



HOKKAIDO UNIVERSITY

Title	The Evolution of Endogenous Pararetrovirus Sequences in the Genomes of Rice and Other Grass Species
Author(s)	陈, 孙禄
Degree Grantor	北海道大学
Degree Name	博士(農学)
Dissertation Number	甲第11819号
Issue Date	2015-03-25
DOI	https://doi.org/10.14943/doctoral.k11819
Doc URL	https://hdl.handle.net/2115/60898
Type	doctoral thesis
File Information	Sunlu_Chen.pdf



**The Evolution of Endogenous Pararetrovirus Sequences
in the Genomes of Rice and Other Grass Species**

（イネ及びイネ科植物種のゲノムに内在する
パラレトロウイルス配列の進化）

北海道大学 大学院農学院
生物資源科学専攻 博士後期課程

陳 孫祿

Dedicated to
my mother and father:
Judi Liao & Zhonggai Chen

ABBREVIATIONS

AIC	Akaike information criterion
cDNA	complementary DNA
CP	capsid/coat protein
DSB	double strand break
eGPV1	endogenous grass pararetrovirus 1
eGPV2	endogenous grass pararetrovirus 2
EPRV	endogenous pararetrovirus sequence
eRPV	endogenous recombinant pararetrovirus
eRTBVL	endogenous rice tungro bacilliform virus-like
EVE	endogenous viral element
IGR	intergenic region
LTR	long terminal repeat
ML	maximum likelihood
MP	movement protein
MYA	million years ago
NHEJ	non-homologous end joining
ORF	open reading frame
PBS	primer binding site
PCR	polymerase chain reaction
PPT	polypurine tract
PR	protease
PVCV	petunia vein clearing virus
RT/RH	reverse transcriptase/RNase H
RTBV	rice tungro bacilliform virus

CONTENTS

Chapter 1	General introduction	1
1.1	Endogenous viral elements.....	1
1.2	Pararetroviruses & endogenous pararetrovirus sequences (EPRVs).....	3
1.3	Rice tungro bacilliform virus & endogenous rice tungro bacilliform virus-like sequences (eRTBVLs)	9
Chapter 2	Multiple origins of endogenization of eRTBVL sequences during rice speciation and the corresponding virus genealogy	12
2.1	Introduction.....	12
2.2	Materials and Methods.....	13
2.2.1	Plant materials and taxonomy	13
2.2.2	Genomic data mining	14
2.2.3	Phylogenetic analysis	15
2.2.4	Endogenization pattern analysis	15
2.2.5	Orthologous eRTBVL screening by primer sets.....	16
2.2.6	Endogenization order analysis.....	16
2.2.7	Recombination analysis.....	17
2.3	Results	18
2.3.1	The eRTBVL sequences in the genome of wild rice	18
2.3.2	Organization of the eRTBVL families.....	19
2.3.3	Phylogenetic relationships of eRTBVL sequences.....	22
2.3.4	Proportions of eRTBVL families in the three rice genomes.....	23
2.3.5	Distribution of eRTBVL sequences in cultivated and wild rice accessions	25
2.3.6	Temporal relationships of endogenization of eRTBVL families.....	28
2.3.7	Recombination of ancient pararetroviruses	30
2.3.8	The most ancient endogenization event of an eRTBVL family	34
2.3.9	Relationship between the viruses of eRTBVLs and RTBV.....	34
2.4	Discussion.....	38

2.4.1	Genealogy of the viral lineages of eRTBVLs led by recombination.....	38
2.4.2	Geographic distributions of viral lineages of eRTBVLs	42
Chapter 3 Widespread and independent endogenizations of eRTBVL superfamily in other grass genomes		44
3.1	Introduction	44
3.2	Materials and methods	45
3.2.1	Genomic data mining and paleoviral genome reconstruction	45
3.2.2	Phylogenetic analysis	46
3.2.3	Orthologous insertion analysis	46
3.3	Results	47
3.3.1	Sequences of eRTBVL superfamily present throughout grass genomes.....	47
3.3.2	Diverse viral genomic structures of eRTBVL superfamily	48
3.3.3	Abundant independent endogenization events of eRTBVL superfamily	49
3.4	Discussion	54
Chapter 4 The eGPV1 and eGPV2 sequences in grass genomes reveal a coevolved mutualism between nonautonomous viruses.....		56
4.1	Introduction	56
4.2	Materials and Methods	57
4.2.1	Genomic data mining and paleoviral genome reconstruction	57
4.2.2	Structure detection and domain prediction.....	58
4.2.3	Phylogenetic and recombination analyses.....	58
4.2.4	Orthologous insertion analysis	59
4.2.5	Plant materials and polymerase chain reaction.....	60
4.3	Results	61
4.3.1	Unusual <i>Petuvirus</i> -related EPRV, eGPV1 in the rice genome	61
4.3.2	Another unusual EPRV, eGPV2 naturally lacking an RT/RH domain but possessing an IGR highly similar to eGPV1	62
4.3.3	Multiple paleogenomes of nonautonomous pararetroviruses in <i>Oryza</i>	

genomes	69
4.3.4 Co-evolution of two functionally complementary nonautonomous pararetroviruses	73
4.3.5 IGR swapping establishes and maintains the mutualism between nonautonomous pararetroviruses.....	78
4.4 Discussion.....	82
Chapter 5 Recombinant pararetrovirus-derived eRPV sequences support an interspecies recombination-mediated viral modular evolution.....	84
5.1 Introduction.....	84
5.2 Materials and methods	85
5.2.1 Genomic data mining	85
5.2.2 Phylogenetic analysis	85
5.2.3 Orthologous insertion analysis	86
5.3 Results	86
5.3.1 The eRPV sequences derived from a recombinant virus recently.....	86
5.3.2 Interspecies illegitimate recombination led modular evolution and subsequently symbiosis switch.....	87
5.4 Discussion.....	92
Chapter 6 General discussion	94
6.1 Outcome of ancient grass pararetrovirus and origin of extant one.....	94
6.2 Impact of ancient pararetroviruses on host immunity	96
6.3 Revisit the mechanism of integration of EPRVs into grass genomes	98
Chapter 7 Summary	101
Supplementary Figures	104
Supplementary Tables.....	118
References.....	154
Acknowledgements	167

LIST OF FIGURES

Figure 1.1	A scheme for the origin of retroelements including pararetroviruses	4
Figure 1.2	Linear genomic structures of typical plant pararetroviruses	5
Figure 1.3	Diagram of the replication cycle of typical plant pararetroviruses	6
Figure 1.4	Rice tungro disease symptoms (A) and pathogens (B)	10
Figure 1.5	Genomic structures of RTBV and the virus of eRTBVLs	11
Figure 1.6	A sketch of the integration of the virus sequences of eRTBVLs into host AT-repeat regions.....	11
Figure 2.1	eRTBVL families identified by a phylogenetic analysis based on the RT/RH region in eRTBVL sequences	20
Figure 2.2	Structure of the eRTBVL-X cluster on chromosome 8 of Nipponbare	21
Figure 2.3	Global view of phylogenetic relationships based on the sequences of seven regions of the eRTBVL sequences	22
Figure 2.4	Endogenization patterns of three eRTBVL families	24
Figure 2.5	Endogenization patterns of three eRTBVL families for Nipponbare- and 93-11-specific eRTBVLs	24
Figure 2.6	Orthologous eRTBVL distribution in a collection of cultivated and wild rice accessions	27
Figure 2.7	Detailed structure of the viral linear genome of constructed eRTBVL consensus sequences.....	32
Figure 2.8	Recombination analysis of the eRTBVL families.....	33
Figure 2.9	Distribution of orthologous eRTBVL-D loci in <i>Oryza</i> genomes	36
Figure 2.10	Phylogenetic relationship between eRTBVL and RTBV sequences.....	37
Figure 2.11	Schematic representation of possible integration events of eRTBVLs during rice and pararetrovirus evolution.....	41
Figure 3.1	Phylogenetic analysis of eRTBVL sequences in grass genomes and exogenous RTBV sequences based on RT/RH domain.....	51

Figure 3.2 Three types of virus genomic structures of eRTBVL sequences in grass genomes	52
Figure 3.3 Sequence comparison between the ORFx1-ORFx2 and the ORFx	53
Figure 4.1 Reconstruction of the viral genomes of eGPV1 and eGPV2 sequences in three <i>Oryza</i> genomes	64
Figure 4.2 Phylogenetic relationship of eGPV1/eGPV2 and pararetrovirus sequences based on the RT/RH and CP domains.....	65
Figure 4.3 Phylogenetic relationship of eGPV1/eGPV2 and pararetrovirus sequences based on the MP and PR domains	66
Figure 4.4 Dot plot of pairwise sequence comparison between eGPV1 and eGPV2 sequences	67
Figure 4.5 Insertion events of eGPV1s and eGPV2s sequences into grass genomes..	71
Figure 4.6 Genomic PCR detection of the presence/absence of representative eGPV1-OS/eGPV2-OS insertions among orthologous loci in <i>Oryza</i> genomes	72
Figure 4.7 Reconstruction of the viral genomes of eGPV1 and eGPV2 sequences in the <i>S. bicolor</i> genome	75
Figure 4.8 Pairwise alignments of protein sequences of eGPV1s and eGPV2s.....	76
Figure 4.9 Pairwise alignments of the IGR sequences of eGPV1s and eGPV2s	77
Figure 4.10 Phylogenetic relationships of the IGR sequences of eGPV1s and eGPV2s	80
Figure 4.11 Virus-derived small indel (deletions and insertions) variations in the eGPV1-SB and eGPV2-SB IGR sequences	81
Figure 4.12 A proposed mutualism between two ancient nonautonomous pararetroviruses in the life circles in host cells.....	83
Figure 5.1 The eRPV sequences derived from a recombinant pararetrovirus.....	90
Figure 5.2 Phylogenetic relationship of the homologous parts between eRPV and eGPV2 sequences	91
Figure S1 ML phylogenetic analysis of eRTBVL elements based on ORFx	104

Figure S2	ML phylogenetic analysis of eRTBVL elements based on the MP region	105
Figure S3	ML phylogenetic analysis of eRTBVL elements based on the CP region	106
Figure S4	ML phylogenetic analysis of eRTBVL elements based on the PR region	107
Figure S5	ML phylogenetic analysis of eRTBVL elements based on ORFz	108
Figure S6	ML phylogenetic analysis of eRTBVL elements based on the IGR	109
Figure S7	Informative SNP patterns in the RT/RH sequences of eRTBVL families and presumed breakpoint positions in the alignment	110
Figure S8	Informative SNP patterns in the IGR sequences of eRTBVL families and presumed breakpoint positions in the alignment (for Figure 2.8A)	111
Figure S9	Informative SNP patterns in the IGR sequences of eRTBVL families and presumed breakpoint positions in the alignment (for Figure 2.8B).....	112
Figure S10	Informative SNP patterns in the ORFx sequences of eRTBVL families and presumed breakpoint positions in the alignment	113
Figure S11	Informative SNP patterns in the MP sequences of eRTBVL families and presumed breakpoint positions in the alignment	114
Figure S12	Informative SNP patterns in the CP sequences of eRTBVL families and presumed breakpoint positions in the alignment	115
Figure S13	Informative SNP patterns in the ORFz sequences of eRTBVL families and presumed breakpoint positions in the alignment	116
Figure S14	Motif alignments in the four domains of rice EPRVs and various exogenous pararetrovirus sequences.....	117

LIST OF TABLES

Table 1.1	EVEs identified in plant genomes	2
Table 2.1	Sequence divergences of three eRTBVL families after endogenization based on sole-SNP calling	29
Table 3.1	Summary of sequence distribution of eRTBVL superfamily in 11 grass genomes	50
Table 4.1	Identification of functional domains in eGPV1 and eGPV2 proteins	68
Table S1	Plant materials used in this study	118
Table S2	Information of investigated grass genome databases	120
Table S3	Mapping of W1943-eRTBVLs into the Nipponbare and 93-11 genomes based on flanking sequence alignments.....	121
Table S4	Detection of the eRTBVL-X cluster in cultivated and wild rice accessions by genomic PCR.....	123
Table S5	Genomic locations of orthologous eRTBVL-D loci in <i>Oryza</i> genomes.....	126
Table S6	Sequence distribution of eRTBVL superfamily in grass genomes.....	127
Table S7	Distribution of eGPV1 and eGPV2 segments in Poaceae genomes.....	143
Table S8	Mapping of eGPV1 and eGPV2 loci in the <i>O. sativa</i> genome into other <i>Oryza</i> genomes	148
Table S9	Mapping of eGPV1 and eGPV2 loci in the genomes of <i>O. glaberrima</i> , <i>O. glumaepatula</i> , <i>O. meridionalis</i> and <i>O. punctata</i> into the genome of <i>O. sativa</i>	149
Table S10	Primers used in this study.....	151

Chapter 1 General introduction

1.1 Endogenous viral elements

During the viral incidental invasion into the host germline cells, viral sequences can integrate into the host germline genomes and some sequences can be inherited as part of the genetic makeup of the host species across generations in a Mendelian fashion; this process is called endogenization and these viral-derived sequences are named endogenous viral elements (EVEs) (Feschotte and Gilbert, 2012). Originally, EVEs were characterized exclusively as endogenous retroviruses, which were derived from retroviruses, the only known eukaryotic viruses requiring chromosomal integration for replication (Feschotte and Gilbert, 2012). Surprisingly, EVEs embedded in various eukaryotic genomes were recently identified to be derived from those viruses without integration ability, even from no-retroviral RNA viruses (Staginnus and Richert-Pöggeler, 2006; Taylor and Bruenn, 2009; Belyi *et al*, 2010; Horie *et al*, 2010; Katzourakis and Gifford, 2010; Liu *et al*, 2010; Chiba *et al*, 2011; Liu *et al*, 2011). The EVEs identified in plant genomes were summarized in Table 1.1. Natural fossils of viruses are difficult to obtain, but the vertical inheritance of EVEs has provided an unprecedented opportunity to explore the deep evolution of viruses (Patel *et al*, 2011; Feschotte and Gilbert, 2012; Katzourakis, 2013). The mining and studying of EVEs in eukaryotic genomes therefore are contributing to the emerging field of paleovirology, which provides fresh insights into the origin of extant viruses and the impact of ancient viruses on host biology (Emerman and Malik, 2010; Patel *et al*, 2011).

Each EVE sequence was initiated from a single integration event in a single host individual and gradually reached fixation in a population by genetic drift and/or natural selection (Chessa *et al*, 2009; Johnson, 2010; Katzourakis and Gifford, 2010). Some EVEs have been domesticated by hosts to serve in certain host important physiology (Feschotte and Gilbert, 2012), such as a partitivirus coat protein sequence which

contributed to a host gene (*ILR2*) to regulate the activities of the phytohormone auxin in *Arabidopsis thaliana* (Chiba *et al*, 2011). Some EVEs were thought to be involved in the resistance to exogenous virus infections (Aswad and Katzourakis, 2012), such as an endogenous bornavirus-like element (itEBLN) which can inhibit the replication of Borna disease virus in ground squirrel (Fujino *et al*, 2014). Independent integration events of EVEs have led to different EVE profiles in the genomes of host species, which can help us to resolve unclear phylogenetic relationships of host animal and plant species (Chiba *et al*, 2011).

Table 1.1 EVEs identified in plant genomes (modified from Feschotte and Gilbert, 2012)

Family or genus	Genome type	Host genome	Number per haploid genome	References
Potyviridae	+ssRNA	Grape	Several	Tanne and Sela (2005)
Betaflexiviridae	+ssRNA	Cucumber	1	Chiba <i>et al</i> (2011)
Rhabdoviridae	-ssRNA	Widespread	1 or 2	Chiba <i>et al</i> (2011)
Varicosavirus	-ssRNA	Widespread	1 or 2	Chiba <i>et al</i> (2011)
Partitiviridae	dsRNA	Widespread	1 to 4	Liu <i>et al</i> (2010), Chiba <i>et al</i> (2011)
Totiviridae	dsRNA	Widespread	1 to 6	Liu <i>et al</i> (2010)
Geminiviridae	ssDNA	Widespread	5 to 120	Bejarano <i>et al</i> (1996), Ashby <i>et al</i> (1997), Liu <i>et al</i> (2011),
Pararetrovirus	dsDNA	Widespread	A dozen to a thousand	Staginnus and Richert-Pöggeler (2006), Staginnus <i>et al</i> (2009), Chabannes and Iskra-Caruana (2013)

1.2 Pararetroviruses & endogenous pararetrovirus sequences (EPRVs)

Pararetrovirus was first named by Howard Martin Temin (1934 -1994) in 1985, referring to reverse-transcribing double-stranded DNA viruses including hepadnaviruses and caulimoviruses (Temin, 1985). As one of the major threats to plant production and human health, pararetroviruses were thought to have a distant relationship with long terminal repeat (LTR) retrotransposons, and plant pararetroviruses (*Caulimoviridae* family) were proposed to diverge from a same remote origin with the *gypsy* group of LTR retrotransposons (Xiong and Eickbush, 1990) (Figure 1.1). Currently the plant pararetrovirus comprises seven genera (virus taxonomy from ICTV, 2013). Their virus particles are isometric (*Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Solendovirus* and *Soymovirus*) or bacilliform (*Badnavirus* and *Tungrovirus*) (Figure 1.2). One pararetroviral particle contains a single molecule of dsDNA which forms a circle with discontinuities (open circle). Although plant pararetroviruses display various genomic structures, all of them encode a movement protein (MP), a capsid (or called coat) protein (CP) harboring a CCHC-type zinc finger, and a POL-like polyprotein comprising protease (PR), reverse transcriptase (RT) and RNase H (RH) but no integrase (Figure 1.2) (Hohn and Rothnie, 2013). Liking that of LTR retrotransposons and retroviruses, RH domain in pararetroviruses is associated with RT domain to form a multifunctional protein, called RT/RH. The intergenic region (IGR) of plant pararetroviruses is a diverse and noncoding region containing *cis*-regulation elements, and is important for viral transcription, translation and replication (Hohn and Rothnie, 2013). Highly diverse additional non-standard genes or open reading frames (ORFs) were also frequently found in plant pararetrovirus genomes (Figure 1.2) and their protein products play an important role in particle assembly, vector transmission and immune suppression (Hohn, 2013).

In the RT-mediated replication process of plant pararetroviruses, as briefly shown in Figure 1.3, the viral DNA is firstly transcribed into a pregenomic RNA in the nucleus,

and then the latter is exported to the cytoplasm, where the translated viral RT/RH (cleaved from a polyprotein by the PR) reverse transcribes the pregenomic RNA into a complementary DNA (cDNA) beginning at the primer binding site (PBS) using an initiator methionine tRNA as a primer; subsequently, the synthesis of plus DNA strand is initiated at the polypurine tract (PPT) (Hull, 2002a). Template switches are requested during the replication, which permit the RT/RH to switch from the 5' end to the 3' end of the pregenomic RNA and allow the synthesis of plus strand to continue at the strand discontinuities (Hull, 2002a; Hohn and Rothnie, 2013). After replication, the CP-derived molecules bind to the replicated DNAs to form unenveloped viral particles (Hohn and Rothnie, 2013). The MP is involved in viral cell-to-cell movement (Scholthof, 2005). All the known pararetroviruses encode all of the essential proteins and thus are autonomous during the viral life cycle in host cells.

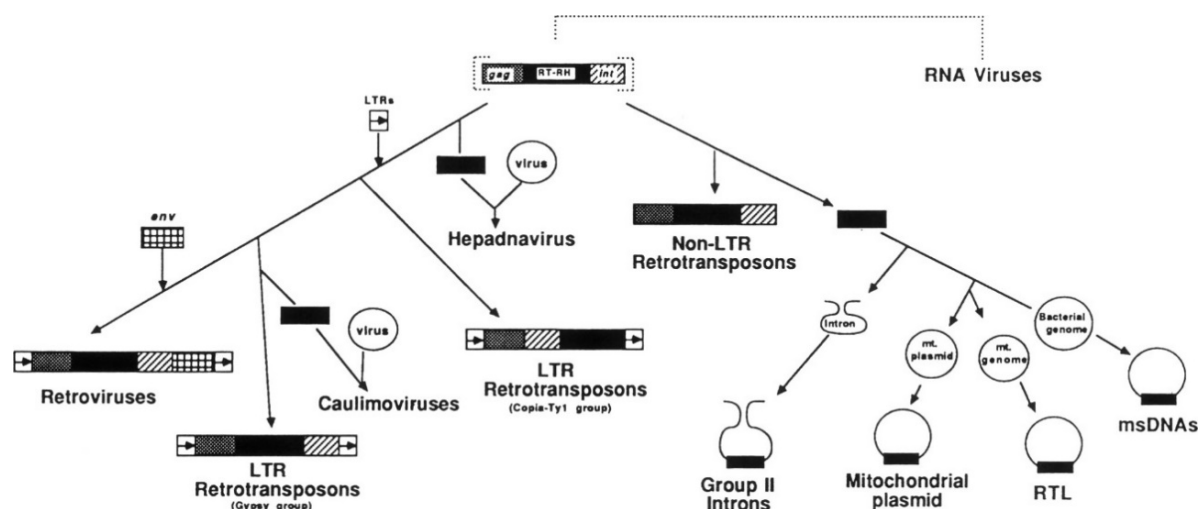


Figure 1.1 A scheme for the origin of retroelements including pararetroviruses (cited from Xiong and Eickbush, 1990)

Important structural features of each category of RT are presented. Shaded boxes correspond to domains of the ORF of the element, solid shading, RT region; stippled, gag region; diagonal shading, integrase domain; cross-hatched, envelope gene. LTRs present in certain elements are diagrammed as an open box with an arrow. For the hepadnaviruses, caulimoviruses, group II introns, the Mauriceville mitochondrial plasmid, RTL sequences and msDNA it is assumed that only a portion of the pol gene, containing the RT domain, entered an already existing element.

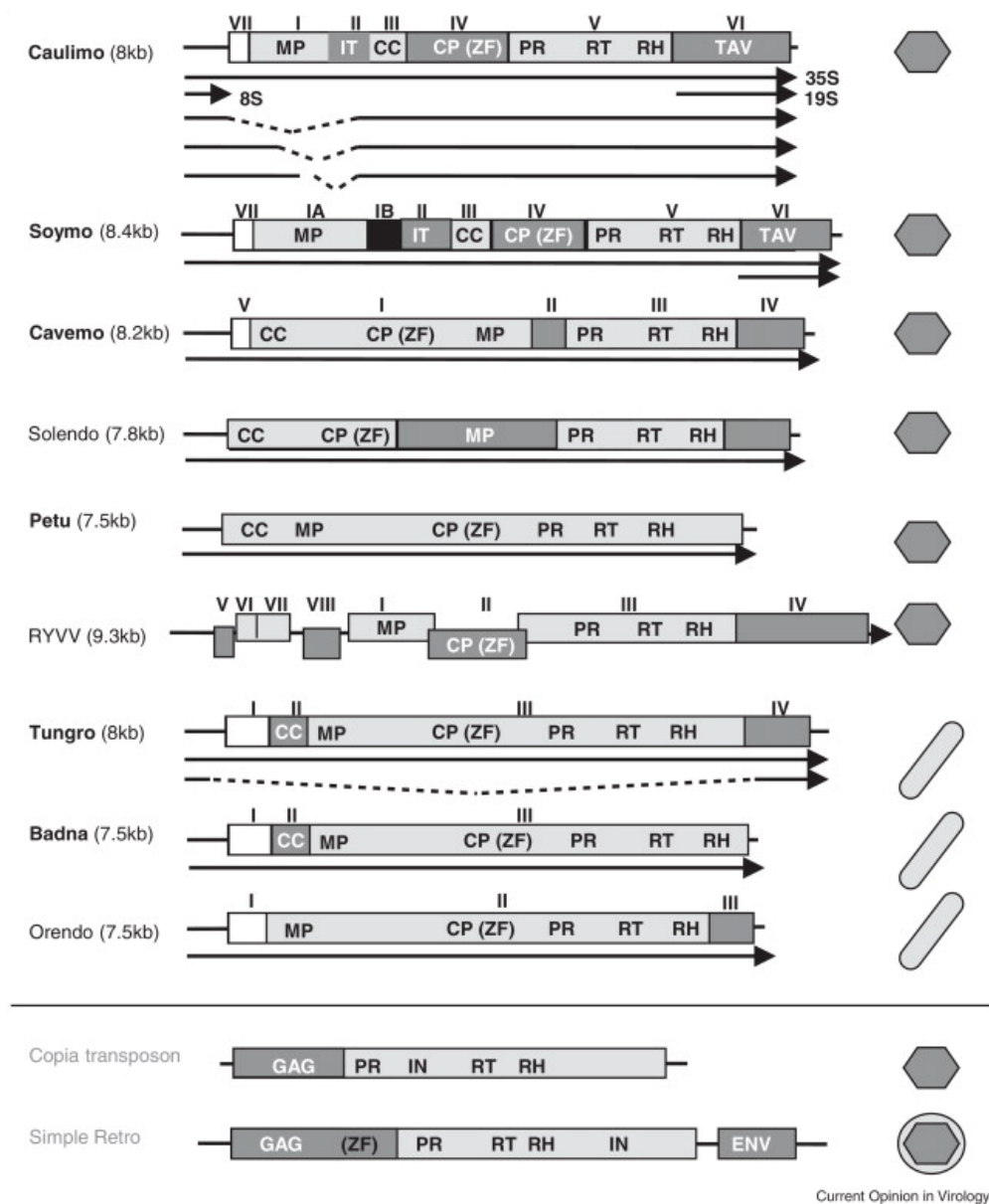


Figure 1.2 Linear genomic structures of typical plant pararetroviruses (cited from Hohn and Rothnie, 2013)

Arrows beneath each viral sequence indicate transcripts (dotted lines indicate introns). IT, insect transmission; CC, coiled coil; ZF, zinc finger; TAV, transactivator/viroplasmin; MP, movement protein; CP, coat protein; PR, protease; RT, reverse transcriptase; RH, RNase H. A canonical retrotransposon and a simple mammalian retrovirus are shown in the lower panel.

