becher, H., Ma, L., Kelly, L.J., Kovarik, A., Leitch, I.J. and Leitch, A.R. (2014) Endogenous pararetrovirus sequences associated with 24 nt small RNAs at the centromeres of *Fritillaria imperialis* L. (Liliaceae), a species with a giant genome. *Plant J.* **80**, 823–833.
- Chabannes, M. and Iskra-Caruana, M.L. (2013) Endogenous pararetroviruses – a reservoir of virus infection in plants. *Curr. Opin. Virol.* **3**, 615–620.
- Chen, S., Liu, R., Koyanagi, K.O. and Kishima, Y. (2014) Rice genomes recorded ancient pararetrovirus activities: virus genealogy and multiple origins of endogenization during rice speciation. *Virology*, **471**, 141–152.
- Geering, A.D.W., Scharaschkin, T. and Teycheney, P.Y. (2010) The classification and nomenclature of endogenous viruses of the family Caulimoviridae. *Arch. Virol.* **155**, 123–131.
- Geering, A.D.W., Maumus, F., Copetti, D., Choise, N., Zwickl, D.J., Zytynski, M., McTaggart A.R., Scalabrin, S., Vezzulli, S., Wing, R.A., Quesneville, H. and Teycheney, P.Y. (2014) Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun.* **5**, 5269.
- Gilbert, C. and Feschotte, C. (2010) Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS Biol.* **8**, e1000495.
- Harper, G., Osuji, J.O., Heslop-Harrison, J.S. and Hull, R. (1999) Integration of banana streak badnavirus into the *Musa* genome: molecular and cytogenetic evidence. *Virology*, **255**, 207–213.
- Harper, G., Hull, R., Lockhart, B. and Olszewski, N. (2002) Viral sequences integrated into plant genomes. *Annu. Rev. Phytopathol.* **40**, 119–136.
- Jefferson, O.A. and Chancellor, T. (2002) The biology, epidemiology, and management of rice tungro disease in Asia. *Plant Dis.* **86**, 88–100.
- Jones, M.C., Gough, K., Dasgupta, I., Rao, B.L.S., Cliffe, J., Qu, R., Shen, P., Kaniewska, M., Blakebrough, M., Davies, J.W., Beachy, R.N. and Hull, R. (1991) Rice tungro disease is caused by an RNA and a DNA virus. *J. Gen. Virol.* **72**, 757–761.
- Kunii, M., Kanda, M., Nagano, H., Uyeda, I., Kishima, Y. and Sano, Y. (2004) Reconstruction of putative DNA virus from endogenous rice tungro bacilliform virus-like sequences in the rice genome: implications for integration and evolution. *BMC Genomics*, **5**, 80.
- Liu, R. and Kishima, Y. (2014) Chapter 12: Establishment of endogenous pararetroviruses in the rice genome. In *Plant Virus–Host Interaction* (Gaur, R.K., Hohn, T. and Sharma, P., eds), pp. 229–240. Boston, MA: Academic Press.
- Liu, R., Koyanagi, K.O., Chen, S. and Kishima, Y. (2012) Evolutionary force of AT-rich repeats to trap genomic and episomal DNAs into the rice genome: lessons from endogenous pararetrovirus. *Plant J.* **72**, 817–828.
- Mette, M., Kanno, T., Aufsatz, W., Jakowitsch, J., van der Winden, J., Matzke, M.A. and Matzke, A.J. (2002) Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *EMBO J.* **21**, 461–469.
- Ndowora, T., Dahal, G., LaFleur, D., Harper, G., Hull, R., Olszewski, N.E. and Lockhart B. (1999) Evidence that badnavirus infection in *Musa* can originate from integrated pararetroviral sequences. *Virology*, **255**, 214–220.
- Noreen, F., Akbergenov, R., Hohn, T. and Richert-Pöggeler, K.R. (2007) Distinct expression of endogenous *Petunia* vein clearing virus and the DNA transposon *dTph1* in two *Petunia hybrida* lines is correlated with differences in histone modification and siRNA production. *Plant J.* **50**, 219–229.
- Richert-Pöggeler, K., Sheperd, R. and Casper, R. (1996) *Petunia* vein-clearing virus, a pararetrovirus that also exists as a retroelement in the chromosome of its host. In: *Abstract Book of the Xth International Congress of Virology, Jerusalem, Israel* 16.
- Richert-Pöggeler, K.R., Noreen, F., Schwarzacher, T., Harper, G. and Hohn, T. (2003) Induction of infectious *petunia* vein clearing (pararetro) virus from endogenous provirus in *petunia*. *EMBO J.* **22**, 4836–4845.
- Sawyer, S.L., Wu, L.I., Emerman, M. and Malik, H.S. (2005) Positive selection of primate TRIM5 α identifies a critical species-specific retroviral restriction domain. *Proc. Natl. Acad. Sci. USA*, **102**, 2832–2837.
- Staginnus, C. and Richert-Pöggeler, K.R. (2006) Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends Plant Sci.* **11**, 485–491.
- Staginnus, C., Iskra-Caruana, M., Lockhart, B., Hohn, T. and Richert-Pöggeler, K. (2009) Suggestions for a nomenclature of endogenous pararetroviral sequences in plants. *Arch. Virol.* **154**, 1189–1193.