Although lacking the process and molecular machinery for integration (Temin, 1985; Harper *et al*, 2002), a growing number of endogenous pararetrovirus (EPRV) sequences have been discovered by virtue of the increasing decryption of plant genomes (Staginnus and Richert-Pöggeler, 2006; Hohn *et al*, 2008; Staginnus *et al*, 2009). Two teams simultaneously and independently reported the first EPRVs at 1998, which are two integrated sequences of banana streak badnavirus in the banana genome (Harper *et al*, 1999; Ndowora *et al*, 1999). Later, Jakowitsch *et al* (1999) reported the EPRVs in the tobacco genome and found thousands of coipes resulting from repeated integrations. Until now, EPRVs have been described in widespread plant species, such as Solanaceae family (Jakowitsch *et al*, 1999; Richert-Pöggeler *et al*, 2003; Gregor *et al*, 2004; Hansen *et al*, 2005; Staginnus *et al*, 2007), Musaceae family (Ndowora *et al*, 1999; Geering *et al*, 2001; Geering *et al*, 2005; Harper *et al*, 2005), Rutaceae family (Yang *et al*, 2003; Roy *et al*, 2014), *Dracaena sanderiana* (Su *et al*, 2007), *Dahlia variabilis* (Pahalawatta *et al*, 2008; Almeyda *et al*, 2014), *Ananas comosus* (Gambley *et al*, 2008; Sether *et al*, 2012), *Vitis vinifera* (Bertsch *et al*, 2009), *Populus trichocarpa* (Bertsch *et al*, 2009), *Ficus carica* (Laney *et al*, 2012), *Fritillaria imperialis* (Becher *et al*, 2014), *Dioscorea cayenensis* (Seal *et al*, 2014; Umber *et al*, 2014). The EPRVs of all the genera of Caulimoviridae family (*Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Solendovirus*, *Badnavirus* and *Tungrovirus*) have been found, except for *Soymovirus* (Chabannes and Iskra-Caruana, 2013). Recently, the EPRVs of animal pararetroviruses (*Hepadnaviridae* family) were identified in the multiple genomes of birds (Gilbert and Feschotte, 2010; Katzourakis and Gifford, 2010) and snakes (Gilbert *et al*, 2014).

EPRVs were thought to be generated from the incidental integrations of pararetrovirus sequences during their invasions into host germline cells, and these incidental integrations could involve illegitimate recombination or double-strand break (DSB) repair via non-homologous end joining (NHEJ) between the pararetrovirus and host genomes (Bill and Summers, 2004; Staginnus and Richert-Pöggeler, 2006; Liu *et*

al, 2012). EPRVs are mostly degenerated and non-infectious due to mutational decay within host genomes, and are generally silenced by host epigenetic regulation such as DNA and histone methylation (Staginnus and Richert-Pöggeler, 2006; Noreen *et al*, 2007). Only three cases of EPRV-derived infections have been reported in respective hybrids of tobacco, petunia and banana (reviewed by Chabannes and Iskra-Caruana, 2013). In these cases, EPRVs in a hybrid genome recombined to reconstruct a full length viral genome (Iskra-Caruana *et al*, 2010), or formed tandem repeats to allow the transcription of pregenomic viral RNA that was then reverse transcribed into viral DNA genomes (Richert-Pöggeler *et al*, 2003).

EPRVs are valuable viral genomic fossils to allow us to reconstruct the deep history of pararetrovirus evolution (Gifford and Tristem, 2003; Emerman and Malik, 2010; Holmes, 2011; Feschotte and Gilbert, 2012). In addition to a viral invasion record, EPRVs could provide novel genetic sources and induce genomic rearrangements and epigenetic variations in host genomes (Hohn *et al*, 2008). In some of the plant genomes, EPRVs are partly or even predominantly located in centromeric/pericentromeric regions, suggesting that EPRVs may play a role in the evolution of centromere sequences of certain plant species (Richert-Pöggeler *et al*, 2003; Hansen *et al*, 2005; Staginnus *et al*, 2007; Becher *et al*, 2014). Furthermore, EPRVs might confer the resistance to exogenous virus infections by homology-dependent gene silencing (Staginnus and Richert-Pöggeler, 2006; Chabannes and Iskra-Caruana, 2013). The study of EPRVs is contributing to the emerging paleovirology of pararetroviruses, which provides novel and deep insights into the origin and evolution of pararetroviruses, and into the interaction and co-evolution between pararetroviruses and hosts (Emerman and Malik, 2010; Patel *et al*, 2011).

1.3 Rice tungro bacilliform virus & endogenous rice tungro bacilliform virus-like sequences (eRTBVLs)

Rice tungro bacilliform virus (*Caulimoviridae*, *Tungrovirus*; RTBV) is the only known extant pararetrovirus that infects cultivated rice (*Oryza sativa* L.). The dsDNA genome of RTBV has one discontinuity in each strand and is assembled by the CP molecules into bacilliform particles (Hull, 1996). The RTBV genome consists of one IGR and four ORFs, ORF1 to ORF4 (Hay *et al*, 1991; Qu *et al*, 1991). The protein production of ORF1 could be related to particle assembly (Hay *et al*, 1994). This ORF is special that ATT start codon (the second ATT codon, no ATG start codon) initiates the translation of the corresponding transcript (Fütterer *et al*, 1996). The protein of ORF2 is suggested to participate in capsid assembly (Herzog *et al*, 2000). The ORF3 encodes a polyprotein including MP, CP, PR and RT/RH. The protein of ORF4 can suppress the cell-to-cell spread of host antiviral silencing (Rajeswaran *et al*, 2014). The RTBV infects rice plants through the vector *Nephotettix virescens* (green leafhopper), however, this vector transmission requires the help from a ssRNA virus, rice tungro spherical virus (*Secoviridae*, *Waikavirus*, RTSV) (Hull, 1996). The coinfection of RTBV and RTSV induces rice tungro disease (Figure 1.4) which is a significant constraint on rice yield in South and South East Asia (Hay *et al*, 1991; Qu *et al*, 1991). The symptoms of rice tungro disease include young leaf discoloration (yellowing or orangening) and stunted growth, which are primarily caused by RTBV (Hull, 1996).

The Plant Breeding Laboratory of Hokkaido University found the EPRVs in the rice genomes at 2004, which are most similar to RTBV, therefore referred to as endogenous RTBV-like sequences (ERTBVs or eRTBVLs) (Kunii *et al*, 2004; Liu *et al*, 2012). Based on the similarity with RTBV, viral genome of eRTBVLs is putatively composed of one IGR and three ORFs (ORFx, y and z), among which the longest ORFy encodes presumed MP, CP, PR and RT/RH (Figure 1.5) (Kunii *et al*, 2004). The examined eRTBVL sequences in the Nipponbare genome can be divided into three clusters based

on the phylogenetic tree of RT region (Kunii *et al*, 2004). The spatio-temporal and phylogenetic origins of eRTBVLs are still a mystery. Different from the other reported EPRVs, eRTBVLs are often flanked by AT di-nucleotides repeats in rice genomes (Kunii *et al*, 2004). Liu *et al* (2012) used an *in silico* approach to prove that the integrations of eRTBVLs tend to occur in AT repeats which could be a hotspot of DSB (Figure 1.6). Southern blotting results suggested that similar sequences of eRTBVLs also exist in certain *Oryza* AA-genome species (Kunii *et al*, 2004). These sequences in different *Oryza* AA-genome species exhibit variations in copy number and methylation level (Kunii *et al*, 2004).

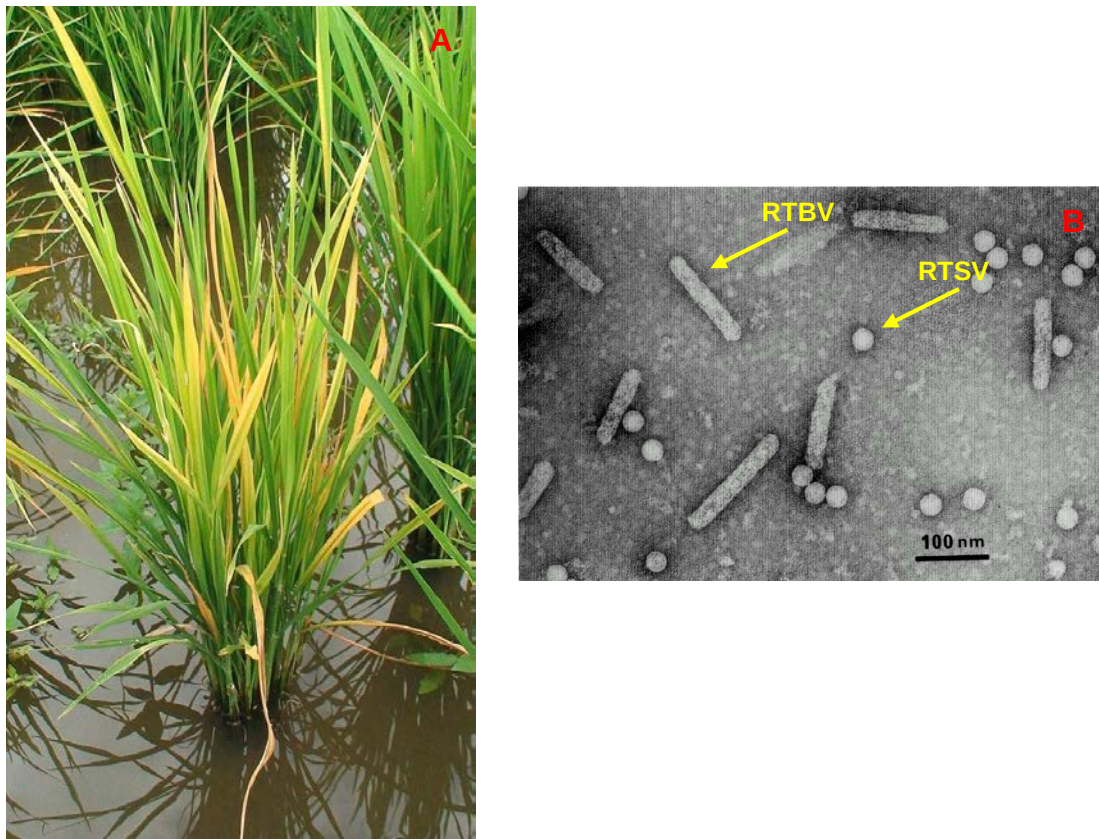


Figure 1.4 Rice tungro disease symptoms (A) and pathogens (B) (cited from IRRI Rice Knowledge Bank: www.knowledgebank.irri.org)

Chapter 2 Multiple origins of endogenization of eRTBVL sequences during rice speciation and the corresponding virus genealogy

2.1 Introduction

Rice tungro disease, a significant constraint on rice yields in South and South-East Asia, is caused primarily by infection with RTBV, the known extant rice pararetrovirus (Hay *et al*, 1991; Qu *et al*, 1991). The rice EPRVs (eRTBVLs), which share similarity with RTBV sequences, have been found in the cultivated rice genomes (Kunii *et al*, 2004; Liu *et al*, 2012). Intact virus genome sequences were assembled successfully by the eRTBVL fragments collected from the Nipponbare genome database (Kunii *et al*, 2004). Natural virus fossils are difficult to obtain, but the vertical inheritance of EPRV sequences, such as eRTBVLs, has provided an unexpected opportunity for genetic and evolutionary study of pararetroviruses on a long-term scale. The age of an endogenization event can be estimated by examining the orthologous EPRV insertions at a specific locus among host species with known phylogenetic relationships (Gayral *et al*, 2010; Gilbert and Feschotte, 2010). In addition, independent viral invasion events in different host populations generated distinct EPRV endogenization patterns (Chessa *et al*, 2009). Like paleontological fossils, EPRVs embedded in the host genome help in deducing the activities of ancient viruses, their geographical origin, and their integration history based on phylogenetic and geographical data of the host.

The evolutionary history of *Oryza* species, especially *O. sativa*, has been studied deeply and described. The speciation of *Oryza* AA-genome species occurred more than 5 million years ago (MYA) and shared a common progenitor with *Oryza* BB-genome species (Zou *et al*, 2008; Zou *et al*, 2013; Zhu *et al*, 2014). As one of *Oryza* AA-genome species, *O. sativa* (Asian cultivated rice, including two subspecies *indica* and *japonica*) was domesticated from the wild rice species *O. rufipogon* (Oka, 1988; Khush, 1997; Kovach *et al*, 2007; Huang *et al*, 2012). *O. rufipogon* is thought to have diverged into

at least two ecotypes, perennial and annual (the annual type is also called *O. nivara*), 0.16 MYA, as estimated by the molecular clock (Zhu *et al*, 2007; Zheng and Ge, 2010). Each ecotype contains geographically diverse populations. *Japonica* was domesticated first from a perennial population of *O. rufipogon* in southern China approximately 10,000 years ago (YA); subsequently, it was crossed with local annual populations of *O. rufipogon* in South East and South Asia to breed *indica* (Jiang and Liu, 2006; Londo *et al*, 2006; Huang *et al*, 2012; Wei *et al*, 2012; Yang *et al*, 2012). The eRTBVL profiles in the *japonica* and *indica* genomes (88 and 74 segments respectively) are distinct; however, based on their flanking sequence alignments, some of them overlap (22 shared segments) (Liu *et al*, 2012). Currently, three rice genomic databases, one for each of the cultivated subspecies, *japonica* and *indica*, and one for the wild species, *O. rufipogon*, are publicly available.

Here, I compared the eRTBVL sequences from the three rice genome databases by comprehensive *in silico* analyses and screened orthologous sequences in diverse cultivated and wild rice accessions, to understand the evolutionary route of the eRTBVL sequences. My data revealed that the eRTBVL sequences occurred in multiple origins associated with rice speciation, and that the divergence between RTBV and the virus of eRTBVL occurred long before the immediate progenitor of *O. rufipogon*. The rice genomes also have recorded the evolutionary dynamics of rice pararetroviruses driven by a limited number of homologous recombinations of viral sequences, providing a valuable fossil record of genetic recombination of paleoviruses. The potential impacts of the recombinations and the geographic relationships in ancient pararetrovirus evolution are discussed.

2.2 Materials and Methods

2.2.1 Plant materials and taxonomy

The global rice core collection (69 accessions, including *japonica* and *indica*; 10

accessions were excluded here because of considerable missing polymerase chain reaction (PCR) data) was obtained from the National Institute of Agrobiological Sciences (NIAS), Japan. This collection minimizes the accession number required to maintain genetic diversity of all the accessions at the NIAS Genebank (Kojima *et al*, 2005). Six additional cultivated accessions maintained by the Laboratory of Plant Breeding, Hokkaido University, were added to the collection (finally 65 cultivated rice accessions). The wild rice collection (30 *O. rufipogon* accessions) and the other *Oryza* species accessions were provided by the National Institute of Genetics, Japan. Detailed accession information is presented in Supplementary Table S1. All the seeds were sterilized, germinated, and planted in greenhouses at Hokkaido University. The taxonomy of these accessions was reconstructed based on the eRTBVL polymorphisms detected in this study (see orthologous eRTBVL screening described below) using the unweighted pair-group method with arithmetic mean (UPGMA) in the SAHN module of the NTSYS-pc 2.1 package (Rohlf, 2000).

2.2.2 Genomic data mining

The genomic data informations of *Oryza* species were listed in Supplementary Table S2. To identify eRTBVLs in W1943, a BLASTN search was performed using the BLAST+ 2.2.27 application (Camacho *et al*, 2009) with the previously assembled representative eRTBVL sequences from the Nipponbare genome (DDBJ accession numbers: BR000029-BR000031) (Kunii *et al*, 2004) as queries. The following BLAST parameters were used: word size, 11; gap open, 5; gap extend, 2; penalty, -3; and reward, 2. The contigs spanning high-identity matches (e-values < 1e-10, alignment length > 100 bp) were extracted, and filtered so that only the contigs with > 50 bp flanking at least one end of the eRTBVL loci were kept for mapping. Mapping of the W1943 eRTBVLs onto the *japonica* and *indica* genomes by BLASTN was based on a maximum of 10 kb of flanking sequence(s) of each eRTBVL locus in the filtered contigs. The unambiguously mapped eRTBVLs were compared with their counterparts

in the *japonica* and *indica* genomes (Liu *et al*, 2012), and those that mapped to unique genomic loci were retrieved. The screening of orthologous eRTBVL-D loci in the other *Oryza* genomes was also based on the flanking sequences mapping performed in the W1943 genome.

2.2.3 Phylogenetic analysis

The phylogenetic trees were constructed based on the defined regions in the eRTBVL sequences. All the segments with lengths > 80% of the lengths of each of the regions were retrieved from the Nipponbare, 93-11 and W1943 genomes. The sequences were aligned using ClustalW (Thompson *et al*, 1994) followed by manual editing, and ambiguous regions and missing data were removed. Maximum likelihood (ML) phylogenetic analysis was performed by MEGA 5.05 software (Tamura *et al*, 2011). The best-fitting substitution models for each region were determined in MEGA 5.05 (TN93+G model for the MP, CP and PR regions; T92+G model for the RT/RH region, ORFz and IGR; T92 model for ORFx). ML analysis of the eRTBVL families and RTBV strains was based on the amino acid sequences of the RT/RH region using the WAG+G model, which was chosen after model testing. Support for the ML trees was evaluated by 1000 bootstrap replicates.

2.2.4 Endogenization pattern analysis

To maximize the number of eRTBVL segments and at the same time exclude the short ones with ambiguous taxonomy, all the sequences with lengths > 60% of the lengths of the RT/RH region, ORFz and IGR (the ML trees based on these regions showed clear distant clusters of the eRTBVL-A, -B and -C families) were retrieved from the three rice genomes for the endogenization pattern analysis. To estimate the relative abundance of the eRTBVL-A, -B and -C families in the three genomes, the segments that belonged to the eRTBVL-D and -X families or had ambiguous taxonomy were not taken into account. In the same way, I also compared the relative abundance

of the species-specific eRTBVLs in the three families after removing the known species-shared elements.

2.2.5 Orthologous eRTBVL screening by primer sets

Total DNAs were extracted from the leaf samples of the cultivated and wild rice accessions by Plant DNAzol (Invitrogen, Carlsbad, CA, USA). The DNA concentration in each sample was measured by NanoDrop 2000 (Thermo Fisher Scientific, Wilmington, DE, USA) and all were adjusted to a similar level. Primer pairs were generally designed and used as follows: for amplification 1, a forward primer for the left flank (P1) and a reverse primer for the interior of the eRTBVL locus (P2) were used; for amplification 2, a forward primer for the interior of the eRTBVL locus (P3) and a reverse primer for the right flank (P4) were used; for amplification 3, a primer pair for left and right flanks (P1, P4) were used to amplify the whole length of the eRTBVL locus or the empty donor site. PCR were performed using Ex Taq or LA Taq polymerase (Takara, Shiga, Japan) in a PTC-200 thermal cycling system (GMI, Ramsey, MN, USA). When the whole length of some long eRTBVLs failed to amplify, the reaction was adjusted to amplify the possible empty donor sites in the corresponding rice accessions. The PCR products were resolved on a 1-2% agarose gel and analyzed with a Typhoon 8600 PhosphorImager (GE Healthcare, Little Chalfont, U.K.). The presence/absence of orthologous eRTBVLs in the rice accessions was estimated based on the genomic PCR results using the primer sets (for part of the eRTBVL loci, one or two of the four primers, P1-P4, failed in the amplifications).

2.2.6 Endogenization order analysis

Because the eRTBVL-A, -B and -C sequences formed clear distinct clusters in the ML trees based on the RT/RH region, ORFz and IGR, sequence alignments of these three regions were constructed for each of the eRTBVL-A (-A1 and -A2), -B, and -C families. The single-nucleotide-polymorphisms (SNPs) in these alignments were of

two types: sole-SNPs that were found solely in a single sequence within each alignment, and multiple-SNPs that were found in more than one sequence within each alignment. The multiple-SNPs were largely rooted in the divergence of the viral populations of eRTBVLs prior to integration, whereas the sole-SNPs were largely rooted in the mutation accumulation of eRTBVLs after endogenization. To estimate the temporal order of endogenization events of eRTBVL-A, -B, and -C, I extracted sole-SNPs from each eRTBVL family alignment. Sites with missing data or with gaps in the alignment were excluded. I constructed pseudo-sequences in which the sole-SNPs replaced the initial bases in the consensus (ancestral) sequences (consensus sequence construction was described below) of the RT/RH region, ORFz and IGR of the corresponding eRTBVL families. The pairwise distances between the pseudo-sequences and the original consensus sequences were assessed using MEGA 5.05 with the following parameters: maximum composite likelihood, inclusion of transitions + transversions, gamma distributed among sites. Finally, the pairwise distances were divided by the corresponding sample size of the sequences used for sole-SNP calling. Because some species-shared elements were present in the alignments, I ignored some of the sole-SNPs (the sole-SNPs of the eRTBVLs from a single genome were shared by orthologous sequences in other genomes) in the above divergence analysis. Therefore, I constructed alignments for each eRTBVL family using only the sequences from a single genome, and repeated the sole-SNP calling and mean pairwise distance computation described above.

2.2.7 Recombination analysis

Consensus sequences of the eRTBVL-A1, -A2, -B, -C and -X families were acquired by aligning the sequences from the defined regions for each family (the chimeric sequences that possibly resulted from intra-recombination after integration were excluded). The eRTBVL sequences that spanned at least 80% of the length of each region in the three rice genomes were used in the alignments. The consensus sequences

of each region were combined, and ambiguous regions at the terminals were trimmed to produce the final whole length consensus sequences. These consensus sequences were then aligned by ClustalW and manually edited for use in the recombination analysis. I employed the RDP, GENECONV and BOOTSCAN methods in the RDP3 package (Martin *et al*, 2010) with the default settings, to detect recombination signals. Potential recombination events were rechecked and visualized using the RDP method (windows size, 30 bp). The constructed consensus sequences of eRTBVL families have been submitted to DDBJ: accession numbers, BR001195-BR001199.

2.3 Results

2.3.1 The eRTBVL sequences in the genome of wild rice

To investigate the origin of the endogenous pararetrovirus elements in the rice genome, I performed a strict *in silico* BLAST search (e value < 1e-10) of the eRTBVL sequences against a recently released draft genome of the specific *O. rufipogon* line (W1943) (Huang *et al*, 2012) that is distributed in eastern China (Lu *et al*, 2008). Among the 1607 contigs that were found to be harboring eRTBVL segments, the 206 contigs with > 50 bp that flanked at least one end the eRTBVL sequences (comprising more than 103 eRTBVL loci in this *O. rufipogon* genome) were extracted and mapped to the *japonica* (Nipponbare) and *indica* (93-11) genomes for further examination. After mapping, 57 of the eRTBVL elements in W1943 were anchored successfully with unique and unambiguous matches to the Nipponbare and 93-11 genomes, including the elements that were shared with Nipponbare and/or 93-11 (species-shared eRTBVLs), and W1943-specific eRTBVLs (Supplementary Table S3).

Among these 57 eRTBVL elements in the W1943 genome, 17 were W1943-specific and therefore absent in the Nipponbare and 93-11 genomes, meaning that only their flanking sequences could be mapped into the *O. sativa* genomes (Supplementary Table S3). For the species-shared eRTBVL elements, 17 were found in the same loci in both

the Nipponbare and 93-11 genomes, and 19 and four were found in either the Nipponbare or 93-11 genomes, respectively (Supplementary Table S3). The eRTBVL loci that were shared in the Nipponbare and 93-11 genomes each originated from one insertion (Liu *et al*, 2012). Thus, these species-shared eRTBVLs (orthologous eRTBVLs) must have been inherited from progenitor genomes, implying that most of the eRTBVL elements were integrated into the host genomes before rice domestication.

2.3.2 Organization of the eRTBVL families

A phylogenetic analysis based on the conserved RT/RH region showed that the eRTBVL sequences from the three rice genomes (Nipponbare, 93-11, and W1943) clustered into four apparent families (Figure 2.1). Three of the families were consistent with the three groups, eRTBVL-A, -B and -C, from the Nipponbare genome, described in the previous study (Kunii *et al*, 2004). These three eRTBVL families each comprised species-shared and species-specific elements from the three rice genomes (Figure 2.1). The fourth newly identified family comprised the least members and was named eRTBVL-D (Figure 2.1). The branches in the eRTBVL-D cluster were divergent and far distant from the other three families; in particular, each of the eRTBVL-D sequences had orthologs in all the three genomes (Figure 2.1, Supplementary Table S3), indicating that eRTBVL-D could be the oldest family among them. Because of the apparent clustering that was observed within the eRTBVL families, the four distinct families must have originated from corresponding ancient pararetrovirus lineages and not from differentiation after their integration into the genome. In addition to the four families, I found another eRTBVL cluster, named eRTBVL-X, which comprised 10 eRTBVL segments with a nearly 94-kb stretch between Os08g0282500 and Os08g0285200 loci on chromosome 8 in the Nipponbare genome (Figure 2.2) (Liu *et al*, 2012).

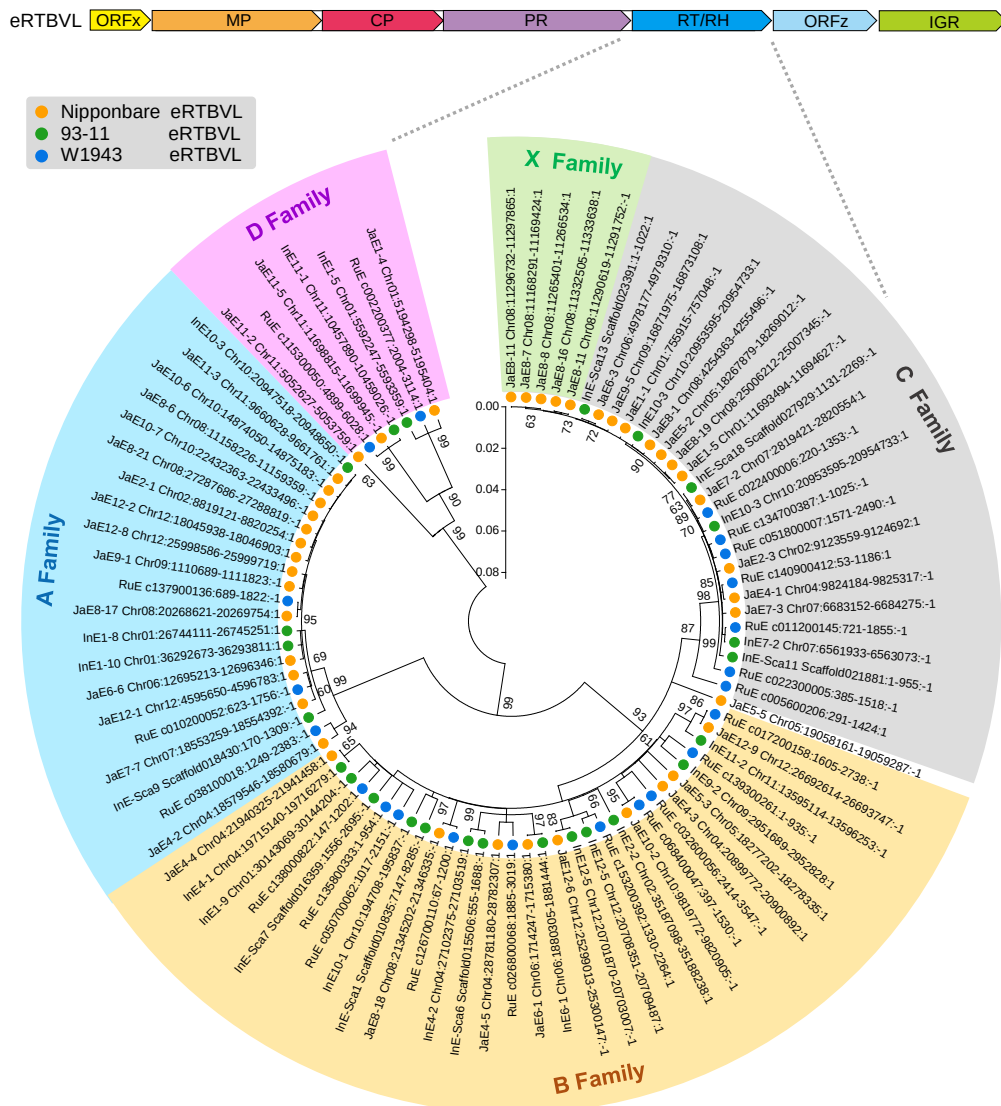


Figure 2.1 eRTBVL families identified by a phylogenetic analysis based on the RT/RH region in eRTBVL sequences

The ML tree was constructed using the nucleotide sequences of the RT/RH region of eRTBVLs from the *O. rufipogon* (W1943), *japonica* (Nipponbare) and *indica* (93-11) genomes (only sequences with lengths > 80% of the length of RT/RH consensus sequence were used; therefore, the orthologous sequences of JaE11-2 that I had collected were not included because they contained deletions). The genomic locations of the eRTBVL sequences are shown in each taxon named as ID, chromosome/contig, position, and strand. The rice genomes to which the sequences belong are labeled with colored solid circles in front of the taxon names (orange, Nipponbare; green, 93-11; blue, W1943). Sequences sorted into the same family are indicated by the same background color. The eRTBVL-X sequences are indicated. The bootstrap consensus tree inferred from 1,000 replicates is displayed in the center of the circle, and bootstrap values greater than 60 are shown above the branches. The tree is midpoint rooted and drawn to scale, with branch lengths measured by the number of substitutions per site. The schematic illustration at the top of the figure shows the structure of the eRTBVL sequences; RT/RH region is indicated.

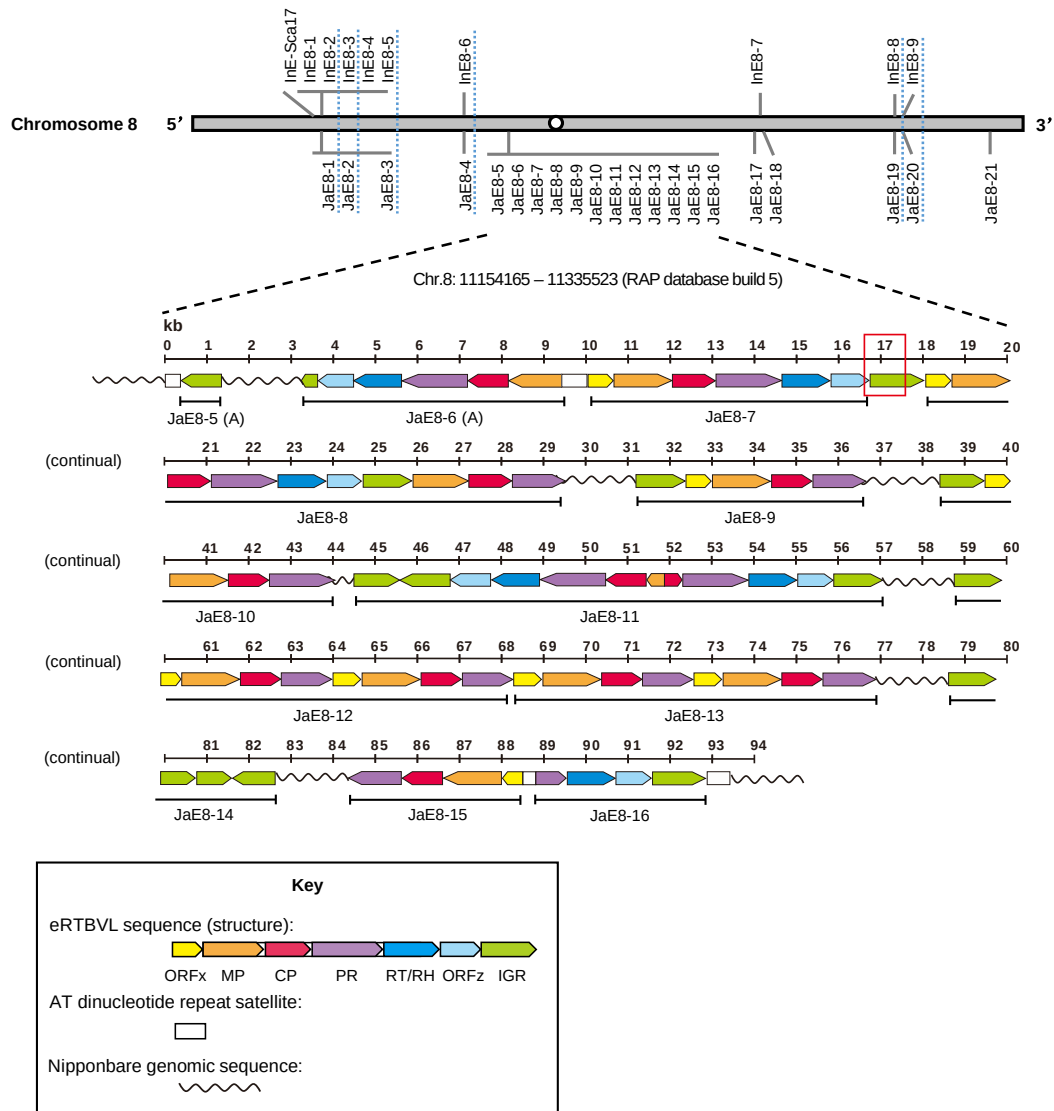


Figure 2.2 Structure of the eRTBVL-X cluster on chromosome 8 of Nipponbare

The gray horizontal bar at the top represents chromosome 8 of rice; the circle indicates the centromere. The distributions of the eRTBVL loci on this chromosome in the 93-11 and Nipponbare lines are indicated above and below the bar respectively, according to Liu *et al.* (2012). Orthologous loci between the two lines are marked by the blue dotted lines on the right side of the eRTBVL IDs. Details of the eRTBVL-X cluster in Nipponbare are expanded. The sequences corresponding to seven regions in the eRTBVL sequences are colored (yellow, ORFx; orange, MP; red, CP; purple, PR; dark blue, RT/RH; light blue, ORFz; green, IGR). Rice genomic sequences are indicated by wavy lines, and AT repeat satellites are indicated by a white rectangle. The region of the sequence in the red box is uncharacterized in the database where it is represented by a gap denoted by the letter N. This region was sequenced for the first time in this study. The eRTBVL segments (from JaE8-5 to JaE8-16) defined by Liu *et al.* (2012) in this cluster are indicated under the sequence using their annotation codes. These eRTBVL segments were sorted into the eRTBVL-X family based on the phylogenetic and recombination analyses; JaE8-5 and JaE8-6 (labeled in brackets) are exceptions and belong to the eRTBVL-A family.

2.3.3 Phylogenetic relationships of eRTBVL sequences

Separate alignments were performed for each of the seven defined regions (ORFx, MP, CP, PR, RT/RH, ORFz, and IGR) in the eRTBVL sequences. These alignments were used to construct phylogenetic trees based on each of these regions. The topologies of five of the trees, those for ORFx, MP, CP, PR, and RT/RH, were similar to each other, while the topologies of the trees for ORFz and IGR were different (Figure 2.3; details in Figure 2.1 and Supplementary Figures S1-S6). In the five topologically similar trees eRTBVL-B was close to eRTBVL-C, but in the two trees based on ORFz and IGR, eRTBVL-B was a sister to eRTBVL-A. The sequences in the eRTBVL-A family were split into two subfamilies, eRTBVL-A1 and -A2, by a phylogenetic analysis based on the IGR, which is a diverse intergenic noncoding region (Figure 2.3, Supplementary Figure S6). The sequence divergence after endogenization for the IGR of eRTBVL-A1 was larger than that for the IGR of eRTBVL-A2 (Table 2.1, see below), which implies that the corresponding virus sequence of the eRTBVL-A1 subfamily may have been the initial genotype and that the virus sequence of the eRTBVL-A2 subfamily was derived from it.

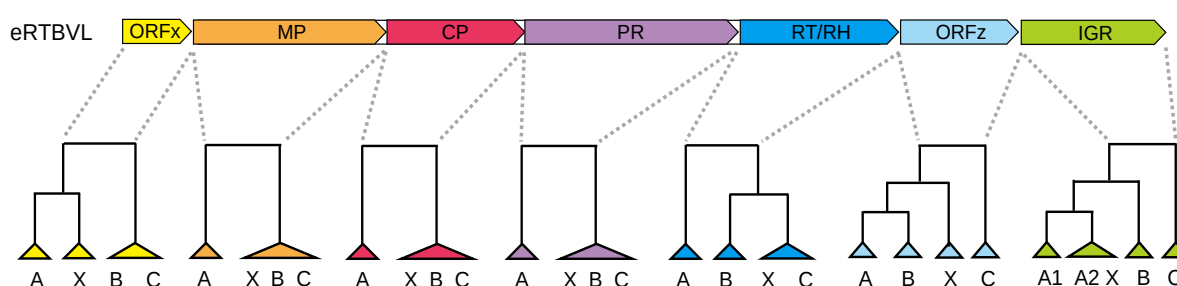


Figure 2.3 Global view of phylogenetic relationships based on the sequences of seven regions of the eRTBVL sequences

The eRTBVL structure is shown schematically with the summarized phylogenetic trees (midpoint rooted for display purposes) for each region displayed beneath. The abbreviated names of the eRTBVL families are shown below the branches. The complete ML phylogenetic trees for each region are presented in Figure 2.3; details in Figure 2.1 and Supplementary Figures S1-S6.

2.3.4 Proportions of eRTBVL families in the three rice genomes

Each of the *japonica*, *indica*, and *O. rufipogon* genomes harbored about 100 copies of the eRTBVL-A, -B, -C and -D families (this study, and Liu *et al*, 2012), but in different proportions. I examined the endogenization patterns of the eRTBVLs in each of the three genomes by comparing their relative abundances in the eRTBVL-A, -B and -C families (Figure 2.4). In the Nipponbare (*japonica*) genome, the eRTBVL-A family showed a significantly higher abundance in the genome than the other two families, and also a higher abundance than the eRTBVL-A family in the other two genomes (Student's t-test, all $p < 0.0124$). In the 93-11 (*indica*) genome, eRTBVL-B showed a higher abundance than the other two families, and also a higher abundance than its counterparts in the other two genomes (all $p < 0.0254$). In the W1943 (*O. rufipogon* in eastern China) genome, eRTBVL-C was the most abundant family (all $p < 0.0495$). I also compared the relative abundances of the species-specific elements in the three families after removing the known species-shared elements, and observed the same trend for the proportions of eRTBVL families in the cultivated rice genomes (Figure 2.5). Thus, the most abundant eRTBVL family in a representative rice genome implies an abundance of the corresponding viruses in the habitats of the host rice species and the dominant geographic prevalence of the virus in the time period.

The major progenitor of *indica* was distributed in South East and South Asia, the geographical origin of *japonica* was the Pearl River area in China (Huang *et al*, 2012; Wei *et al*, 2012), and the specific *O. rufipogon* W1943 line is spread around the Lower Yangtze region in China (Lu *et al*, 2008). Accordingly, it was inferred that the viruses that became members of the eRTBVL-B family were prevalent in South East and South Asia, and the viruses that became members of the eRTBVL-A and eRTBVL-C families were prevalent in southern and eastern China, respectively.

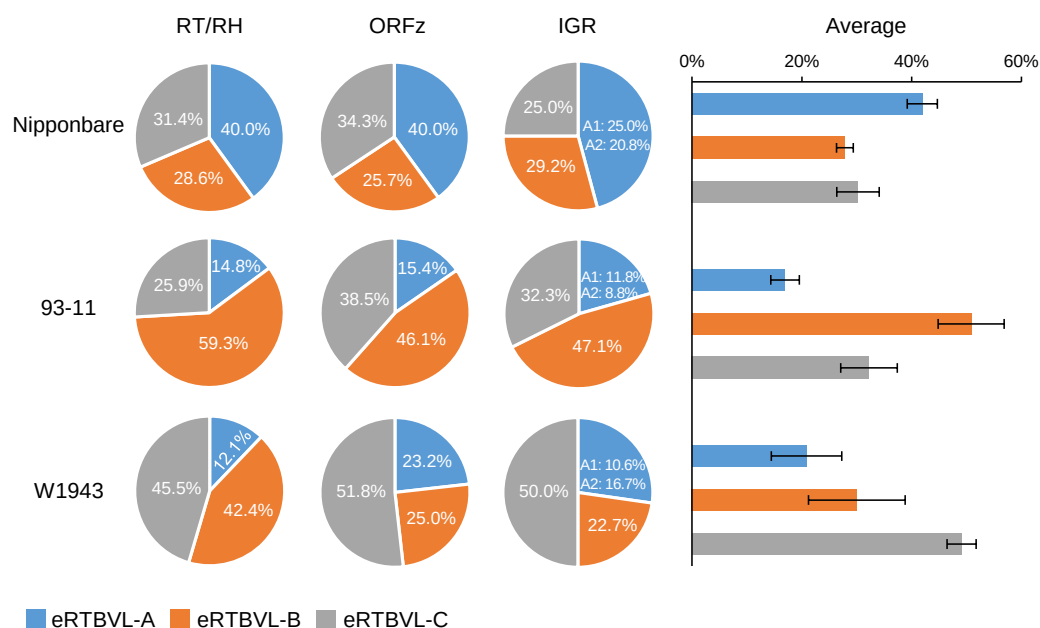


Figure 2.4 Endogenization patterns of three eRTBVL families

Relative abundances of the eRTBVL-A, -B and -C families in each rice genome are indicated in the pie charts. For the RT/RH, ORFz, and IGR regions, the proportions were calculated based on the sequence counts (the sequences all covered >60% of the length of consensus sequence for each region). The average values for the relative abundances are shown in the bar charts. The error bars indicate $\pm$ SEM (standard error of measurement).

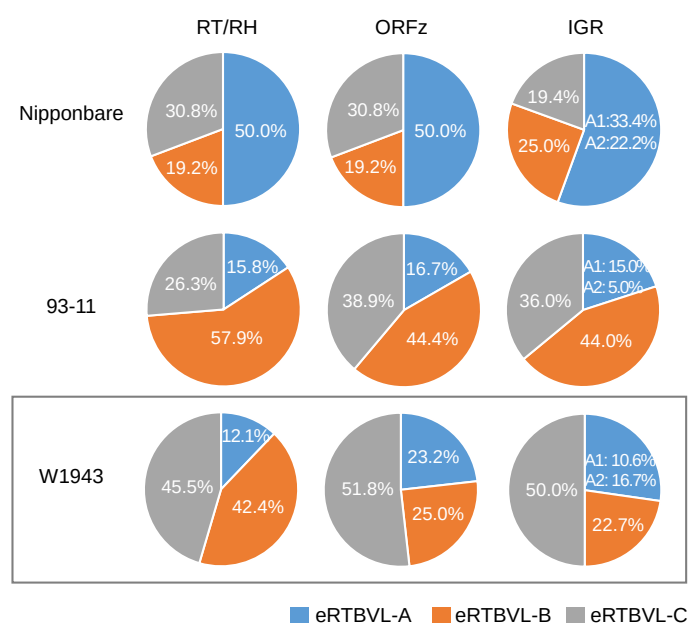


Figure 2.5 Endogenization patterns of three eRTBVL families for Nipponbare- and 93-11-specific eRTBVLs

Relative abundances of the eRTBVL-A, -B and -C families in each rice genome are indicated in the pie charts. For the RT/RH, ORFz and IGR regions, the proportions were calculated according to the sequence counts (sequences that covered >60% of the length of consensus sequence of each region were used).

2.3.5 Distribution of eRTBVL sequences in cultivated and wild rice accessions

To examine the relationship between rice ecotypes and the eRTBVL families, orthologous sequences of the eRTBVL loci that were identified in the three rice genome databases were further explored in a core collection of 65 cultivated (16 *japonica* and 49 *indica*) and 30 wild accessions (Figure 2.6). I screened orthologous eRTBVL insertions by PCR with plural primer pairs that were designed to amplify the eRTBVL sequence interior to the flanking sequences. Because of the abundant repetitive sequences in the flanking regions of a number of eRTBVL loci, the orthologs of only 34 eRTBVL loci, representing Nipponbare/93-11-shared and -specific loci and W1943-specific loci, were amplified successfully (Figure 2.6). All six of the eRTBVL-D elements that I examined had been fixed in almost all the cultivated and wild accessions in the collection, which confirmed eRTBVL-D as the most ancient family. Seven Nipponbare/93-11-shared eRTBVL-A, -B, and -C elements, including unsigned elements, were distributed widely in the cultivated rice accessions and in the Chinese *O. rufipogon* accessions, but were absent in the South East and South Asia accessions of *O. rufipogon* (Figure 2.6). Twelve Nipponbare-specific eRTBVLs were detected in most of the *japonica* accessions and in the Chinese accessions of *O. rufipogon*, but not in South East and South Asia accessions of *O. rufipogon* (Figure 2.6). In contrast, five 93-11-specific elements were observed in a wide range of *indica* accessions and the *O. rufipogon* accessions from South East, South Asia, and China, but were absent in the *japonica* accessions (Figure 2.6). Four W1943-specific elements were absent in all other accessions (Figure 2.6). Hence, the eRTBVL pool for *japonica* might have originated from a Chinese *O. rufipogon* population (but not from the line of W1943), and the eRTBVL pool for *indica* could be mixed with those of *O. rufipogon* populations from South East and South Asia in addition to China. Interestingly, the Nipponbare/93-11-shared elements (except the eRTBVL-D elements) could have been introduced into both *japonica* and *indica* cultivars only through Chinese *O. rufipogon* (perennial) lineages, and not through *O. rufipogon* (annual) lineages in South East or South Asia,

which did not possess these Nipponbare/93-11-shared eRTBVL elements. The eRTBVLs (except eRTBVL-D) from *O. rufipogon* (annual) lineages in South East and South Asia were distributed mostly toward *indica* cultivars. This flow of eRTBVLs is in accordance with the history of rice domestication (Huang *et al*, 2012).

The *in silico* analysis showed that the eRTBVL-X elements formed a cluster consisting of 10 sequence segments with a nearly 94-kb stretch on chromosome 8 in the Nipponbare and W1943 genomes (Figure 2.2, Supplementary Table S3), in which eRTBVL-A and -C were the dominant families, respectively (Figure 2.4). I performed a genomic PCR to validate and detect orthologous insertions of the eRTBVL-X elements in the cultivated and wild rice collections, as described above. The eRTBVL-X cluster was detected in seven accessions of *japonica* and in all Chinese accessions of *O. rufipogon* (Supplementary Table S4). These observations suggest that endogenization of the eRTBVL-X family occurred initially in Chinese *O. rufipogon* populations, and then the corresponding locus was inherited by a limited number of *japonica* cultivars.

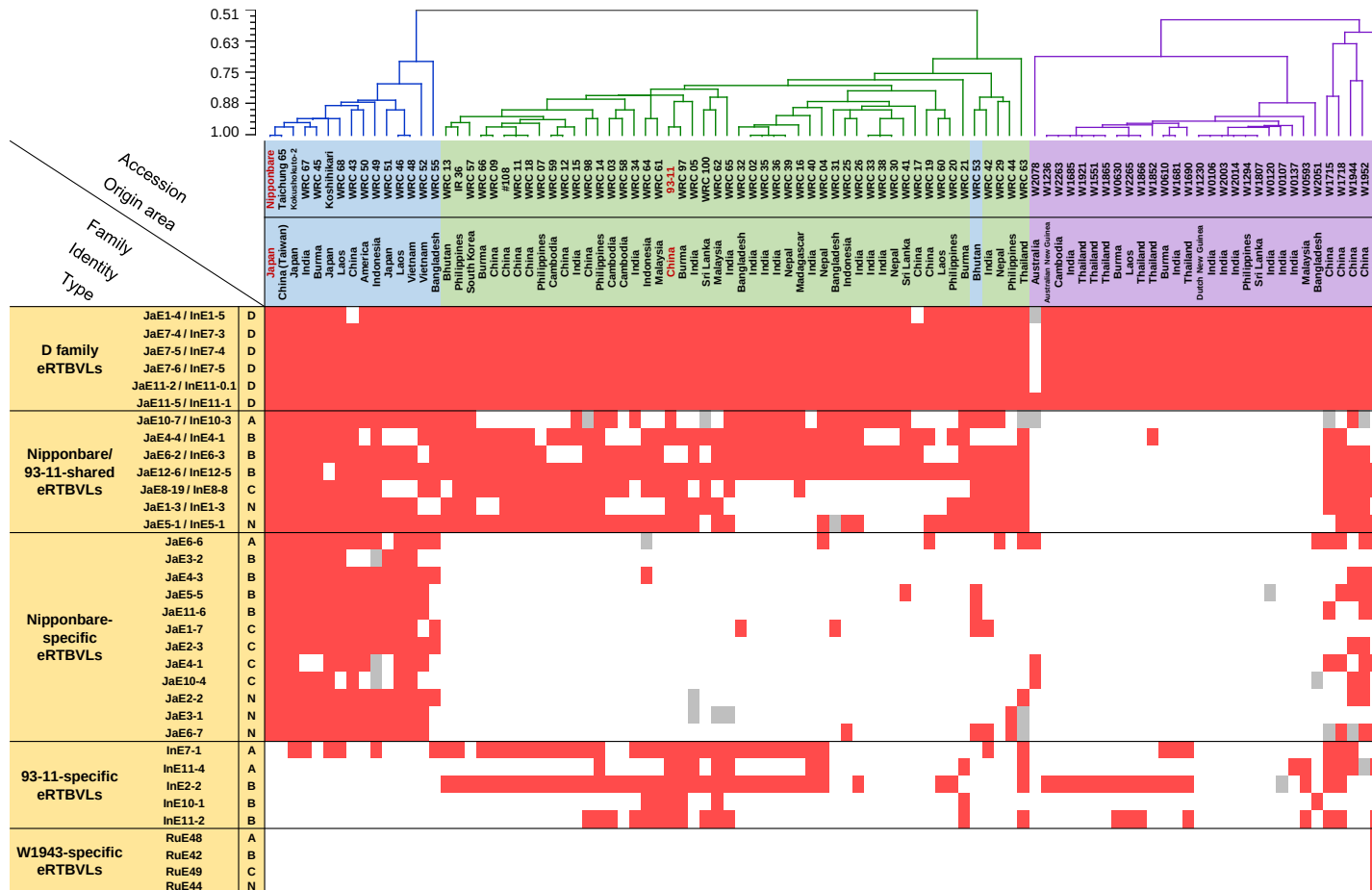


Figure 2.6 Orthologous eRTBVL distribution in a collection of cultivated and wild rice accessions

Genomic PCR was used to estimate the presence/absence of orthologous eRTBVLs in the rice genomes. Red/white rectangles indicate presence/absence of orthologs; gray rectangles indicate an ambiguous result. The taxonomy of the rice accessions constructed using these eRTBVL profiles is displayed at the top; blue indicates the *japonica* group, green indicates the *indica* group, and purple indicates the *O. rufipogon* group. The scale indicates the similarity coefficient among these accessions. The rice accession name and its area of origin are shown below the branches. The colored backgrounds indicate the accessions to which the subspecies or species belong, according to the databases and references (see Supplementary Table S1). The red font marks the three accessions, the genome databases of which were investigated in the present study. The eRTBVL elements that were examined fell into five groups: two groups of species-shared elements (eRTBVL-D family elements and the other Nipponbare/93-11 shared elements), and three groups of species-specific elements (elements specific to Nipponbare, 93-11, and W1943). Details of these element are available in Liu *et al* (2012) and Supplementary Table S3. Elements for which there was no family information (because of short length) are indicated by the letter N.

2.3.6 Temporal relationships of endogenization of eRTBVL families

To unravel the order of endogenization of the eRTBVL-A, -B, and -C families, I analyzed the single-nucleotide-polymorphisms (SNPs) that were contained solely in a single segment (one copy of the eRTBVL sequence) within each set of sequence alignments (called as sole-SNPs), to compare the sequence diversity of the eRTBVL families after their endogenization (see Methods). The sequence diversities based on these sole-SNPs were largely a result of the mutations that had accumulated after the endogenization of the viruses into the rice genomes. In all the datasets derived from the sole-SNP data, I found that the nucleotide sequence diversity within the eRTBVL-B family was always larger than the diversities within the eRTBVL-A and -C families (Table 2.1). Accordingly, the higher number of mutations after the endogenization of eRTBVL-B should be primarily responsible for the larger nucleotide diversity of the eRTBVL-B sequences compared with the diversities of the eRTBVL-A and -C sequences. Thus, the endogenization of eRTBVL-B has preceded that of eRTBVL-A and -C. No consistent trend of differences in nucleotide diversity was detected between the eRTBVL-A and -C families, which indicated the temporal closeness of their endogenization (Table 2.1).

Table 2.1 Sequence divergences of three eRTBVL families after endogenization based on sole-SNP calling

Dataset ^a	eRTBVL family	RT/RH region		ORFz		IGR	
		Divergence ^b	Sample size ^c	Divergence ^b	Sample size ^c	Divergence ^b	Sample size ^c
Nipponbare	A (A1/A2) ^d	0.0024	14	0.0028	14	0.0056/0.0044	9/7
	B	0.0055	10	0.0084	8	0.0102	13
	C	0.0033	11	0.0019	11	0.0049	10
93-11	A (A1/A2) ^d	0.0050	4	0.0053	4	NA/NA	NA/NA
	B	0.0057	13	0.0087	11	0.0089	13
	C	0.0043	6	0.0037	7	0.0031	7
W1943	A (A1/A2) ^d	NA	NA	0.0047	7	0.0113/NA	4/NA
	B	0.0062	10	0.0095	12	0.0122	9
	C	0.0060	6	0.0027	10	0.0056	16
Combined	A (A1/A2) ^d	0.0014	21	0.0011	25	0.0042/0.0039	14/10
	B	0.0030	33	0.0051	31	0.0051	35
	C	0.0023	23	0.0015	28	0.0039	33

The SNPs that appeared solely in one segment (sole-SNPs) in the alignment of sequences from one eRTBVL family were used to estimate the divergences of the eRTBVLs after endogenization for the eRTBVL-A, -B and -C families (see Methods).

^aThe alignment datasets for sole-SNP calling were constructed using the eRTBVL sequences of the RT/RNase H (RT/RH) region, ORFz, and the intergenic region (IGR) (>80% of the length of consensus sequence of each region) from the corresponding single rice genome (Nipponbare, 93-11, or W1943) and from the three genomes combined.

^bNumber of base substitutions per site.

^cNumber of sequences in an alignment for sole-SNP calling.

^dIn the IGR datasets, the divergence analyses were performed separately for the eRTBVL-A1 and -A2 subfamilies.

NA, not available (when the sequence sample size ≤ 3 , the data were not considered).

2.3.7 Recombination of ancient pararetroviruses

The results of the phylogenetic trees based on the ORFx to RT/RH sequences were inconsistent with the results for the trees based on the ORFz and IGR sequences (Figure 2.3); the former revealed a close relationship between eRTBVL-B and -C, while the latter revealed a relationship between eRTBVL-A and -B (Figure 2.3). A reason for this inconsistency is the possible exchange of the virus sequences. The multiple copies in each of the eRTBVL families allowed us to construct precise genomic consensus (ancestral) sequences of the corresponding viruses of the eRTBVL families (a full-length eRTBVL-D consensus sequence is not available). All the viral consensus sequences that I constructed comprised the three proper ORFs and an IGR (Figure 2.7). The full consensus sequences for the eRTBVL-A1, -B, and -C families were aligned and examined using three methods (RDP, GENECONV, and BOOTSCAN) (Martin *et al*, 2010) to detect the most likely signals for homologous recombination events between sequence pairs. The outcomes that were detected by all three methods, represented by the RDP plot, are displayed in Figure 2.8A. High pairwise similarity was identified between eRTBVL-B and -C, but the similarity dropped from the RT/RH to IGR where the pairwise identity between eRTBVL-B and -A rose. This result indicated that the similarity with eRTBVL-B was exchanged between eRTBVL-A and -C across the genome. Examination of the SNP pattern in the raw (not consensus) eRTBVL sequences supported the exchanged similarity (Supplementary Figures S7 and S8). Because earlier endogenization of eRTBVL-B suggested its corresponding virus could emerge earlier than the viruses of eRTBVL-A and -C (Table 2.1), the viruses of eRTBVL-A and -C might have evolved from a homologous recombination event between the virus of eRTBVL-B and another, as yet unidentified, virus lineage.

The corresponding virus of eRTBVL-A2 was apparently derived from the virus of eRTBVL-A1, as mentioned above. The recombination events predicted by the three methods clearly showed that the viral sequence of eRTBVL-A2 could have resulted

from the homologous recombination between eRTBVL-A1 and -C viral IGR sequences (Figure 2.8B, SNP pattern in Supplementary Figure S9). The viral lineages of eRTBVL-A and -C might be close in both temporal and spatial distribution as described above, which would have allowed the two viruses to meet and exchange genetic information. Such a recombination event will produce two daughter virus sequences, each of which will inherit one of the two exchanged parental viral sequences. However, the sister counterpart for eRTBVL-A2 was unlikely to have been trapped by the three rice genomes.

To detect other potential recombinations of rice paleoviruses, I examined the eRTBVL-X locus, which was grouped with eRTBVL-A in the phylogenetic trees based on ORFx and IGR, and with eRTBVL-C in the tree based on RT/RH region (Figure 2.3). Recombination analyses of the eRTBVL consensus sequences using the three methods all predicted nearly the same recombination events, which indicated that viruses of the eRTBVL-X family might have resulted from the homologous recombination between the viruses of the eRTBVL-A (possibly -A2) and -C families (Figure 2.8C). Five crossover sites were identified (Figure 2.8C), where homologous recombination breakpoints most probably occurred. One crossover site was located in the terminal end of the MP region, and the other four sites were within the ORFx, CP region (two sites), and ORFz. Examination of the informative SNPs among the eRTBVL-A, -C, and -X sequences validated the putative multiple recombination breakpoints in these regions (an example in Figure 2.8D, complete data in Supplementary Figures S10-S13). I also detected a few eRTBVL-X-specific SNP loci (Figure 2.8D; Supplementary Figures S10, S12 and S13), which helped us exclude the alternative possibility that the eRTBVL-X family was the result of intra-recombinations of eRTBVL sequences within the rice genomes after integration. Taken together, I propose that the viruses of eRTBVL-X arose in the habitat of Chinese *O. rufipogon* populations after multiple recombinations between viruses of eRTBVL-A and -C.

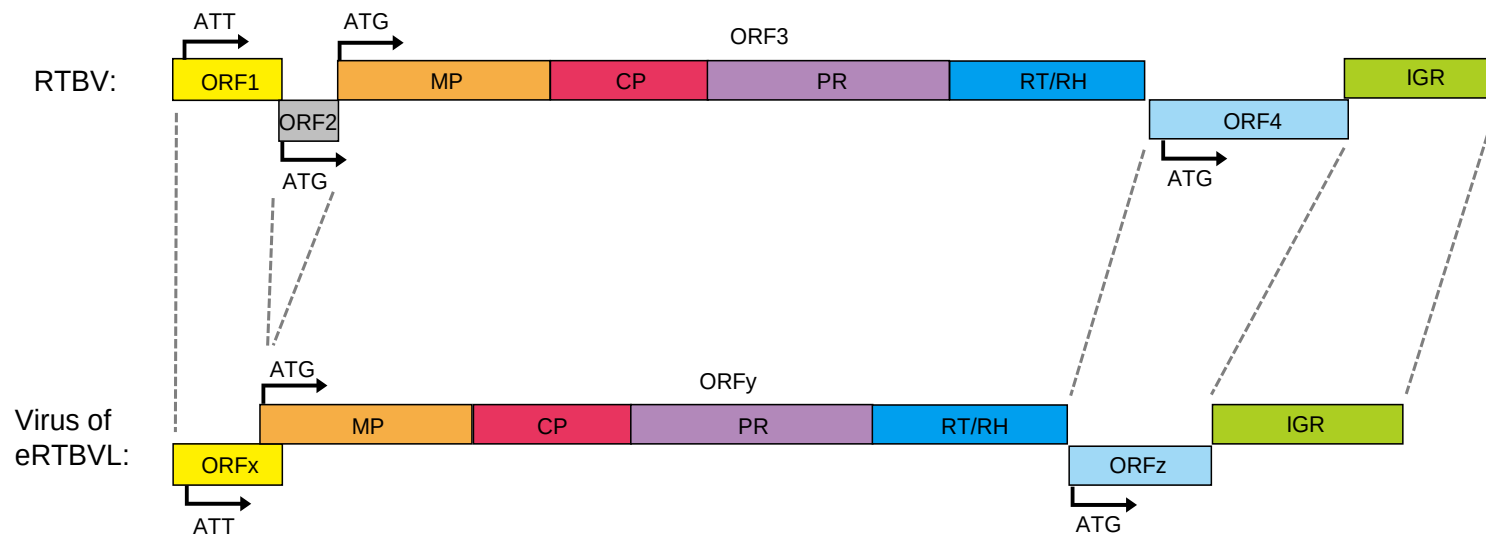


Figure 2.7 Detailed structure of the viral linear genome of constructed eRTBVL consensus sequences

Top: The RTBV genome sequence (NCBI accession: NC_001914.1). Bottom: Potential linear genomic structure of the virus based on the eRTBVL sequences. The start codon of each ORF is indicated.

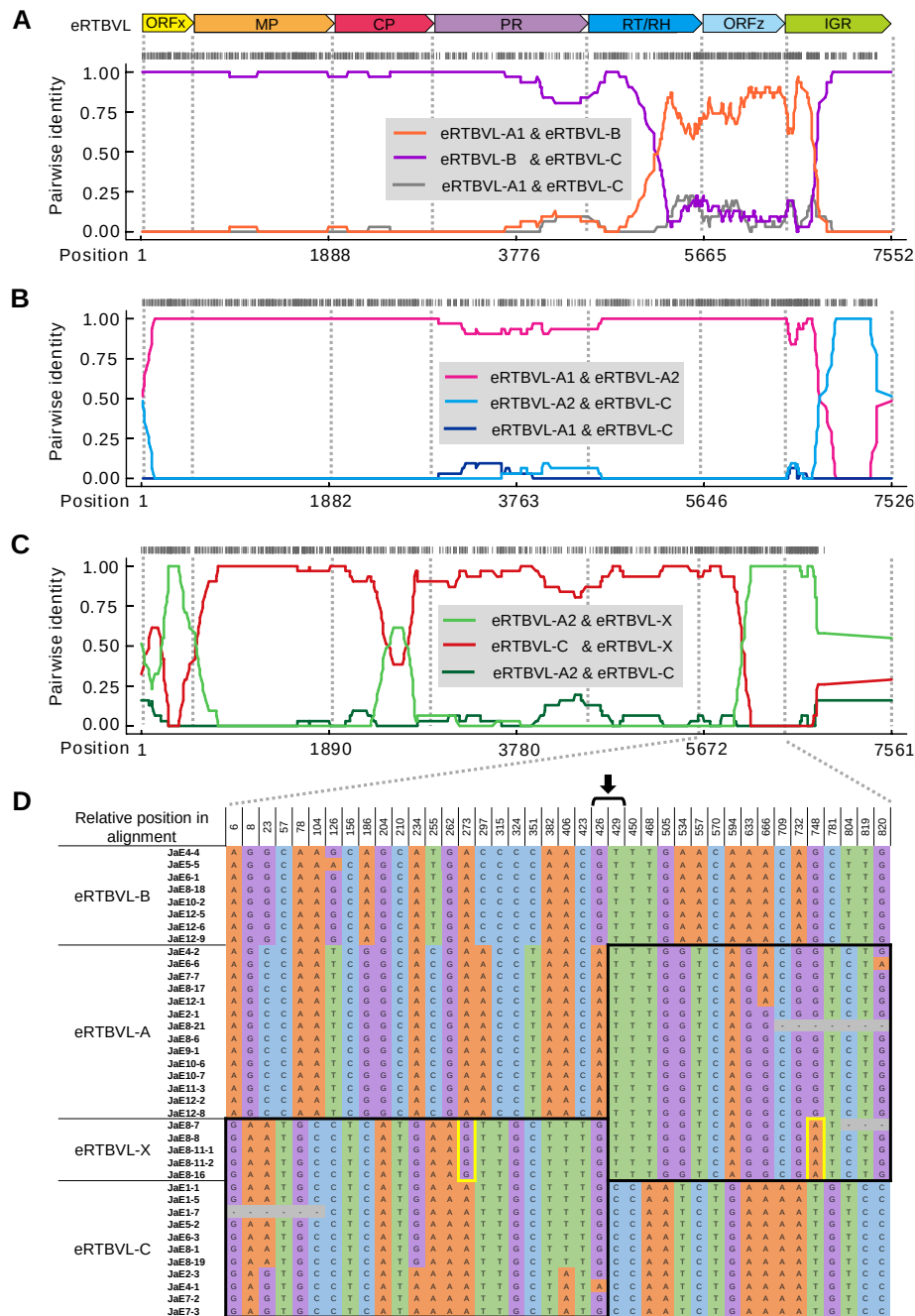


Figure 2.8 Recombination analysis of the eRTBVL families

The structure of the eRTBVL sequences is displayed schematically at the top of the figure. Gray dotted lines indicate the borders of each region in the eRTBVL alignments. The pairwise identity on the vertical axis is the average pairwise sequence identity within a 30-bp sliding window, which was moved one bp at a time across the alignment of the consensus eRTBVL sequences. The position in the alignments is indicated along the horizontal axis. Informative SNP sites in each alignment are indicated by bars at the top of each plot. The solid colored lines are the pairwise identities between two consensus sequences from the eRTBVL families. **(A)** Recombination analysis for the eRTBVL-A1, -B and -C sequences. **(B)** Recombination analysis for the eRTBVL-A1, -A2 and -C sequences. **(C)** Recombination analysis for the eRTBVL-A2, -C and -X sequences (the pairwise identities at the second half of IGR are invalid because there are no informative SNP sites as indicated by the bar). **(D)** Snapshot of an informative SNP pattern in the ORFz of the eRTBVLs in the Nipponbare genome. Black lines outline high pairwise identity sites between the eRTBVL-X and -A/-C sequences, and yellow lines outline the eRTBVL-X-specific SNPs. Hyphens indicate missing data. The arrow indicates the putative breakpoint of the recombination between the eRTBVL-A and -C sequences. Only a section of the informative aligned SNP sites is shown; the complete informative SNP patterns in all the alignments are provided in Supplementary Figures S7-S13.

2.3.8 The most ancient endogenization event of an eRTBVL family

The examination of orthologous eRTBVL-D sequences in the cultivated and wild accessions has confirmed eRTBVL-D as the most ancient family (Figure 2.6). To further explore the endogenization time of the eRTBVL-D family, I screened their orthologous sequences in the other *Oryza* genomes using flanking sequence mapping in the available genomic databases and PCR detection in the corresponding genomic DNAs. The counterparts of 12 loci of the eRTBVL-D family shared by *japonica*, *indica* and *O. rufipogon* were also identified in the genomes of the other *Oryza* AA-genome species: *O. glaberrima*, *O. barthii* (the progenitor of *O. glaberrima*), *O. glumaepatula* and *O. meridionalis* (Figure 2.9, detailed genomic locations see Supplementary Table S5). The orthologous sequence of one of the eRTBVL-D loci was successfully identified in the genome of *Oryza* BB-genome species, *O. punctata* (Figure 2.9, Supplementary Table S5). None of orthologous sequences was found in the genome of *Oryza* FF-genome species, *O. brachyantha* (Figure 2.9).

These observations suggested that the endogenization of the eRTBVL-D family initially occurred before the divergence of *Oryza* AA-genome and BB-genome species, and the time could be at least 5 MYA according to the previous reports of *Oryza* divergence (Zou *et al*, 2008; Zou *et al*, 2013; Zhu *et al*, 2014). Noticeably, most of the eRTBVL-D sequences could be endogenized in a direct common progenitor of all the current *Oryza* AA-genome species and, in fact, the members of the eRTBVL-D family could be derived from multiple independent endogenization events of paleovirus families.

2.3.9 Relationship between the viruses of eRTBVLs and RTBV

The endogenization periods of eRTBVL-A, -B and -C revealed here indicated that the corresponding viruses of these families could not be the progenitors of RTBV because of the low-level of nucleotide similarity between them (44–51%) (Kunii *et al*,

2004). Another, much older eRTBVL-D family was identified for the first time in this study. To explore whether the virus lineage of eRTBVL-D was a common progenitor of eRTBVL and RTBV, I aligned the predicted amino acid sequences translated from the conserved RT/RH regions of the eRTBVL sequences and diverse RTBV strains to construct a phylogenetic tree (Figure 2.10). I found that the eRTBVL families were grouped in one major branch, while the RTBV strains formed another major branch. This result suggested that the corresponding viruses of eRTBVLs could be a sister to RTBV and that their divergence occurred before the speciation of *Oryza* AA-genome species, at least 5 MYA (Zou *et al*, 2008; Zou *et al*, 2013; Zhu *et al*, 2014).

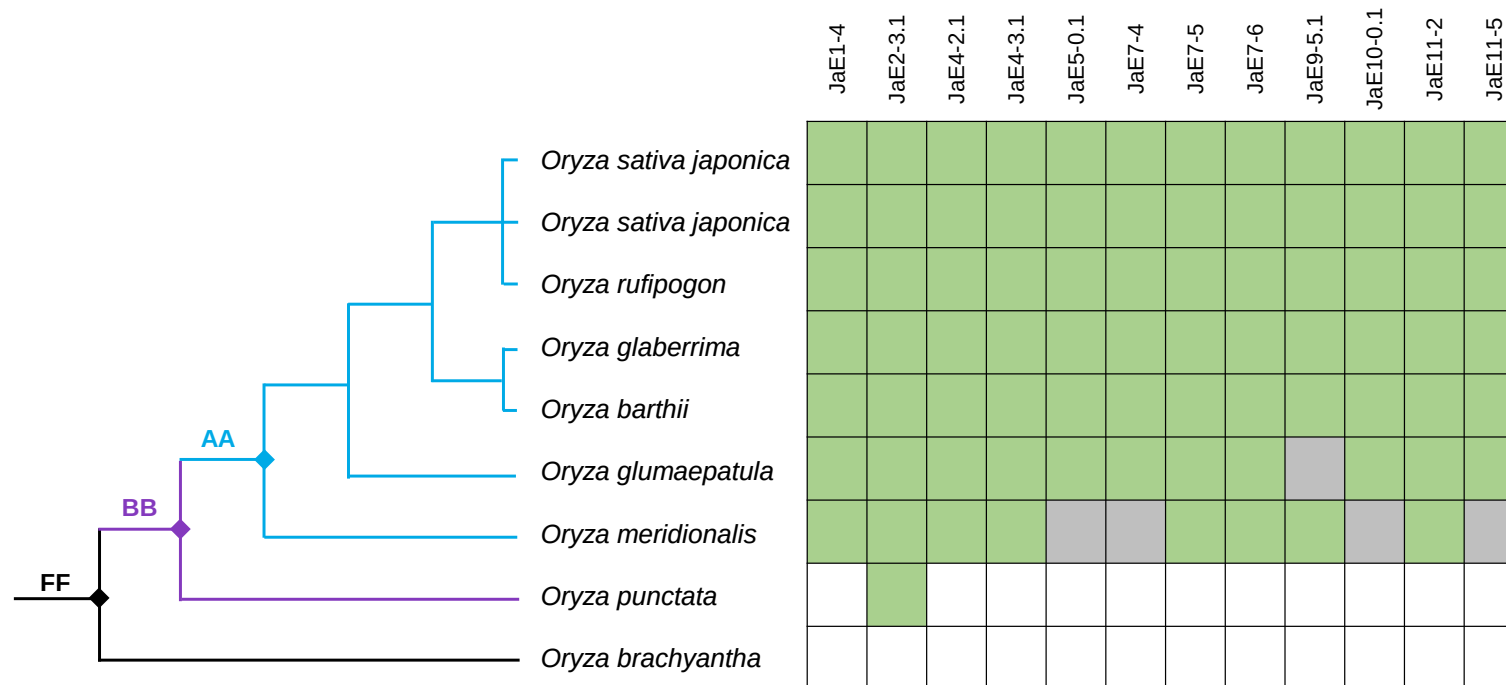


Figure 2.9 Distribution of orthologous eRTBVL-D loci in *Oryza* genomes

The phylogeny of *Oryza* species is displayed at the left according to previous references (Chen *et al*, 2013; Zhu *et al*, 2014). The presence/absence patterns of orthologous eRTBVL-D loci (totally 12 sets) in the corresponding *Oryza* species are shown in the right (green, presence; grey, unknown; white, absence). Detailed genomic locations are provided in Supplementary Table S5.

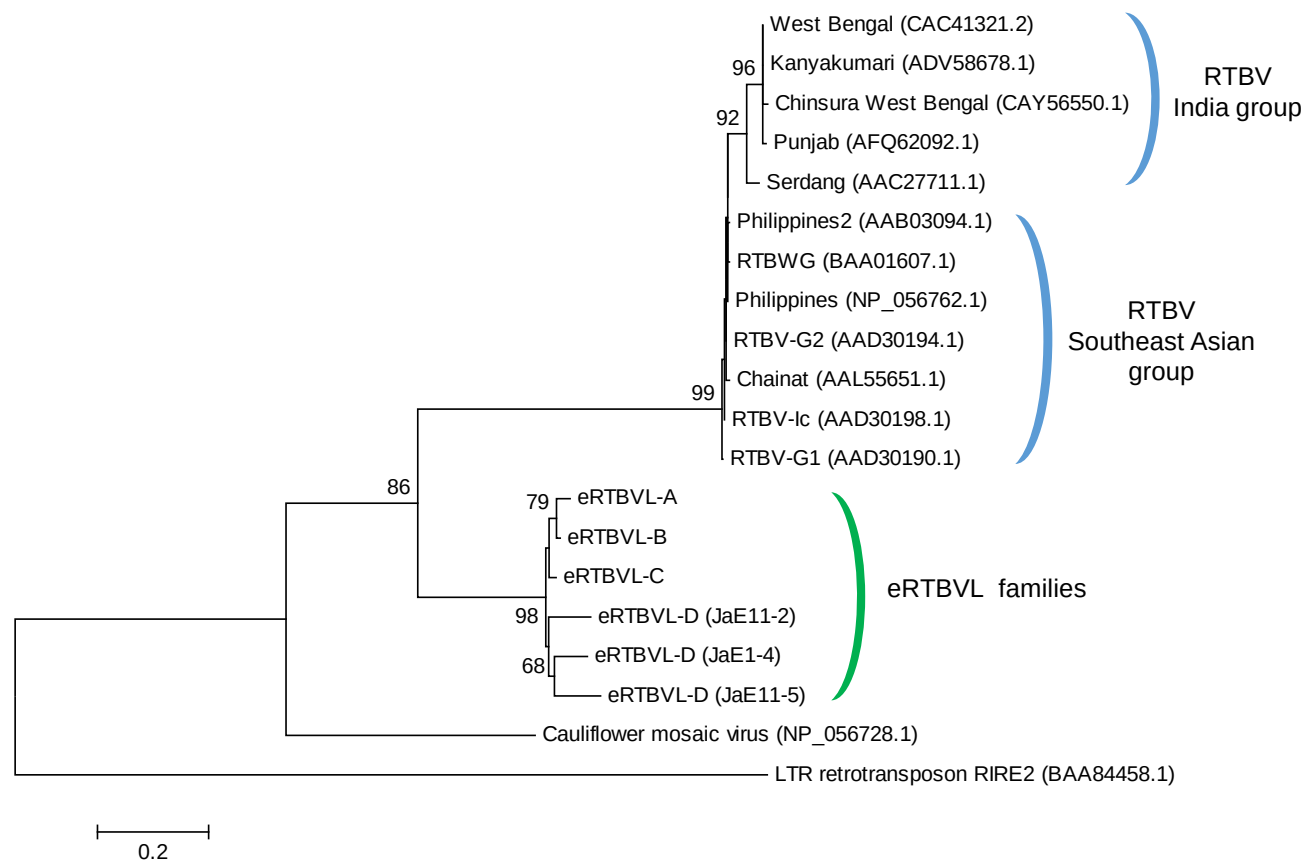


Figure 2.10 Phylogenetic relationship between eRTBVL and RTBV sequences

The ML tree was constructed based on the translated amino acid sequences from the RT/RH regions of the eRTBVL sequences (consensus sequences were used for the eRTBVL-A, -B and -C families) and various RTBV strains. The tree with the highest log likelihood is shown, and bootstrap values greater than 60 are shown above the branches (based on 1000 bootstrap replicates). The tree was drawn to scale, with branch lengths measured in number of substitutions per site. A rice LTR retrotransposon and cauliflower mosaic virus were used as outgroups. The NCBI accession numbers of the RTBV and outgroup sequences are indicated on the tree. The RTBV group names are from Sharma *et al* (2011).

2.4 Discussion

The discovery of integrated copies from pararetroviruses has renewed our realization of the widespread genetic flow from virus to plant genomes (Staginnus and Richert-Pöggeler, 2006; Gayral *et al*, 2010), but the spatio-temporal origin and deep evolution of EPRVs remains largely unknown. These integrated virus sequences are regarded as fossils in the genome. The eRTBVL sequences offer a special paleontological record of the population structure and abundance of the active viruses prior to endogenization and the period of their integration. The data in this study allowed us to infer the genealogy of the viruses of the eRTBVL families and their different geographical origins associated with the evolution of the rice species. In particular, the prediction of recombination events between the different viral lineages helped us to understand the evolutionary relationships between the viruses and the host rice plants.

2.4.1 Genealogy of the viral lineages of eRTBVLs led by recombination

The eRTBVL-D family was identified as the oldest family distinct from the younger eRTBVL-A, -B and -C families (Figures 2.1 and 2.6, summarized in Figure 2.11). Each of the eRTBVL-D sequences had orthologs in nearly all the *O. sativa* and *O. rufipogon* accessions examined here (Figure 2.6). The viral eRTBVL-D sequences, therefore, had begun to immobilize in the common progenitor genome of *Oryza* AA-genome and BB-genome species, at least since 5 MYA (Zou *et al*, 2008; Zou *et al*, 2013; Zhu *et al*, 2014). The other viral lineages of the eRTBVL families emerged relatively recently, after the divergence of the *O. rufipogon* ecotypes (Figures 2.9 and 2.11). The nucleotide divergences after endogenization (Table 2.1) suggested that the emergence of the viral lineage of eRTBVL-B sequences could have preceded those of eRTBVL-A and -C sequences (Figure 2.11). Although an alternative scenario (the virus lineage of eRTBVL-B was younger than the lineages of eRTBVL-A and -C and was a recombinant from the latters) can't be completely excluded, the current scenario of a more ancient virus lineage of eRTBVL-B is reasonable and congruent with my

observations (see below). *Japonica* and *indica* were thought to be domesticated in different processes; *japonica* was established mainly from perennial *O. rufipogon* in southern China, while *indica* was a subsequent hybrid of the ancient *japonica* and the annual *O. rufipogon* in South East and South Asia (Huang *et al*, 2012). Thus, the *japonica* accessions had only a limited genomic contribution from the annual *O. rufipogon* in South East and South Asia. The results of the present study support the domestication process of a genetic flow from *japonica* to *indica* proposed previously (He *et al*, 2011; Yang *et al*, 2012), because my data indicate that Nipponbare and 93-11 shared eRTBVLs that were hardly detectable in the *O. rufipogon* accessions from South East and South Asia (Figure 2.6).

My data suggested that the viral lineages of eRTBVL-A and -C sequences might have arisen from a homologous recombination event (Figure 2.8A). Homologous recombination could have occurred between the viral genome of eRTBVL-B sequences (or closely related virus sequences) and the genome of an unknown virus lineage (Figures 2.8A and 2.11; Table 2.1). Most of the eRTBVL sequences in the three rice genomes fell into the four families, and no other new families were found. If an unknown virus lineage was involved in the homologous recombination, it had not been integrated into these three rice genomes. The eRTBVL-A family was divided into the eRTBVL-A1 and -A2 subfamilies (Supplementary Figure S6). The virus of eRTBVL-A1 subfamily seemed to be the initial genotype of this lineage (Table 2.1), and the virus of eRTBVL-A2 subfamily was generated by homologous recombination with the virus of eRTBVL-C family (Figures 2.8B and 2.11). Afterwards, the viral lineage of eRTBVL-X sequences originated from a further recombination between the viruses of eRTBVL-A2 and -C sequences (Figures 2.8C and 2.11). Unlike the other eRTBVL loci, the eRTBVL-X locus was observed in a limited number of *japonica* and the Chinese *O. rufipogon* accessions. Thus, the eRTBVL-X sequences are a specific indicator for the accessions associated with the later process of the domestication of *japonica*.

However, I failed to detect some of the possible virus sequences involved in the homologous recombination events in the rice databases investigated here; for example, a parental virus sequence for the recombination with the viral genome of eRTBVL-B that resulted in the viral genomes of eRTBVL-A and -C, and a sister virus sequence of eRTBVL-A2 that had resulted from a recombination between the viral genomes of eRTBVL-A1 and -C. The reasons why these apparently missing virus lineages of eRTBVL sequences did not leave a mark on the rice genomes are difficult to understand.

Recombination is a powerful force that can drive rapid genetic alterations in the DNA and RNA viruses (Roossinck, 1997; Hu *et al*, 2003). Many viruses, including dsDNA viruses, frequently exchange their genomes. In cauliflower mosaic virus (CaMV), for example, the co-infection of different CaMV sequences with neutral markers into a single host induced recombinants that made up over 50% of the recovered viral genomes (Froissart *et al*, 2005). The phylogenetic study using 67 global isolates suggested that CaMV had spread from a single population about 400–500 YA, and that recombinations had occurred continuously as CaMV spread (Yasaka *et al*, 2014). If this is so, the invasions of a massive number of virus lineages could leave copies like the eRTBVL-X sequences that had been accidentally integrated into the host genome, although the major lineages for the eRTBVL-A to -C sequences could be explained by only a few recombination events. Why the rice genomes contained only a few virus lineages of eRTBVL remains unresolved. The rice genomes appear to contain integrated episomal DNA segments, including the viral sequences in the eRTBVL families, into the AT-rich regions in a nonselective manner (Liu *et al*, 2012). The frequencies of occurrence of these eRTBVL families in the genomes might correspond to the number of infecting viruses. Thus, the viruses that infected the plants during rice speciation could be dominated by the viral lineages of eRTBVL-A, -B and -C families, and their fitness should be higher than the other virus recombinants.

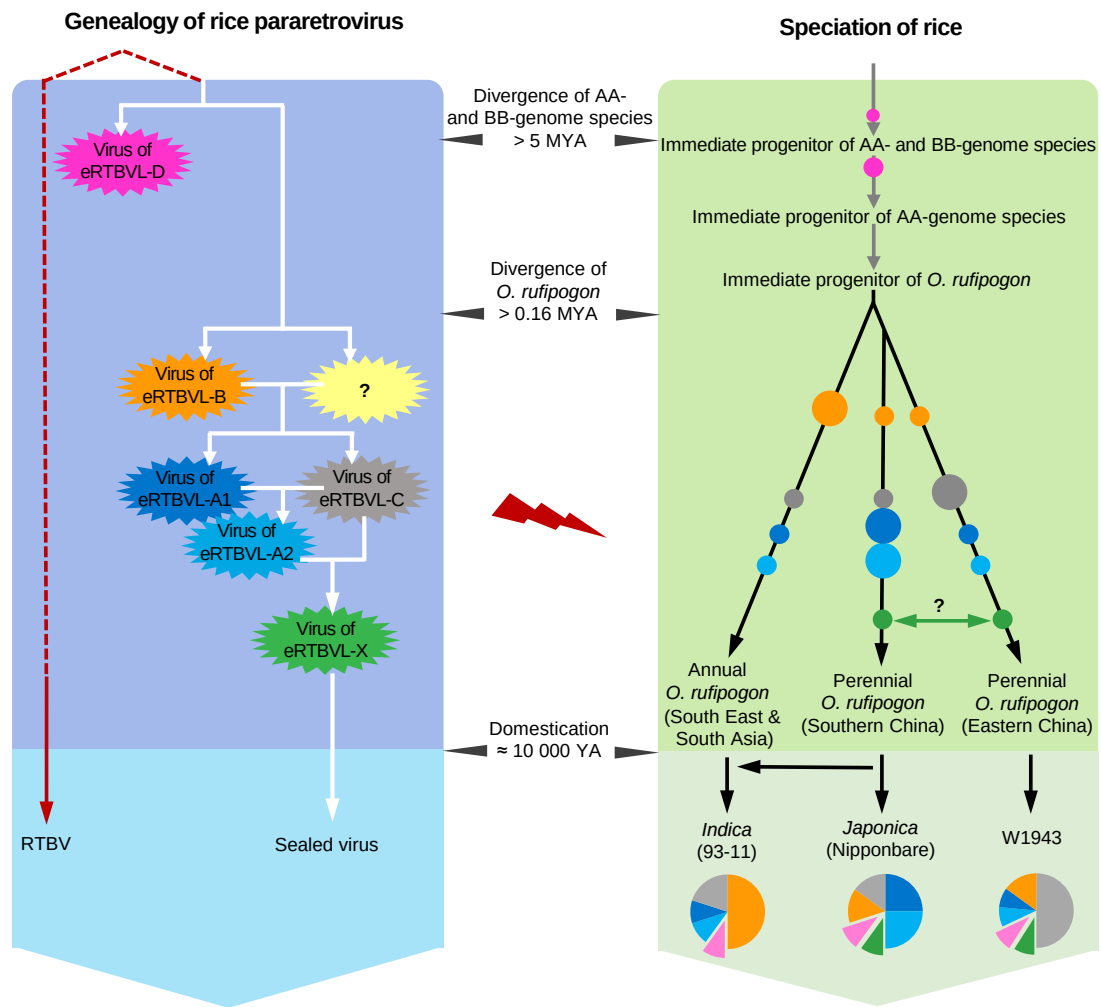


Figure 2.11 Schematic representation of possible integration events of eRTBVLs during rice and pararetrovirus evolution

The genealogy of virus lineages of eRTBVLs is summarized in the left panel using white lines. The colored stars mark the occurrences of the various virus lineages. Among them, the yellow star represents the proposed unknown ancient virus lineage. The red dotted line indicates the unknown evolution of RTBV after its divergence from a common ancestor. The main processes of the rice speciation inferred from previous studies are summarized in the right panel (Lu *et al*, 2008; Zheng and Ge, 2010; Huang *et al*, 2012). The grey lines indicate the time flow from the immediate common progenitor of *Oryza* AA-genome and BB-genome species to the immediate progenitor *O. rufipogon*, and black lines indicate that from the immediate progenitor of *O. rufipogon* to the currently recognized accessions. Colored circles on the lines indicate the possible onset of endogenization; the colors correspond to the colors that indicate the virus lineages in the left panel. The size of the circle indicates the number of virus copies from these viral lineages that were integrated into the three rice genomes, reflecting the different geographic origin and prevalence of the virus lineages in different rice habitats. The abundances of the eRTBVL families shown in the pie charts at the bottom of the panel are not absolute. The timescales between the panels indicate three of the major branches in the rice speciation; MYA, million years ago.

2.4.2 Geographic distributions of viral lineages of eRTBVLs

Considering that the different major eRTBVL families were characteristic of the Nipponbare, 93-11, and W1943 genomes, the geographical origins for the virus lineage that infected the initial host plants could be estimated. The virus corresponding to eRTBVL-B, which was the largest family in the 93-11 genome, might have been prevalent in South East and South Asia, especially India and Indochina known as the ancestral centers of *O. rufipogon* diversity (Londo *et al*, 2006), where the annual *O. rufipogon* was the dominant rice population (Figures 2.4 and 2.11). Similarly, the virus corresponding to eRTBVL-A, which was the major type in the Nipponbare genome, might have been prevalent in southern China where perennial *O. rufipogon* was the dominant population, and the virus corresponding to eRTBVL-C, which was the major type in the W1943 genome, might have been prevalent in eastern China (Figures 2.4 and 2.11). The specific *O. rufipogon* line W1943 was reported to be the northern most distribution of *O. rufipogon* at present time (26°14'N, 116°36'E) (Lu *et al*, 2008). The temporally earlier emergence of the virus lineage of eRTBVL-B hence may be coupled with the expansion and divergence of *O. rufipogon* from South Asia into eastern China. Molecular evidence has indicated that the geographical origin of *japonica* was in the Pearl River region of southern China (Huang *et al*, 2012; Wei *et al*, 2012), while archaeological evidence has indicated that the Lower Yangtze region was one of the regions where rice cultivation originated (Zong *et al*, 2007; Fuller *et al*, 2009). In my study, I found that the profile of eRTBVL sequences in the *O. rufipogon* accession W1943 from the Lower Yangtze region are different from the profile of the sequences in Nipponbare (Figure 2.4, Supplementary Table S3), suggesting that the local *O. rufipogon* accessions from the Lower Yangtze region were unlikely to have been used in domestication as the major founder of *japonica*.

The three genomes from Nipponbare, 93-11, and W1943 contained all of the eRTBVL families (except eRTBVL-X) identified in this study. So far, no rice

accessions that contain only a single eRTBVL family have been found, which makes it difficult to identify a single specific origin for each eRTBVL. Most pararetroviruses can be transmitted by vector insects (Hohn, 2013). The RTBV infects rice plants through the vector *Nephotettix virescens* (green leafhopper); therefore, it is possible that the viral origins of the eRTBVLs were transmitted by insects that are related to the green leafhopper. The geographic distribution of vectors can strongly influence the genetic flow from virus to plant genomes. The different proportions of the eRTBVL families that I detected in the three representative rice genomes could have been influenced by the number of vector insects carrying each virus and/or the virus populations. Thus, in South East and South Asia, the insect population could have been carriers of large numbers of the viruses corresponding to eRTBVL-B, which was the largest family in the 93-11 genome. In southern China, the insect population could have been carriers of large numbers of the viruses corresponding to the major eRTBVL family in the Nipponbare genome, while in eastern China the insect population could have been carriers of the viruses corresponding to the major eRTBVL family in the W1943 genome. This proposal together with the recombination relationships reported here might indicate that the two vector populations in southern and eastern China were often in contact, which could have led to the exchanges among the virus genomes.

Chapter 3 Widespread and independent endogenizations of eRTBVL superfamily in other grass genomes

3.1 Introduction

EPRVs are endogenous copies of pararetroviruses in host genomes resulting from accident integration and provide valuable and novel information for the studies of virus and host evolution (Staginnus and Richert-Pöggeler, 2006; Feschotte and Gilbert, 2012). In the *O. sativa* genome, a family of EPRVs, eRTBVLs, has been identified (Kunii *et al*, 2004) and was thought to be derived from a sister of the current pararetrovirus (RTBV) infecting rice. With the help of RTSV, RTBV also can infect other grass (Poaceae family) species in addition to rice (Khan *et al*, 1991). It remains unknown whether other grass species, especially non-*Oryza* species, were also the hosts of the viruses of eRTBVLs and were also genomically integrated by similar virus-derived sequences. The EPRVs derived from RTBV itself or its progenitor are unidentified. The further screening of eRTBVL- or RTBV-related EPRVs will reveal the in-depth evolutionary history of rice pararetroviruses.

EPRVs in host plant genomes have been deeply studied at the species- and genus-wide levels, but not yet at a family-wide level. For example, the EPRV sequences of banana streak virus have been examined in large numbers of banana varieties and several close relatives (Ndowora *et al*, 1999; Geering *et al*, 2001; Geering *et al*, 2005; Harper *et al*, 2005; Gayral and Iskra-Caruana, 2009; Gayral *et al*, 2010). I and former researchers provided a clear endogenization timeline and evolution map of eRTBVL sequences mainly in *Oryza* AA-genome species (Kunii *et al*, 2004; Liu *et al*, 2012). Orthologous loci of the eRTBVL-D family also presented in the genome of *O. punctata* (*Oryza* BB-genome species) in addition to the genomes of *Oryza* AA-genome species. Other eRTBVL families resulting from independent integration events may exist in the genomes of *Oryza* non-AA-genome species and non-*Oryza* grass species. Here, I

extended the screening of homologous sequences of eRTBVLs to all the other currently available grass genomes and identified eRTBVL homologous sequences in all the examined genomes. These sequences formed an eRTBVL superfamily and were derived from multiple independent integrations of a sister genus of *Tungrovirus*. The distinct events of pararetrovirus integration in different grass genomes may have been shaping grass genomic diversity and evolution.

3.2 Materials and methods

3.2.1 Genomic data mining and paleoviral genome reconstruction

The detailed information of grass assembled genomic sequences was listed in Supplementary Table S2. To identify homologous eRTBVL sequences in these grass genomic sequences, BLASTn (parameters: word size, 11; gap open, 5; gap extend, 2; penalty, -3; reward, 2) and BLASTp (parameters: word size, 3; gap open, 11; gap extend, 1) searches were performed using the BLAST+ 2.2.27 application (Camacho *et al*, 2009) with the eRTBVL consensus sequences (DDBJ accession numbers: BR000029-BR000031) and the corresponding translated amino acid sequences as queries, respectively. The BLAST hit sequences (e-values < 1e-10, alignment length > 100 bp) and their both 5 kb flanking sequences were extracted to construct the possible viral genomes. Typical viral linear genomic sequences for the eRTBVL-similar sequences from each of the grass genomes were assembled in Vector NTI Advance 11.5 toolkit (Invitrogen, USA).

These assembled full-length viral sequences were then used as queries in a BLASTn search to identify the distribution of eRTBVL homologous sequences in grass genomes (e-values < 1e-10 and length >100 bp). Generally, the identified sequences with an interval shorter than 2 kb in a same chromosome or scaffold sequence were recognized as a single eRTBVL segment. The functional domains of the translated protein sequences were predicted by HHpred server (all standard HHM databases chosen,

default parameters) (Söding *et al*, 2005) and aligned to the domains of the eRTBVL sequences identified in the *O. sativa* genome for confirmation. The patterns of sequence conservation between the N-terminals of ORFx1 and ORFx proteins, and between the C-terminals of ORFx2 and ORFx proteins were generated with WebLogo version 3.3 (Crooks *et al*, 2004) using the alignments of translated protein sequences from all the ORFx1, ORFx2 and ORFx sequences with lengths > 80% of the full-lengths in grass genomes.

3.2.2 Phylogenetic analysis

All the sequences with lengths > 80% of the full-length of RT/RH region were retrieved from the grass genomes. The sequences were aligned using ClustalW (Thompson *et al*, 1994) followed by manual editing, and ambiguous regions and missing data were removed (854 positions in the final dataset). ML phylogenetic analysis was performed by MEGA 5.05 (Tamura *et al*, 2011). The best-fitting substitution model was determined in MEGA 5.05 and the model of GTR+G was chosen. All the available sequences of RTBV strains from NCBI nucleotide database were used as outgroup. Supports for the tree branches were evaluated by 1000 bootstrap replicates and the ML tree with the highest log likelihood was midpoint rooted and displayed.

3.2.3 Orthologous insertion analysis

The datasets of genome collinearity (genome-wide alignments) between *Oryza* species were obtained from Gramene database (Monaco *et al*, 2014). Firstly, the left and right 5 kb flanking sequences of each locus of eRTBVLs in the *O. sativa* genome were mapped into the genomes of *O. punctata* and *O. brachyantha* using BLASTn algorithm in BLAST+ 2.2.27 utility. The mapping results were rechecked using the genome collinearity data. Subsequently, the both 5 kb flanking sequences of each locus in the *O. punctata* and *O. brachyantha* genomes were mapped into the *O. sativa*

genome and were rechecked in the same route.

3.3 Results

3.3.1 Sequences of eRTBVL superfamily present throughout grass genomes

To further examine the phylogenetic origins of the viruses of eRTBVLs and RTBV, I performed a systematic screening of eRTBVL homologous sequences across the other ten genomes of grass species which have been genome-widely sequenced and represent all the four grass subfamilies (Supplementary Table S2). Surprisingly, I identified eRTBVL-similar sequences (e value < 1e-10, length > 100 bp) in all the grass genomes I examined. The number of identified segments varies from two to hundreds in different genomes (Table 3.1). The genomes of *Zea mays* and *Sorghum bicolor* from Panicoideae subfamily harbored the least segment number (named eRTBVL-ZM and eRTBVL-SB respectively) (Table 3.1). Based on the phylogenetic analysis of the sequences of the conserved RT/RH domain (Figure 3.1), all these eRTBVL-similar sequences were grouped into a cluster with eRTBVL-A, -B, -C and -D together, and I therefore called this cluster of eRTBVL-related sequences as eRTBVL superfamily; all the sequences of RTBV strains were grouped into another cluster, which suggested that all the viruses of eRTBVL superfamily had a common phylogenetic progenitor and supported that these viruses were from a sister genus of *Tungrovirus*.

In the cluster of eRTBVL superfamily, there are four major subclusters (Figure 3.1). The subcluster I harbored most of the eRTBVL sequences from the genomes of Ehrhartoideae subfamily and all the sequences from the genome of Bambusoideae subfamily, and contains the eRTBVL-A, -B, -C and -D sequences (eRTBVL-OS) previously identified (Figure 3.1). The subcluster II mainly included the sequences from the genomes of Pooideae subfamily, and the subcluster III included most of the sequences from the genomes of Panicoideae subfamily (Figure 3.1). Noticeably, some of the eRTBVL sequences from the genomes of *Panicum virgatum* (eRTBVL-PV) and

O. punctata (eRTBVL-OP) were grouped into the subcluster IV, which suggested that the eRTBVL sequences in each of the two genomes were derived from distinct virus lineages of eRTBVL superfamily (Figure 3.1). In each subcluster of eRTBVL superfamily, the sequences from a single genome generally formed multiple minor but obvious subclades (Figure 3.1). These observations implied a considerable diversity within eRTBVL superfamily and independent endogenization events of some elements of eRTBVL superfamily.

3.3.2 Diverse viral genomic structures of eRTBVL superfamily

By virtue of numerous copies of eRTBVL sequences, I reconstructed the individual possible viral genomes of eRTBVL superfamily in each genome. Interestingly, the eRTBVL sequences within and among grass genomes showed distinct types of viral genomic structures (Figure 3.2). Most eRTBVL sequences in the examined Ehrhartoideae genomes displayed the same structures as eRTBVL-OS (named as type 1), but a portion of eRTBVL-OP sequences showed different structures (type 2) (Figure 3.2). Compared to type I eRTBVLs, the major structural difference of type 2 eRTBVLs is the existence of two ORFs (ORFx1 and ORFx2) in front of the ORFy (Figure 3.2). Sequence comparison analysis revealed that the first and second halves of the ORFx exhibited significant sequence similarity to the first half of the ORFx1 and the second half of the ORFx2, respectively (Figure 3.3). RTBV genomes also possess two ORFs in front of the ORF3, the counterpart of the ORFy, but no obvious nucleotide or protein sequence similarity was found between the ORF2 of RTBV and the ORFx2 of type 2 eRTBVLs. The eRTBVL sequences in the examined Pooideae and Bambusoideae genomes only showed the structures of type 2 (Figure 3.2).

In the Panicoideae genomes, the eRTBVL sequences displayed the structures of another new type (type 3) as well as type 2 (Figure 3.2). Because of limited and short eRTBVL sequences in the *Z. mays* and *S. bicolor* genomes, it is difficult to judge the

structure type of the eRTBVL-ZM and eRTBVL-SB sequences. Surprisingly, for the type 3 structure, I failed to make a single sequence of viral genome, but built two separate sequences, suggesting split viral genomes of type 3 eRTBVLs (Figure 3.2). The two subgenomes of split viral genomes are credible, because both subgenomes started with PBS sequences and ended with IGR sequences. The subgenome A contained the ORFx and partial ORFy (corresponding to the first half of the ORFy of type 1 and 2 eRTBVLs), and the subgenome B carried the ORFz and partial ORFy (corresponding to the second half of the ORFy of type 1 and 2 eRTBVLs); the redundant PR domains appeared in both subgenomes (Figure 3.2). The type 3 eRTBVLs revealed a novel genomic organization of pararetroviruses, and such split pararetroviral genome has not been identified in the nature.

3.3.3 Abundant independent endogenization events of eRTBVL superfamily

In the phylogeny of the viruses of eRTBVL superfamily based on the RT/RH domain, the type 1 eRTBVLs only appeared in the subcluster I and the type 3 eRTBVLs only in the subcluster III (Figure 3.1). However, the type 2 eRTBVLs were present in three of the subclusters except the subcluster III (Figure 3.1), suggesting the highest divergence of the viruses with type 2 genomic structures. The profiles of eRTBVL sequence and structure diversities supported independent endogenization events of most eRTBVL sequences among different grass genomes and within a grass genome. To further confirm this, I examined the eRTBVL sequences in the three Ehrhartoideae genomes by orthologous locus mapping. Except only one locus of eRTBVLs shared by both the *O. sativa* and *O. puctata* genomes (see Chapter 2), the endogenization of eRTBVL-OS loci was independent of that of eRTBVL-OP loci. No eRTBVL locus in the *O. brachyantha* genome has orthologous one in the *O. sativa* and *O. puctata* genomes.

Table 3.1 Summary of sequence distribution of eRTBVL superfamily in 11 grass genomes

Genome	Common name	Subfamily	Genomic data size	eRTBVL segment number	eRTBVL total length (kb)
<i>Oryza sativa (japonica)</i>	Rice (AA-genome)	Ehrhartoideae	363 Mb	94	385.6
<i>Oryza punctata</i>	Wild rice (BB-genome)	Ehrhartoideae	381 Mb	75	188.0
<i>Oryza brachyantha</i>	Wild rice (FF-genome)	Ehrhartoideae	253 Mb	19	14.9
<i>Hordeum vulgare</i>	Barley	Pooideae	4.46 Gb	124	97.9
<i>Triticum aestivum</i>	Bread wheat	Pooideae	4.25 Gb	208	491.1
<i>Brachypodium distachyon</i>	Purple false brome	Pooideae	262 Mb	74	258.7
<i>Phyllostachys heterocycla</i>	Moso bamboo	Bambusoideae	1.94 Gb	352	666.4
<i>Zea mays</i>	Maize	Panicoideae	1.95 Gb	2	3.6
<i>Sorghum bicolor</i>	Sorghum	Panicoideae	701 Mb	5	3.5
<i>Setaria italica</i>	Foxtail millet	Panicoideae	391 Mb	80	308.0
<i>Panicum virgatum</i>	Switchgrass	Panicoideae	1.60 Gb	376	731.0

Detailed information of eRTBVL segments is provided in Supplementary Table S6, and genomic database sources in Supplementary Table S2. In the genome of *Oryza sativa (japonica)*, 88 eRTBVL segments were identified by Liu *et al* (2012) and additional six segment (eRTBVL-D) were identified in Chapter 2 of this study.

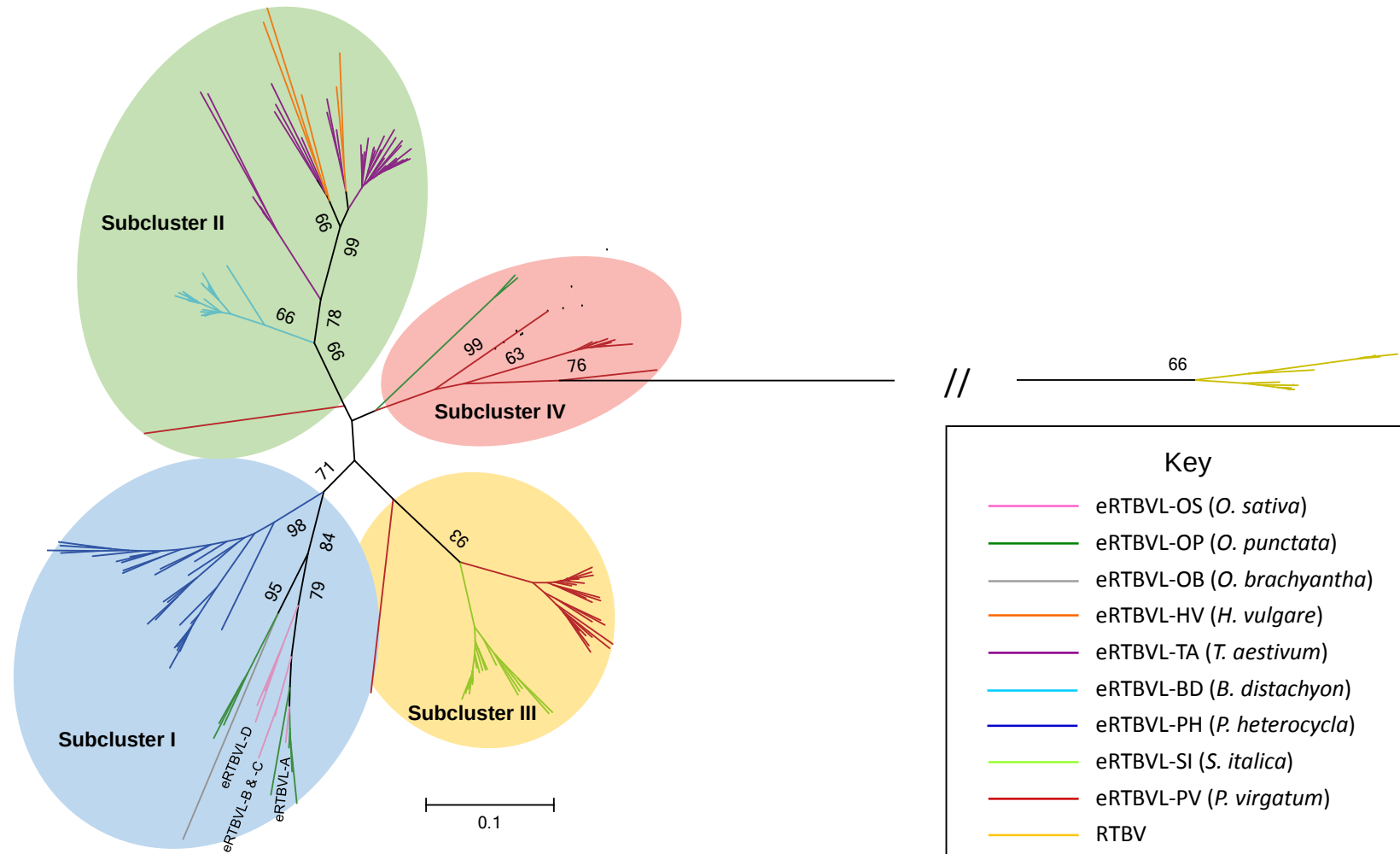


Figure 3.1 Phylogenetic analysis of eRTBVL sequences in grass genomes and exogenous RTBV sequences based on RT/RH domain

Sequences from different grass genomes are indicated by different branch colors. The four subclusters are shown by different background colors. The consensus sequences of eRTBVL-OS were used and their branches are marked. Bootstrap supports greater than 60% (from 1000 replicates) for major nodes are showed in percentage at the branches. The tree is midpoint rooted and drawn to scale, with branch lengths measured by the number of substitutions per site. The accession numbers of RTBV sequences can be found in Figure 2.10.

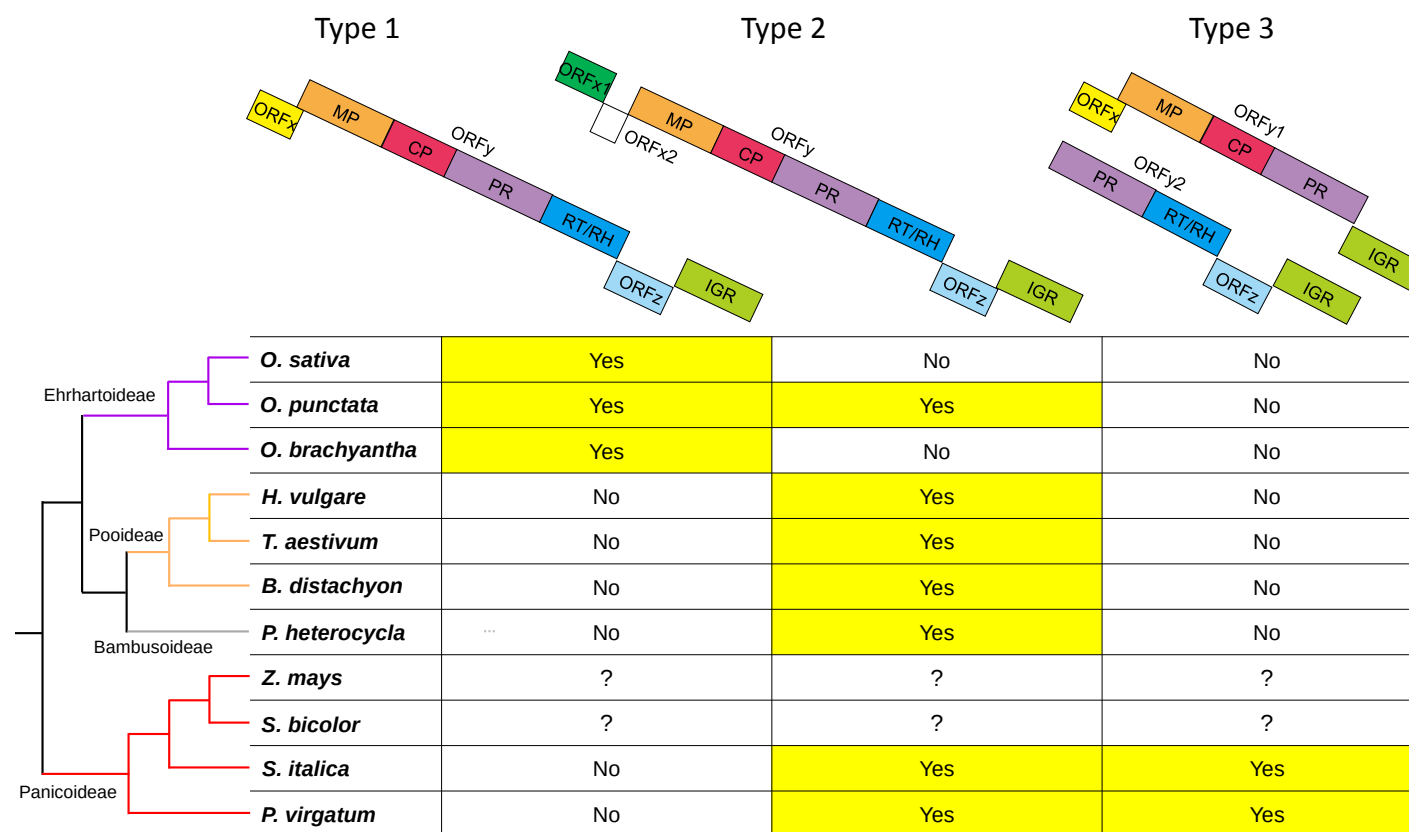


Figure 3.2 Three types of virus genomic structures of eRTBVL sequences in grass genomes

The virus genomic structures of eRTBVL sequences are shown in the top. IGR and each ORF are indicated by boxes, and the presumed MP, CP, PR, and RT/RH domains are also displayed. The phylogeny of grass species is shown at the left, which was built based on previous reports (Catalán *et al*, 1997; Swigoňová *et al*, 2004; Zou *et al*, 2008; Bennetzen *et al*, 2012; Zhang *et al*, 2012; Peng *et al*, 2013). Purple, orange, grey and red branches indicate the species of Ehrhartoideae, Pooideae, Bambusoideae and Panicoideae subfamily, respectively. The distributions of three eRTBVL types with different structures in these grass genomes are summarized in the table. The structure types of the eRTBVL sequences in the *Z. mays* and *S. bicolor* genomes are difficult to judge because of limited and short eRTBVL sequences.

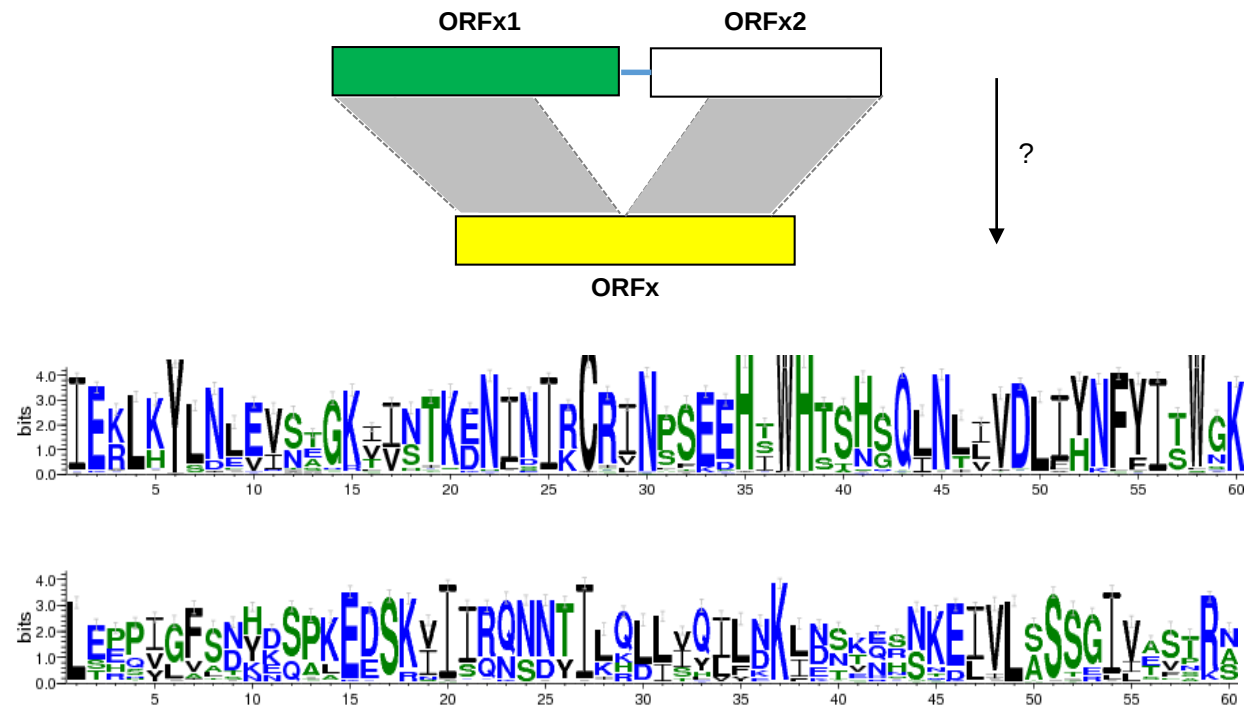


Figure 3.3 Sequence comparison between the ORFx1-ORFx2 and the ORFx

The regions of sequence homology between ORFx1-ORFx2 and ORFx are indicated by grey parallelograms in the top. The pattern of sequence conservation between the N-terminals of ORFx1 and ORFx proteins is displayed with a sequence logo in the middle, and that between the C-terminals of ORFx2 and ORFx proteins in the bottom (60 amino acid positions are shown in each alignment, all the available sequences were used for analysis). In the sequence logos, the overall height of each stack indicates the sequence conservation at that position (measured in bits), and the height of symbols within the stack represents the relative frequency of the corresponding amino acid at that position.

3.4 Discussion

Currently the plant pararetrovirus family comprises seven genera: *Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Solendovirus*, *Soymovirus*, *Badnavirus* and *Tungrovirus* (virus taxonomy from ICTV, 2013). The virus of eRTBVL-OS was a sister of RTBV, the only known virus species of *Tungrovirus*. Geering *et al* (2010) suggested a new genus for the related virus of eRTBVL-OS. Here, the eRTBVL superfamily I identified uncovered the existence of this previously unknown genus of pararetroviruses. This viral genus should be a sister of *Tungrovirus* and was able to infect widespread species in grass family and possibly other relative species. The viruses corresponding to eRTBVL superfamily may be extinct now, or may still be circulating in certain grass populations or species, or may have switched to new plant hosts, non-grass species. If these viruses still hide in the nature, a possibility is that no obvious symptoms of infection make them be overlooked and uncharacterized. It has been suggested that EPRVs could play a role in viral resistance by homology-dependent gene silencing (Mette *et al*, 2002; Staginnus and Richert-Pöggeler, 2006). If eRTBVL sequences were employed by grass species to counter the corresponding exogenous viruses, this virus-derived resistance may lead no or weak symptoms of exogenous viral infections.

Three different genomic structures of the viruses of eRTBVL superfamily were observed here, and other types of viral structures might exist. The type 2 eRTBVLs had the highest sequence diversity, suggesting that the viruses of type 2 could be from the original species of this genus. If so, the viruses of type 1 could be generated from an event of deletion in the ORFx1 and ORFx2 of the viruses of type 2, and this deletion might lead the gene fusion of the ORFx1 and ORFx2 into the ORFx; strikingly, the conservation of protein sequences was maintained after this presumed gene fusion (Figure 3.3). The viruses of type3 could be subsequently generated from a genome split event of the viruses of type 1, and the split event might happen in the PR domain. The alternative evolutionary scenario of the viruses of eRTBVL superfamily cannot be

excluded currently, and further evidence is needed to unveil their origins and divergences. The phylogenetic analysis (Figure 3.1) suggested that host switch frequently occurred, such as the virus lineages of type 2 infecting *O. punctata* which could previously live on the other grass hosts, possibly on *P. virgatum*.

Liu *et al* (2012) suggested that the integrations of episomal DNAs including eRTBVL-OS into AT-rich regions in the rice genomes have generated new or modified transcripts. The sequences of eRTBVL superfamily in the examined grass genomes were generally flanked by AT-rich host genomic sequences, and hence abundant independent endogenization events of eRTBVL superfamily among grass species could have shaped the evolution and divergence of grass species recently. The IGR sequences of pararetroviruses contain promoters that can recruit host transcription factors to transcribe their genomes (Hohn and Rothenie, 2013), and pararetroviral promoter sequences have long been used to drive the expression of exogenous genes (Odell *et al*, 1985). The LTRs of endogenous retroviruses and retrotransposons have been reported to evolve into promoters of host genes (Feschotte and Gilbert, 2012; Rebollo *et al*, 2012), but the evidence that the IGRs of EPRVs have been co-opted for host promoters remains absent. I supposed that the promoter sequences of certain eRTBVL segments can recruit host transcription factors to transcribe the flanking sequences to produce chimeric transcripts. The presumed chimeric transcripts could be long noncoding RNAs possibly with small ORFs (eRTBVL insertions generally present within host intergenic sequences), which may be contributing to the *de novo* origin of proteins (Ruiz-Orera *et al*, 2014) and to the divergence of grass species. Further study is needed to test this hypothesis in the future.

Chapter 4 The eGPV1 and eGPV2 sequences in grass genomes reveal a coevolved mutualism between nonautonomous viruses

4.1 Introduction

As one of the major threats to plant production and human health, pararetroviruses have a distant relationship with LTR retrotransposons (Xiong and Eickbush, 1990). Both of them encode an RT/RH protein to mediate reverse transcription for replication (Temin, 1985). Interestingly, some LTR retrotransposons are nonautonomous for their life circle in host cells, i.e., lost most or all of the coding capacity but are able to amplify themselves by hijacking the protein machinery of autonomous retrotransposons which are functionally and structurally intact (Feschotte *et al*, 2002; Jiang *et al*, 2002; Wicker *et al*, 2007). A hallmark of the parasitism of nonautonomous LTR retrotransposons on their autonomous partners is a substantial sequence similarity of their LTR and neighbor regions, where *cis*-regulation elements are located (Boeke and Corces, 1989; Jiang *et al*, 2002; Du *et al*, 2010).

Plant pararetroviruses (*Caulimoviridae*) were thought to diverge from a same remote origin with the *gypsy* group of LTR retrotransposons (Xiong and Eickbush, 1990), and have a circular dsDNA genome with discontinuities. Although a wide variety of pararetroviral genomic structures, all the known plant pararetroviruses encode four essential proteins (MP, CP, PR and RT/RH) and thus are autonomous during the viral life cycle in host cells. In its replication process, the pararetroviral DNA is firstly transcribed into a pregenomic RNA and the latter is then reverse transcribed by viral RT/RH (cleaved from a polyprotein by PR) into a complementary DNA; this reverse transcription begins at the PBS using an initiator methionine tRNA as primer; subsequently, the synthesis of plus DNA strand is initiated at the PPT (Hull, 2002a). After replication, the viral DNAs are assembled into particles by CP and transmit cell to cell with the help of MP (Hohn and Rothnie, 2013). The IGR of plant

pararetroviruses is a diverse and noncoding region containing *cis*-regulation elements, and is important for viral transcription, translation and replication (Hohn and Roehle, 2013). No naturally nonautonomous species of pararetroviruses was reported until now, probably due to the very limited identification and characterization of pararetroviruses in nature.

EPRVs are viral genomic fossils in host genomes and recorded certain evolutionary processes of pararetrovirus populations (Feschotte and Gilbert, 2012). The mining and studying of EPRVs and other endogenous viral elements in sequenced plant and animal genomes therefore are offering unprecedented fresh insights into viral genetic and genomic evolution (Patel *et al*, 2011; Feschotte and Gilbert, 2012; Katzourakis, 2013). Here, I reported comprehensive paleogenomic data of nonautonomous pararetroviruses by screening EPRVs in depth across 14 genomes of grass species. I provided evidence that two naturally defective but functionally complementary paleovirus species developed into a mutualism and kept co-evolution by noncoding region swapping.

4.2 Materials and Methods

4.2.1 Genomic data mining and paleoviral genome reconstruction

The database version and source of 14 grass genomes were provided in Table S2. For identification of eGPV1-OS, the protein sequences of 56 reference species of plant pararetroviruses from the NCBI genome database were used as a query in a tBLASTn search (word size, 3; gap existence, 11; gap extension, 1) by BLAST+ 2.2.27 utility (Camacho *et al*, 2009) against the *O. sativa* genome sequence. The subject segments spanning high-identity matches (e-values < 1e-50 or sequence identity > 40%, non-eRTBVL sequences) and their both 5-kb upstream and downstream sequences were retrieved to assemble a preliminary consensus sequence of the viral genome in Vector NTI Advance 11.5 toolkit (Invitrogen, USA). This preliminary sequence was then used as a query in a BLASTn search (word size, 11; gap existence, 5; gap extension 2;

penalty, -3; reward, 2) to collect the sequences from all identified loci (e-values < 1e-10 and length >100 bp) to make the final consensus (ancestral) sequence. For identification of eGPV2-OS, I retrieved the segments displaying a region-specific sequence identity higher than 70% to the eGPV1-OS consensus sequence in the BLASTn search and their both 5-kb upstream and downstream sequences to assemble the viral consensus sequence in the same way. To screen homologous sequences of eGPV1s and eGPV2s in the other grass genomes, the BLASTn and BLASTp searches were performed using the above consensus sequences and the translated amino acid sequence of encoding regions as queries respectively. The viral consensus sequences for eGPV1-SB and eGPV2-SB were constructed using the sequences from the *S. bicolor* genome in the same means described above.

4.2.2 Structure detection and domain prediction

Putative ORFs of eGPV1s and eGPV2s were detected using the consensus sequences via ORF Finder utility at the NCBI website (www.ncbi.nlm.nih.gov/projects/gorf). The longest ORFs are reserved and further validated by the multiple sequence alignments of the related sequences of eGPV1s and eGPV2s from grass genomes. Each of the translated protein sequences were subjected to HHpred server (Söding *et al*, 2005) and all the standard HHM databases were chosen for homology domain detection (default parameter). The hits matching to longer than 10 subject residues were retrieved.

4.2.3 Phylogenetic and recombination analyses

The ML phylogenetic analyses were performed using PhyML version 3.0 (Guindon *et al*, 2010). For MP, CP, PR, and RT/RH regions, nucleotide sequences of respective domains of eGPV1s and/or eGPV2s were retrieved from grass genomes (those longer than 90% of the whole-length of each region were chosen), and were aligned by ClustalW (Thompson *et al*, 1994). The nucleotide sequence alignments were then translated into amino acid sequence alignments after removing gaps in BioEdit version

7.2.5 (Hall, 2011). After adding the protein sequences of these domains from other diverse source (eRTBVL, exogenous pararetroviruses, virgaviruses and retrotransposons) into the corresponding protein alignments, the sequence datasets were realigned using MUSCLE (Edgar, 2004) followed by manual editing. The best fitting substitution models were determined by the Akaike Information Criterion (AIC) calculated using ProtTest version 3.2 (Darriba *et al*, 2011). The LG+I+G, WAG+I+G+F, WAG+G+F, and LG+I+G+F models were selected as best fitting models for MP, CP, PR and RT/RH alignments respectively. For the ML phylogenetic analysis of IGR sequences, the IGR nucleotide sequences corresponding to the 3'-terminal of the linear viral genomes of eGPV1s and eGPV2s (those longer than 90% of the whole-length of IGR were chosen) are extracted and aligned by ClustalW plus manual rechecking. Because of a very weak similarity between the IGR sequences of eGPV1^{Or}/eGPV2^{Or} and eGPV1^{So}/eGPV2^{So}, the sequence alignments and ML trees were built respectively for eGPV1^{Or}/eGPV2^{Or} and eGPV1^{So}/eGPV2^{So} sequences. The jModelTest version 2.1.4 was used to judge the best fitting substitution model based on AIC (Darriba *et al*, 2012) for them. The GTR+G and HKY+G models were chosen as the best model used in PhyML for eGPV1^{Or}/eGPV2^{Or} IGR and eGPV1^{So}/eGPV2^{So} IGR alignments, respectively. In all the ML trees, the branch supports were calculated by 1,000 bootstrap replicates. ML trees were visualized with FigTree version 1.4.1 (<http://tree.bio.ed.ac.uk/software/figtree>). The recombination analysis of eGPV1 and eGPV2 raw sequences was performed in RDP version 3.44 (Martin *et al*, 2010) with the default settings.

4.2.4 Orthologous insertion analysis

The datasets of genome collinearity (genome-wide alignments) between *Oryza* species were obtained from Gramene database (Monaco *et al*, 2014). A portion of eGPV1 and eGPV2 segments in an *Oryza* genome were next or near to another eGPV1/eGPV2 segments due to genomic rearrangement, therefore the two (or more)

segments, between which the interval is less than 5 kb in the same chromosome or scaffold, were generally taken as one locus for the mapping of insertion flanking sequences. Firstly, the left and right 5-kb flanking sequences of each locus of eGPV1s and eGPV2s in the *O. sativa* genome were mapped into the other four genomes using BLASTn algorithm in BLAST+ 2.2.27 utility (Camacho *et al*, 2009). The mapping results were rechecked using the genome collinearity data. Subsequently, the both 5-kb flanking sequences of each locus in the *O. glaberrima*, *O. glumaepatula*, *O. meridionalis* and *O. punctata* genomes were mapped into the *O. sativa* genome and were rechecked in the same route. Because some insertion flanking sequences from *O. glumaepatula* and *O. meridionalis* were full of uncharacterized bases (letter N in genomic sequences), the length of these flanking sequences was extended up to 15 kb for examination.

4.2.5 Plant materials and polymerase chain reaction

Seeds of *O. sativa* (cv. Nipponbare), *O. glaberrima* (Acc. IRGC103777), *O. glumaepatula* (Acc. W1169), *O. meridionalis* (Acc. W2112), *O. punctata* (Acc. W1514) and *O. brachyantha* (Acc. W1402) were planted in the greenhouse at Hokkaido University (detailed materials information in Supplementary Table S1). Total DNAs were extracted from the leaf samples using cetyltrimethyl ammonium bromide extraction buffer and the concentrations of DNAs were assessed in a NanoDrop 2000 spectrophotometer (Thermo, USA) and were diluted into the same magnitude. PCRs were performed using Ex Taq or LA Taq polymerase (TaKaRa, Japan) in a PTC-200 thermal cycling system (GMI, USA) as the following program: denaturation at 94°C for 3 min, followed by 34 cycles of 94°C for 30 s, 52°C-58°C for 30 s, and 72°C for 1-3 min, final elongation at 72°C for 10 min. The PCR products were resolved on a 1-2% agarose gel and stained by ethidium bromide, and the results were viewed with an AE-6933FXES Printgraph (ATTO, Japan). Sequencing was performed on a 3730 DNA Analyzer (Applied Biosystems, USA) using a BigDye Terminator v3.1 cycle

sequencing kit (Applied Biosystems, USA) according to the protocol. The assembly of generated sequences were conducted by Vector NTI Advance 11.5 toolkit (Invitrogen, USA).

4.3 Results

4.3.1 Unusual *Petuvirus*-related EPRV, eGPV1 in the rice genome

The Plant Breeding Laboratory, Hokkaido University, previously reported an EPRV, eRTBVLs, in the rice (*O. sativa*) genome which was derived from a sister of extant rice pararetrovirus (RTBV) (Kunii *et al*, 2004 and this study). When I further examined the potential EPRVs in the *O. sativa* genome by a tBLASTn search using all the available reference protein sequences of plant pararetroviruses as queries, I unexpectedly found that the loci best hitting to the sequence of petunia vein clearing virus (PVCV) (e-value < 1e-80) were unrelated to eRTBVLs. By aligning the upstream and downstream sequences of these loci, I assembled a circular consensus sequence with 6558 bp in length (Figure 4.1A). Using this consensus sequence in a BLASTn search, I investigated the distribution of this element in the *O. sativa* genome and found 11 pronounced segments in six chromosomes (Figure 4.1A, Supplementary Table S7). Two long ORFs and an IGR were recognized in this element, and the two ORFs (4269 bp and 1131 bp in length respectively) shared an overlap with 11 bp (Figure 4.1A). To identify functional domains in the ORF proteins, the translated amino acid sequences were subjected to a sensitive detection based on profile hidden Markov models (HHpred; Söding *et al*, 2005). This recognized a potential MP domain in the N-terminal, a PR domain, and an RT/RH domain in the C-terminal of ORF1 polyprotein (Figure 4.1A, Table 4.1). No significant known domain was detected in the ORF2 protein. The conserved motifs in these domains were validated by multiple sequence alignments (Supplementary Figure S14). I then conducted phylogenetic analyses of these segments and diverse pararetrovirus-related sequences for each domain, and found that these segments formed a cluster closer to pararetrovirus and eRTBVL sequences than

outgroups (Figures 4.2A and 4.3), especially closer to PVCV in the tree of the conserved RT/RH domain (Figure 4.2A). These structural and phylogenetic analyses together demonstrated that this element is a novel *Petuvirus*-related EPRV in the *O. sativa* genome, and I termed it as endogenous grass pararetrovirus 1 (eGPV1) *O. sativa* lineage (eGPV1-OS). Strikingly, the essential CP domain for viral particle assemble was undetectable in eGPV1-OS (Figure 4.1A).

4.3.2 Another unusual EPRV, eGPV2 naturally lacking an RT/RH domain but possessing an IGR highly similar to eGPV1

During the investigation of the eGPV1-OS distribution in the *O. sativa* genome, I observed that some fragments showed obvious similarity (70.1%-98.9%) to exactly specific sections of eGPV1-OS, especially the IGR and MP sections, suggesting another element locally similar to eGPV1-OS in the *O. sativa* genome. I reconstructed the structures and made a circular consensus sequence with 6650 bp in length for this element, and found it comprising a long ORF (5523 bp) and an IGR (Figure 4.1B). I identified 23 segments of this element dispersed in nine chromosomes of *O. sativa* (Figure 4.1B, Supplementary Table S7). HHpred detection predicted a MP domain, a CCHC type zinc finger (CP domain) and a PR domain in the ORF ployprotein (Figure 4.1B, Table 4.1), conserved motif of which were subsequently confirmed in the multiple sequence alignments (Supplementary Figure S14). Phylogenetic analyses of the MP, CP and PR domains suggested them closer to pararetrovirus sequences (Figures 4.2B and 4.3). In conclusion, this element is another family of EPRVs, and hereafter was named endogenous grass pararetrovirus 2 (eGPV2) *O. sativa* lineage (eGPV2-OS).

Not sole, but dozens of copies of eGPV1-OS and eGPV2-OS embedded in the *O. sativa* genome indicated that both the corresponding viruses of eGPV1 and eGPV2 should be able to replicate. Surprisingly, an RT/RH domain for replication was failed to be detected in eGPV2-OS (but present in eGPV1-OS) and even no degenerate

RT/RH-like residues were distinguished by the amino acid alignment (Figure 4.1A and B). In addition, an essential CP domain harboring a zinc finger motif was absent in eGPV1-OS as mentioned above (but present in eGPV2-OS) (Figure 4.1A and B). The absences of RT/RH and CP domains do not result from deletions in the EPRVs after the viral insertions into plant genomes, but are characters of exogenous paleoviruses because of consensus patterns at the corresponding regions of eGPV1-OS and eGPV2-OS (Figure 4.1A and B). The comparison between the eGPV1-OS and eGPV2-OS consensus sequences supported the absences of respective domains, and revealed an extraordinarily high local similarity (97.2% of nucleotide identity) mainly at IGR itself and downstream region, including the MP domain (Figure 4.4A). Both eGPV1-OS and eGPV2-OS encoded a PR domain, and the nucleotide similarity between the two PR domains was extremely low (Figure 4.4A). Consequently, my genomic screening revealed two distinct families of EPRVs which were derived from unusual naturally defective pararetrovirus species.

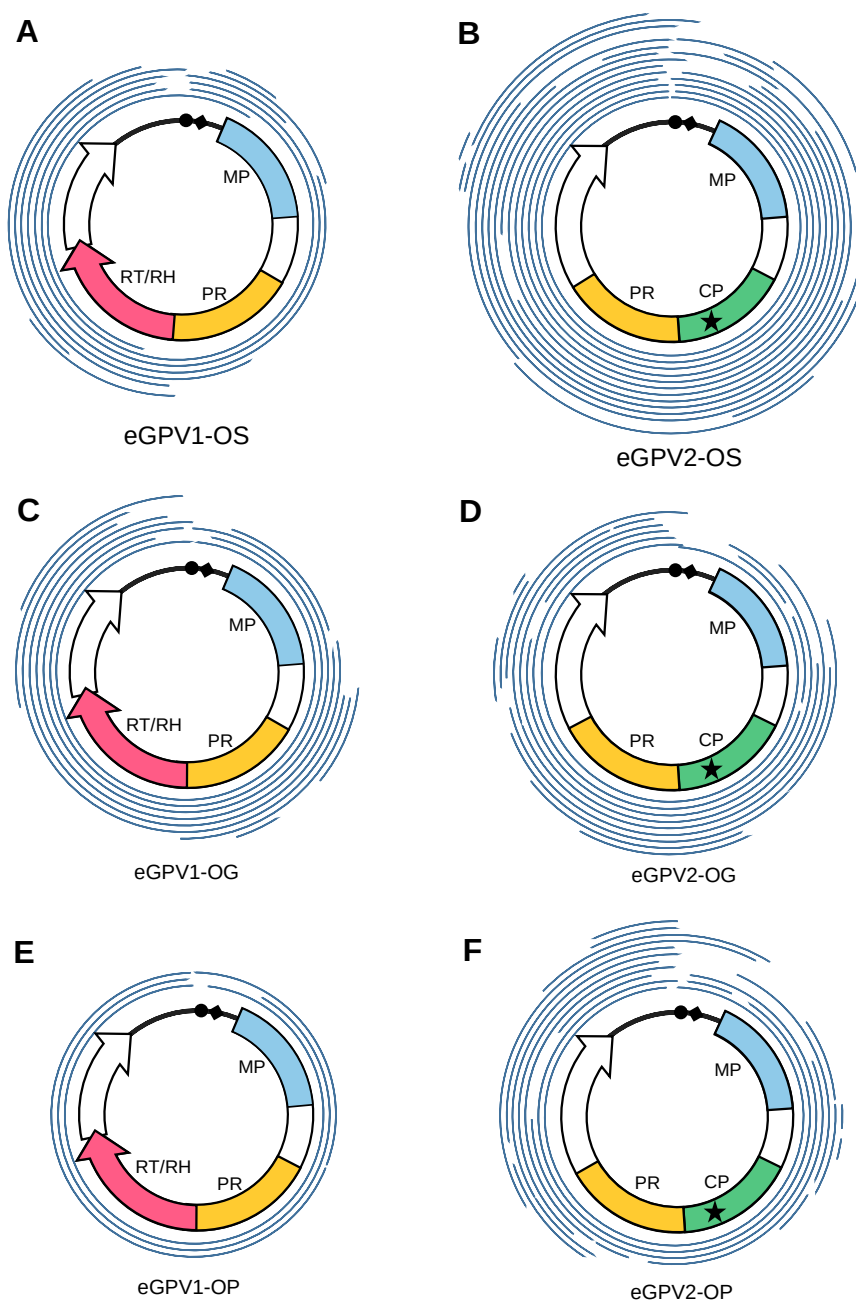


Figure 4.1 Reconstruction of the viral genomes of eGPV1 and eGPV2 sequences in three *Oryza* genomes

The viral genomes were reconstructed for the sequences of eGPV1-OS (**A**) and eGPV2-OS (**B**) in the *O. sativa* genome, and for those of eGPV1-OG (**C**) and eGPV2-OG (**D**) in the *O. glaberrima* genome, and for those of eGPV1-OP (**E**) and eGPV2-OP (**F**) in the *O. punctata* genome. Viral genomic structures are displayed in the center of each panel. The ORFs are represented with arrows, and the predicted domains in the ORF are outlined in colour. The black stars represent the zinc finger motif in the CP domain. The IGR is represented with black curved lines, and the PBS and PPT in the IGR are indicated by black dot and diamond respectively. The segments used for reconstruction are shown by blue curved lines in the outer of the viral genomic structures and are matched to the corresponding positions of viral genomes.

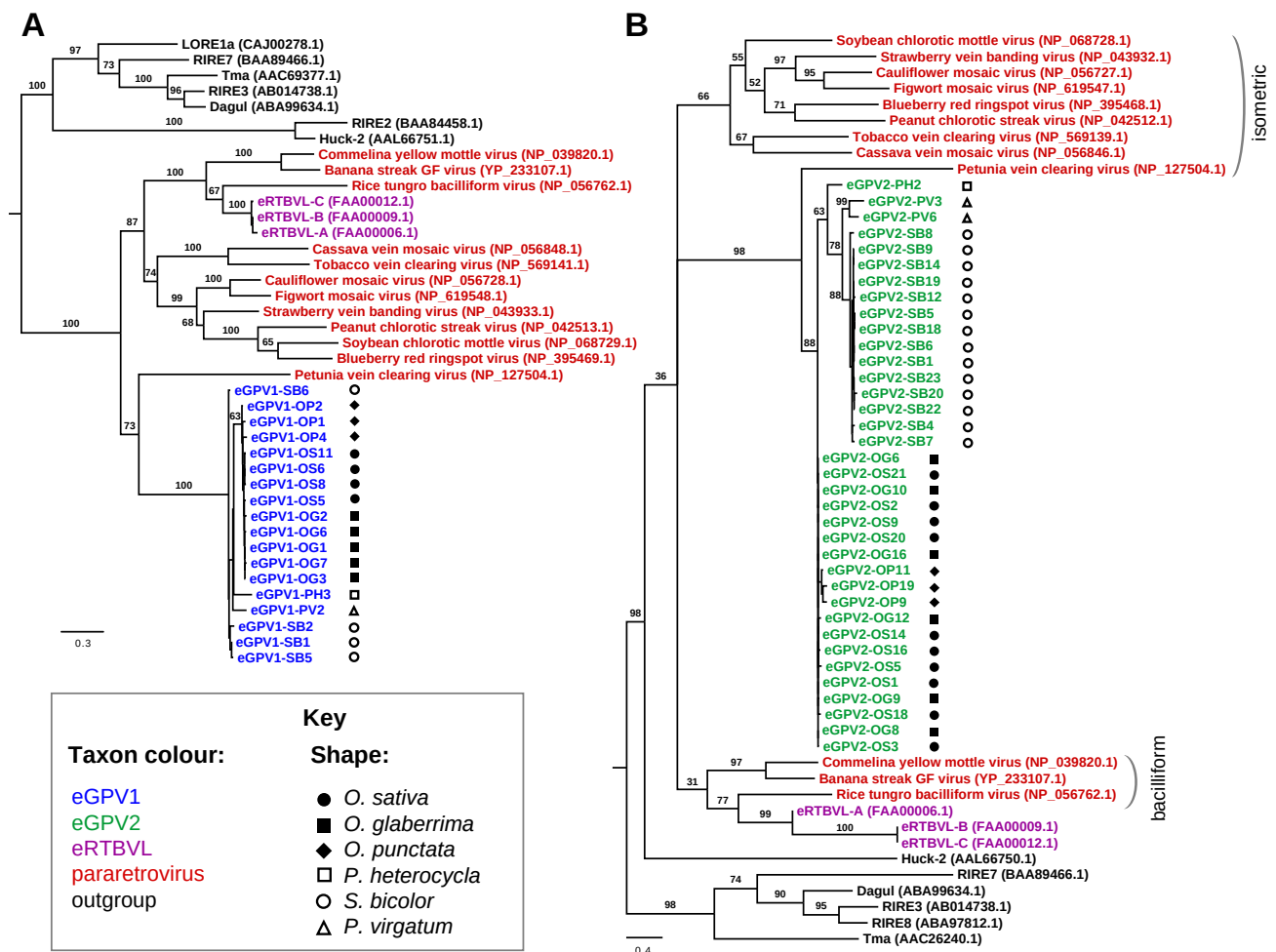


Figure 4.2 Phylogenetic relationship of eGPV1/eGPV2 and pararetrovirus sequences based on the RT/RH and CP domains

ML trees were constructed based on the amino acid alignments of the RT/RH (A) and CP (B) domains. Highly truncated sequences are not included in the phylogenetic analysis. The trees are midpoint rooted for display purpose. Bootstrap supports (1000 replicates) for major nodes are showed in percentage at the branches. Gypsy retrotransposon sequences were used as outgroups. Scale bars represent the genetic distance in substitutions per site. The taxa in blue, green, purple, red and black indicated the eGPV1, eGPV2, eRTBVL, exogenous pararetrovirus and outgroup sequences, respectively. The location informations of eGPV1/eGPV2 sequences are available in Supplementary Table S7, and the NCBI accession numbers of the other sequences are enclosed by brackets. The eGPV1/eGPV2 sequences from a same grass genome are labeled by a same shape after the taxon name (solid circle, *O. sativa*; solid square, *O. glaberrima*; solid diamond, *O. punctata*; hollow square, *P. heterocycla*; hollow circle, *S. bicolor*; hollow triangle, *P. virgatum*).

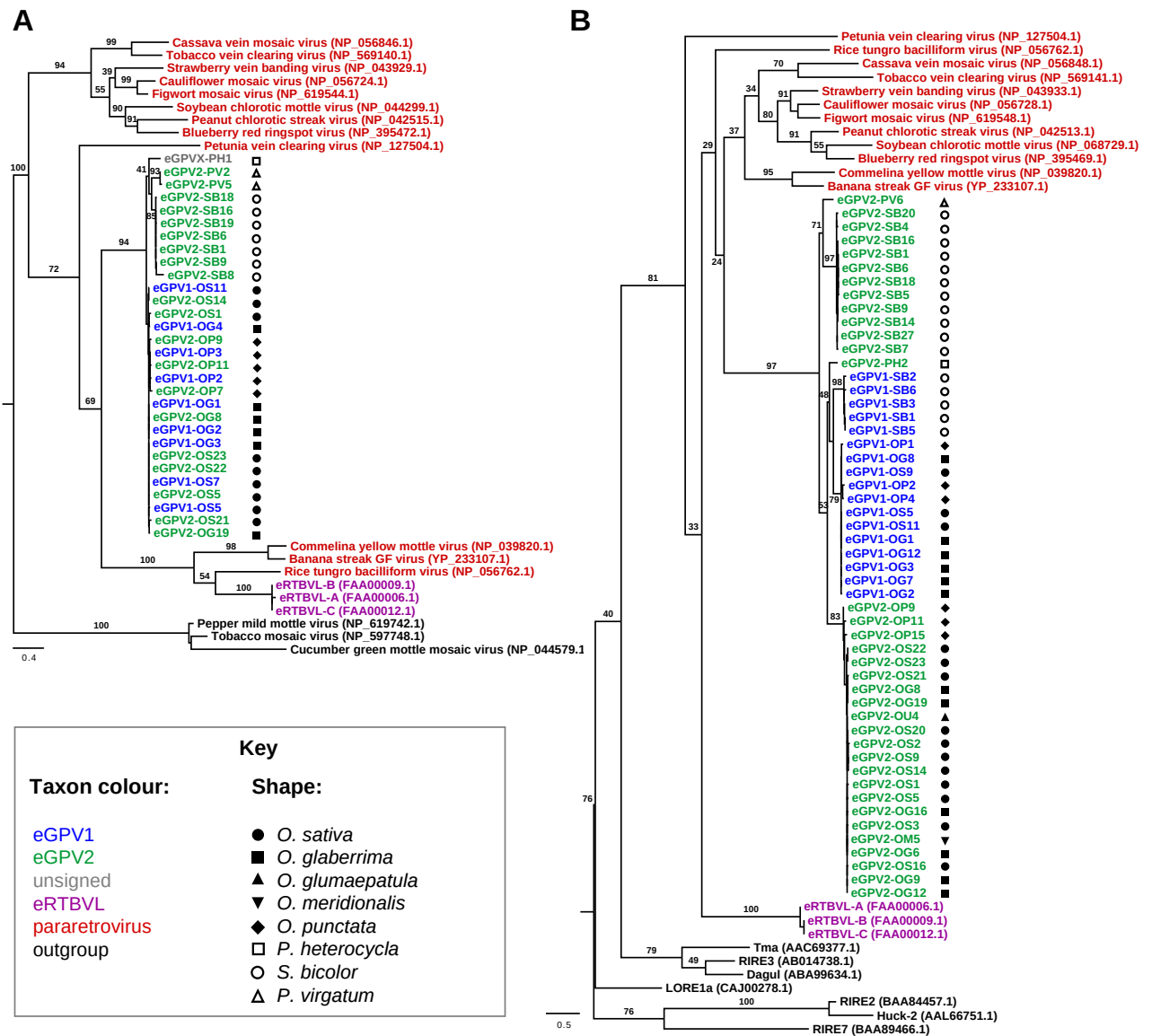


Figure 4.3 Phylogenetic relationship of eGPV1/eGPV2 and pararetrovirus sequences based on the MP and PR domains

ML trees were constructed based on the amino acid alignments of MP (A) and PR (B) domains. Highly truncated sequences are not included in the phylogenetic analysis. All the trees are midpoint rooted for display purpose. Bootstrap supports (1000 replicates) for major nodes are shown in percentage at the branches. Virgaviridae family virus (MP domain) or gypsy retrotransposon (PR domain) sequences were used for outgroups. Scale bars represent the genetic distance in substitutions per site. The taxa in blue, green, grey, purple, red and black indicated the eGPV1, eGPV2, unsigned, eRTBVL, pararetrovirus and outgroup sequences, respectively. The location informations of eGPV1/eGPV2 sequences are available in Supplementary Table S7, and the NCBI accession numbers of the other sequences are enclosed by brackets. The eGPV1 and eGPV2 sequences from a same grass genome are labeled by a same shape after the taxon name (solid circle, *O. sativa*; solid square, *O. glaberrima*; solid triangle, *O. glumaepatula*; solid inverted triangle, *O. meridionalis*; solid diamond, *O. punctata*; hollow square, *P. heterocycla*; hollow circle, *S. bicolor*; hollow triangle, *P. virgatum*).

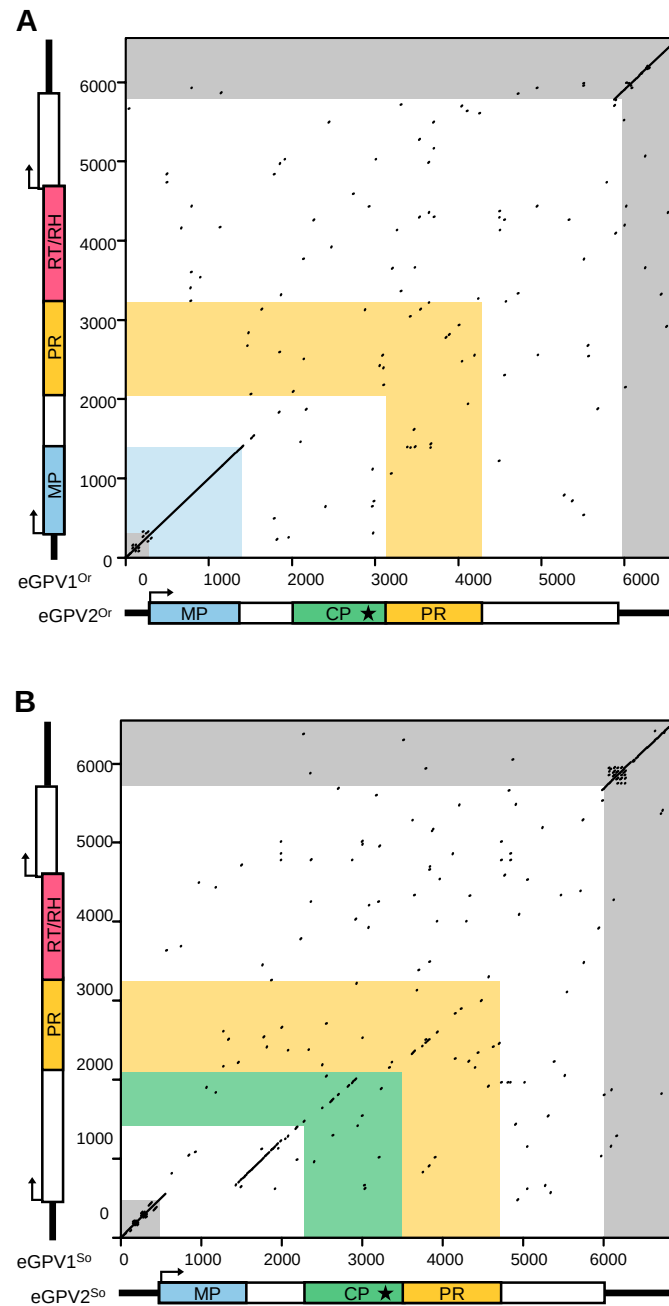


Figure 4.4 Dot plot of pairwise sequence comparison between eGPV1 and eGPV2 sequences

The pairwise comparisons (word size, 10; threshold, 45) are produced between eGPV1^{Or} and eGPV2^{Or} using the eGPV1-OS and eGPV2-OS consensus sequences (**A**), and between eGPV1^{So} and eGPV2^{So} using the eGPV1-SB and eGPV2-SB consensus sequences (**B**). The axes measure the lengths of each sequence. Linear viral genomic structures of eGPV1s and eGPV2s are displayed at the left of the vertical axis and below the horizontal axis. Rectangle with arrow indicated the ORF and black thick line showed the IGR. Black stars represent the zinc finger motif in the CP domain.

Table 4.1 Identification of functional domains in eGPV1 and eGPV2 proteins

Query	Domain number	Domain name/description	Query residues	Subject residues	Probability (%)	E value	P value
eGPV1-OS	PF01107	MP: Viral movement protein	53-212	14-179	99.9	9.80E-26	8.30E-31
	PF02160	Cauliflower mosaic virus peptidase (A3)	783-968	3-201	99.6	7.00E-15	5.90E-20
	4OL8	Reverse transcriptase/ribonuclease H	941-1417	13-478	100	1.70E-77	1.40E-82
eGPV2-OS (ORF1)	PF01107	MP: Viral movement protein	52-215	13-182	99.9	1.80E-25	1.50E-30
	g.40.1	Retrovirus zinc finger-like domains	920-944	15-39	96.4	0.0038	3.20E-08
	PF02160	Peptidase_A3: Cauliflower mosaic virus peptidase	1143-1277	6-144	99.7	1.70E-17	1.40E-22
eGPV1-SB	PF02160	Peptidase_A3: Cauliflower mosaic virus peptidase	714-896	6-201	99.4	1.70E-12	1.40E-17
	4OL8	Reverse transcriptase/ribonuclease H	881-1343	25-476	100	3.20E-76	2.70E-81
eGPV2-SB (ORF1)	PF01107	MP: Viral movement protein	48-216	9-183	99.9	1.60E-25	1.30E-30
	g.40.1	Retrovirus zinc finger-like domains	937-962	16-41	97.5	9.80E-05	8.20E-10
	PF02160	Peptidase_A3: Cauliflower mosaic virus peptidase	1158-1332	6-197	99.8	2.00E-19	1.70E-24

Domain identification was fulfilled with HHpred (Söding *et al*, 2005). The hits with the best *P* value for each block of query residues are displayed.

4.3.3 Multiple paleogenomes of nonautonomous pararetroviruses in *Oryza* genomes

If the viruses of eGPV1s and eGPV2s were capable of replication and transmission in despite of a defective genome, their insertions could also occur in the relative species of *O. sativa*. I therefore searched homologous sequences of eGPV1-OS as well as eGPV2-OS in the other five fully sequenced genomes of *Oryza* species (Supplementary Table S2). I discovered totally 86 segments similar to eGPV1-OS or eGPV2-OS in four of these five *Oryza* genomes (*O. glaberrima*, *O. glumaepatula*, *O. meridionalis* and *O. punctata*), except *O. brachyantha* (Supplementary Table S7). The segment sequences identified in the four genomes were congruent with the structures of eGPV1-OS and eGPV2-OS (average nucleotide identity to eGPV1, 93.6%~99.1%; to eGPV2, 90.3%~98.0%) (Supplementary Table S7). Particularly, the relatively numerous long segments in the genomes of *O. glaberrima* and *O. punctata* allow us to use them to *de novo* construct both complete viral genomes of eGPV1 and eGPV2 sequences independently, and these viral consensus sequences I obtained from the two *Oryza* genomes showed the same structures as those obtained from the *O. sativa* genome (Figure 4.1C-F). These sequence-similar eGPV1s/eGPV2s showing same structures were collectively named eGPV1/eGPV2 *Oryza* group (eGPV1^{Or}/ eGPV2^{Or}).

Because *Oryza* AA-genome species (*O. sativa*, *O. glaberrima*, *O. glumaepatula* and *O. meridionalis*) originated from a common progenitor with BB-genome species (*O. punctata*), a question arose whether these eGPV1 and eGPV2 segments in different *Oryza* genomes inherited from a shared progenitor, or resulted from independent insertions into the respective genomes after species divergence. I executed an orthology analysis of eGPV1/eGPV2 loci through a mutual genome-wide comparison (Hayward *et al*, 2013) to determine the species-shared and -specific insertions. Firstly, I mapped the flanking sequences of the eGPV1 and eGPV2 loci in the genome of *O. sativa* into the genomes of other AA- and BB-genome species. This revealed that most of these

eGPV1 and eGPV2 loci were unique in the *O. sativa* genome and absent at the same site in the other genomes, and that four of the loci were shared by other genome(s) (summarized in Figure 4.5, details in Supplementary Table S8). Among the four shared loci, two loci (one for each of eGPV1s and eGPV2s) have orthologs in the genomes of *O. sativa* and *O. glaberrima*, the other two loci (one for each of eGPV1s and eGPV2s) are present in not only *O. sativa* and *O. glaberrima* but also other AA-genome species (Figure 4.5, Supplementary Table S8). I then mapped the flanking sequences of each eGPV1/eGPV2 locus from respective genomes of *O. glaberrima*, *O. glumaepatula*, *O. meridionalis* and *O. punctata* into the *O. sativa* genome. This confirmed the above shared loci in the corresponding genomes, and demonstrated that species-specific eGPV1 and eGPV2 loci were present in all the four genomes of *Oryza* species in addition to *O. sativa* (Figure 4.5, Supplementary Table S9). Accordingly, the viral sequences of eGPV1s and eGPV2s integrated into the *O. punctata* genome after the divergence between the AA- and BB-genome species, and into AA-genome species recurrently during and after the divergence among AA-genome species (Figure 4.5). I subsequently used genomic PCR to confirm these *in silico* observations. The orthologous eGPV1 and eGPV2 loci shared by part or all AA-genome species were confirmed (Figure 4.6), and for species-specific loci, the representative *O. sativa*-specific eGPV1 and eGPV2 loci were chosen for examining and the corresponding empty sites in the other *Oryza* genomes were validated (Figure 4.6).

The observations above suggested that both the viruses of eGPV1s and eGPV2s can replicate and transmit among multiple *Oryza* species and seemed to appear together in most cases. Considering a high degree of IGR similarity between the them, I proposed that these two nonautonomous paleoviruses have evolved a mutualistic symbiosis in coinfection, i.e. that the virus of eGPV2s could borrow the RT/RH proteins from the virus of eGPV1s for genome replication and the virus of eGPV1s could borrow the CP molecules from the virus of eGPV2s for particle assembly.

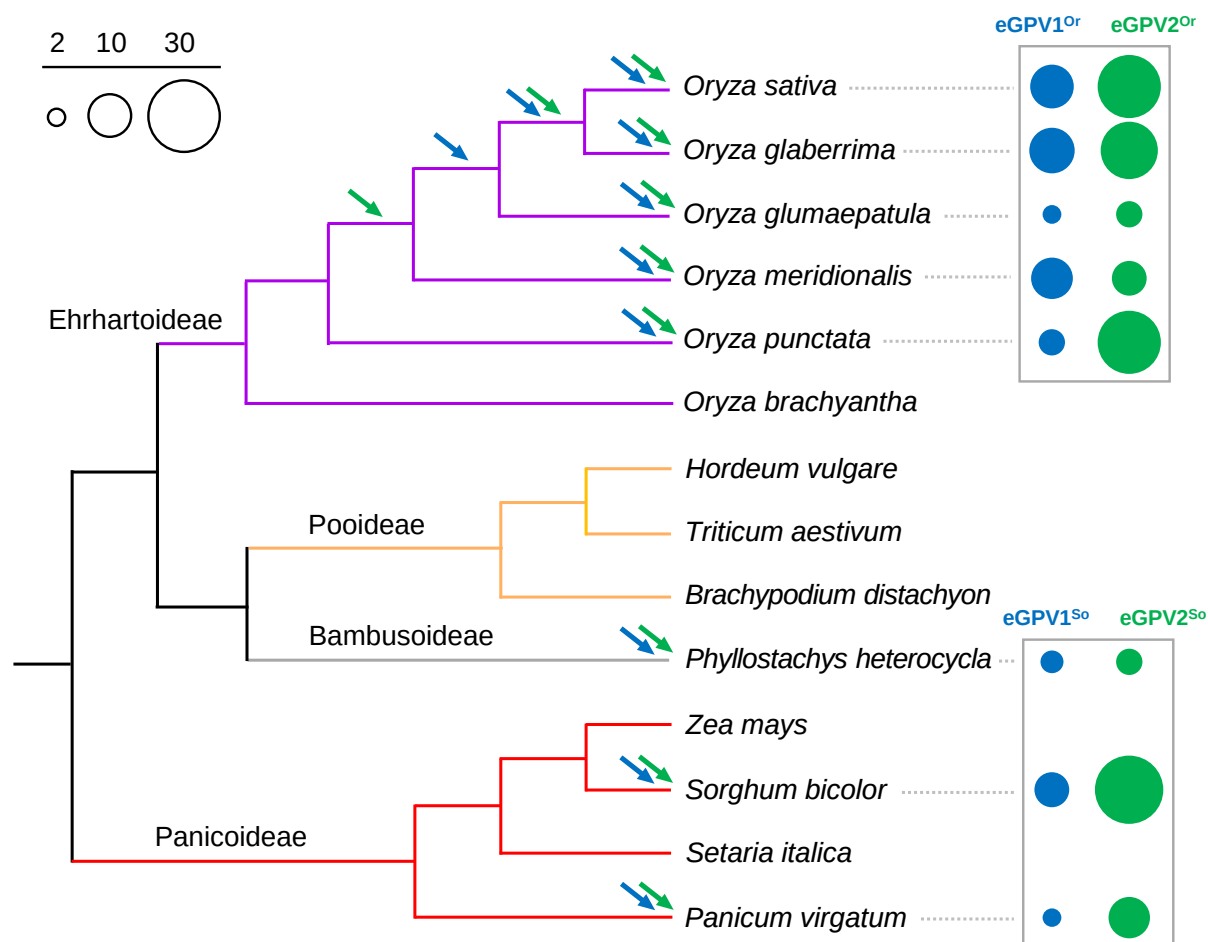


Figure 4.5 Insertion events of eGPV1 and eGPV2 sequences into grass genomes

The phylogeny of grass species were built based on previous reports (Catalán *et al*, 1997; Swigoňová *et al*, 2004; Zou *et al*, 2008; Bennetzen *et al*, 2012; Zhang *et al*, 2012; Peng *et al*, 2013). Purple, orange, grey and red braches indicate the species of Ehrhartoideae, Pooideae, Bambusoideae and Panicoideae subfamily, respectively. The coloured arrows above the braches represent the insertion events of eGPV1 (blue) and eGPV2 (green) segments into the corresponding genomes of grass species. The sizes of coloured circles at the right of taxa illustrate the segment numbers of eGPV1 (blue) and eGPV2 (green) segments in grass genomes according to the scale.

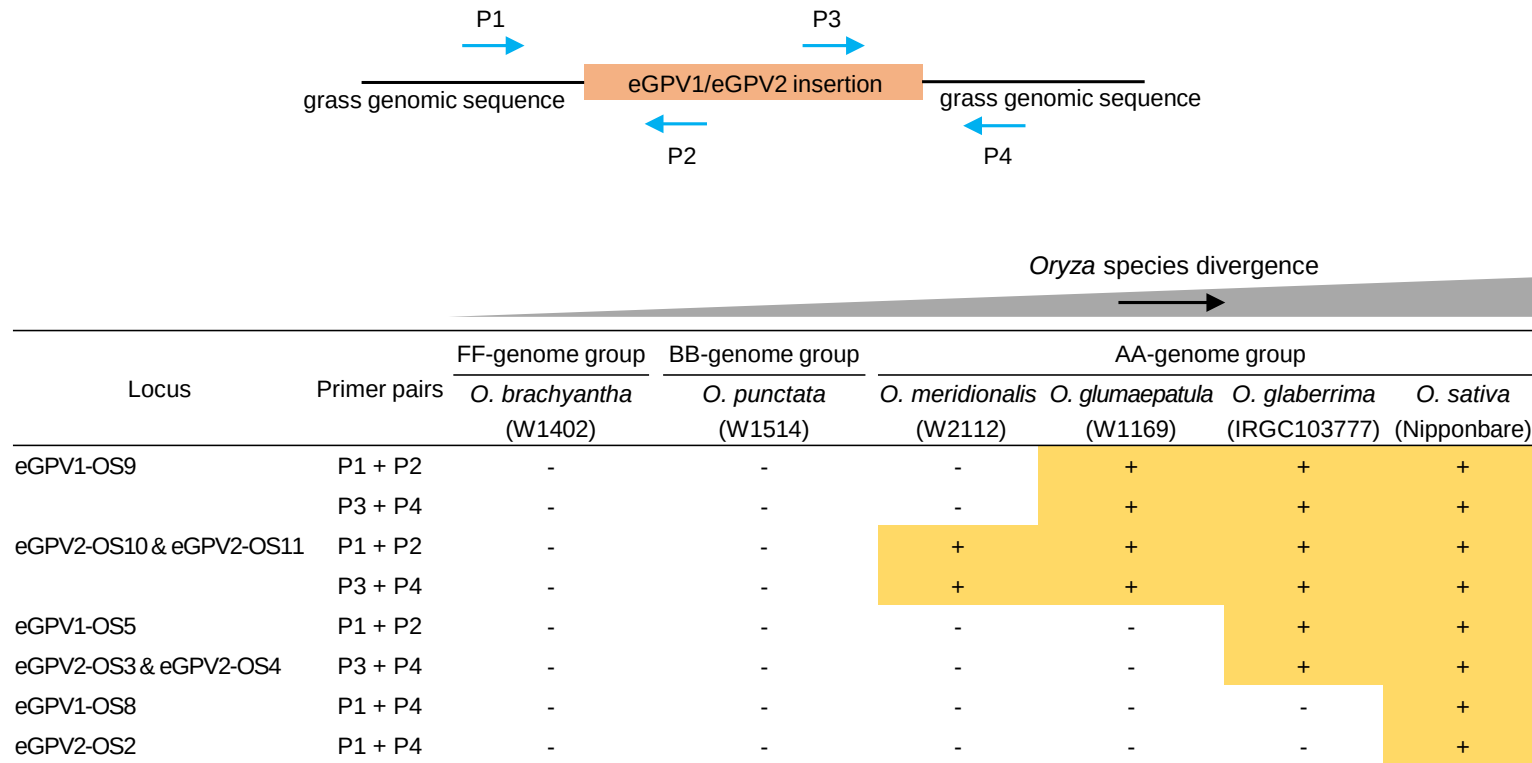


Figure 4.6 Genomic PCR detection of the presence/absence of representative eGPV1-OS/eGPV2-OS insertions among orthologous loci in *Oryza* genomes

The PCR strategy used to detect the presence/absence of an eGPV1-OS/eGPV2-OS insertion at a specific locus is shown in the upper. Blue arrows indicate the primers for PCR detection. The PCR results is shown by a table (plus sign, presence; minus sign, absence) in the lower. The used accessions of respective *Oryza* species are provided in brackets. The detailed information of the detected eGPV1-OS/eGPV2-OS insertions is available in Supplementary Table S7, and primer information in Supplementary Table S10.

4.3.4 Co-evolution of two functionally complementary nonautonomous pararetroviruses

To inspect the evolution of nonautonomous pararetroviruses, I screened eight available genomes of non-*Oryza* grass species (Supplementary Table S2) for eGPV1/eGPV2-related sequences. In two of the four genomes from Panicoideae subfamily, sorghum (*S. bicolor*) and switchgrass (*P. virgatum*), scores of sequences were identified to be similar to the eGPV1^{Or}/eGPV2^{Or} sequences (Figure 4.5, Supplementary Table S7). I *de novo* constructed viral consensus sequences for these segments in the sorghum genome (more and longer segments in this genome), and obtained two separate circular consensus sequences (6544 and 6890 bp respectively; Figure 4.7). One structurally resembles eGPV1^{Or}, but carries neither MP domain nor CP domain; the other structurally equates to eGPV2^{Or} (Figure 4.7; domain prediction in Table 4.1; motif alignments in Supplementary Figure S14); they formed sister groups to eGPV1^{Or} and eGPV2^{Or} respectively in the phylogenies based on the amino acid sequences of each domain (Figures 4.2 and 4.3). Therefore, I call these EPRVs as eGPV1/eGPV2 *S. bicolor* lineage (eGPV1-SB/eGPV2-SB). The identified sequences in the switchgrass genome are phylogenetically and structurally comparable to eGPV1-SB or eGPV2-SB (Figures 4.2 and 4.3, Supplementary Table S7). These sequence in the sorghum and switchgrass genomes were collectively dubbed eGPV1/eGPV2 *Sorghum* group (eGPV1^{So}/eGPV2^{So}). In the genome of bamboo (*P. heterocycla*) from Bambusoideae subfamily, I also identified three and four segments similar to eGPV1 and eGPV2 respectively (Supplementary Table S7), which can be sorted into the *Sorghum* group by phylogenetic analysis (Figures 4.2 and 4.3) although the structure information is incomplete due to limited short segments. No pronounced sequence similar to eGPV1s/eGPV2s was found in all the three genomes from Pooideae subfamily (Figure 4.5). In summary, the eGPV1^{So} sequences were recognized in three non-*Oryza* grass genomes and all lack MP and CP domains, and the eGPV2^{So} sequences were also identified in the same genomes and all hold no RT/RH domain (data not available for

bamboo). The profile of eGPV1^{So}/eGPV2^{So} sequences in the phylogeny of non-*Oryza* grass species suggested independent events of viral insertions (Figure 4.5). Most eGPV1/eGPV2 sequences are truncated and possess disruptive mutations at coding regions, but many eGPV1/eGPV2 sequences in the sorghum genome are relatively complete and even some have integral ORFs without interruptions (Supplementary Table S7), indicating the relatively recent viral insertions in this genome and a possible later emergence of the corresponding viral lineage that infected sorghum.

The structural and phylogenetic analyses above suggested that the virus of eGPV1^{So}/eGPV2^{So} was the relative of the virus of eGPV1^{Or}/eGPV2^{Or}, and demonstrated that the viruses of eGPV1s/eGPV2s had horizontally transmitted to a wide range of grass species and developed into distinct groups and lineages. The protein conservations between eGPV1^{So} and eGPV1^{Or}, and between eGPV2^{So} and eGPV2^{Or} are considerable at most parts, except the first half of ORF1 (MP-related region) between eGPV1^{So} and eGPV1^{Or} (Figure 4.8). The sequence similarities of IGRs between eGPV1^{So} and eGPV1^{Or}, and between eGPV2^{So} and eGPV2^{Or} are weak (Figure 4.9), whereas the IGRs of eGPV1^{So} and eGPV2^{So} exhibit an extremely high sequence similarity (97.1% identity between the consensus sequences) (Figure 4.4B). A considerable sequence similarity was also observed between part downstream regions of IGRs of eGPV1^{So} and eGPV2^{So} (Figure 4.4B), and this similarity pattern is obviously different from the case of eGPV1^{Or} and eGPV2^{Or} (Figure 4.4A). The viral paleogenome comparison therefore implied a co-evolution between the viruses of eGPV1s and eGPV2s during their presumed symbiosis. Noticeably, a region of eGPV1^{So} showed a slight homology to the CP domain of eGPV2^{So}, but the region corresponding to the essential zinc finger motif was deleted in eGPV1^{So} (Figure 4.4B). Taken together, the evidence from comparative molecular biology and comparative genomics of eGPV1s and eGPV2s supported that two nonautonomous pararetroviruses shared *cis*-regulation elements and functionally complemented each other to develop a mutualistic symbiosis.

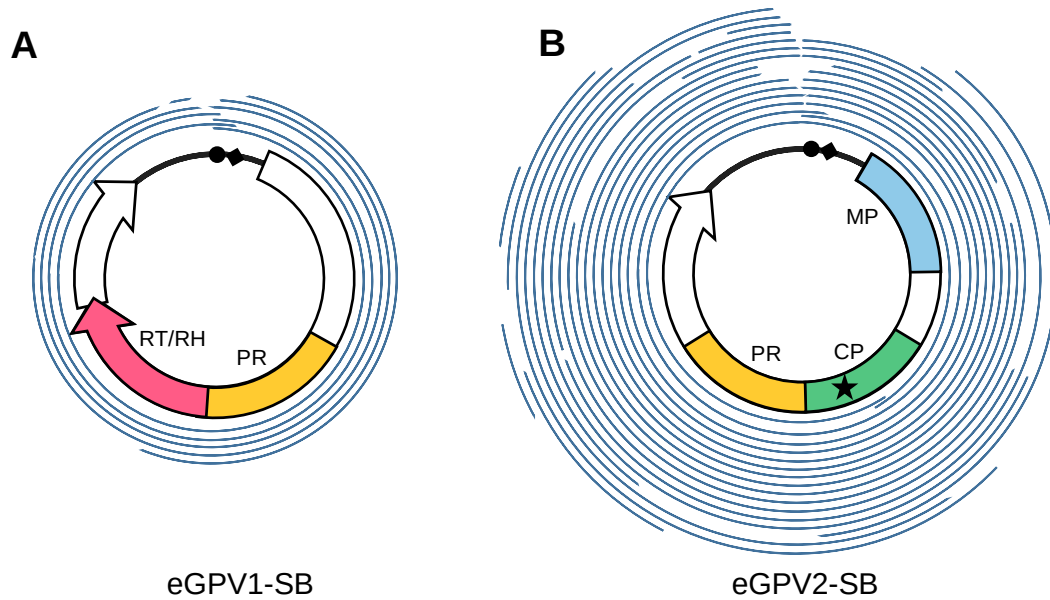


Figure 4.7 Reconstruction of the viral genomes of eGPV1 and eGPV2 sequences in the *S. bicolor* genome

The viral genomes were reconstructed for the sequences of eGPV1-SB **(A)** and eGPV2-SB **(B)** in the *S. bicolor* genome. Viral genomic structures are displayed in the center of each panel. The ORFs are represented with arrows, and the predicted domains in the ORF are outlined in colour. The black stars represent the zinc finger motif in the CP domain. The IGR is represented with black curved lines, and the PBS and PPT in the IGR are indicated by black dot and diamond respectively. The segment sequences used for reconstruction are showed by blue curved lines in the outer of the viral genomic structures and are matched to the corresponding positions of viral genomes.

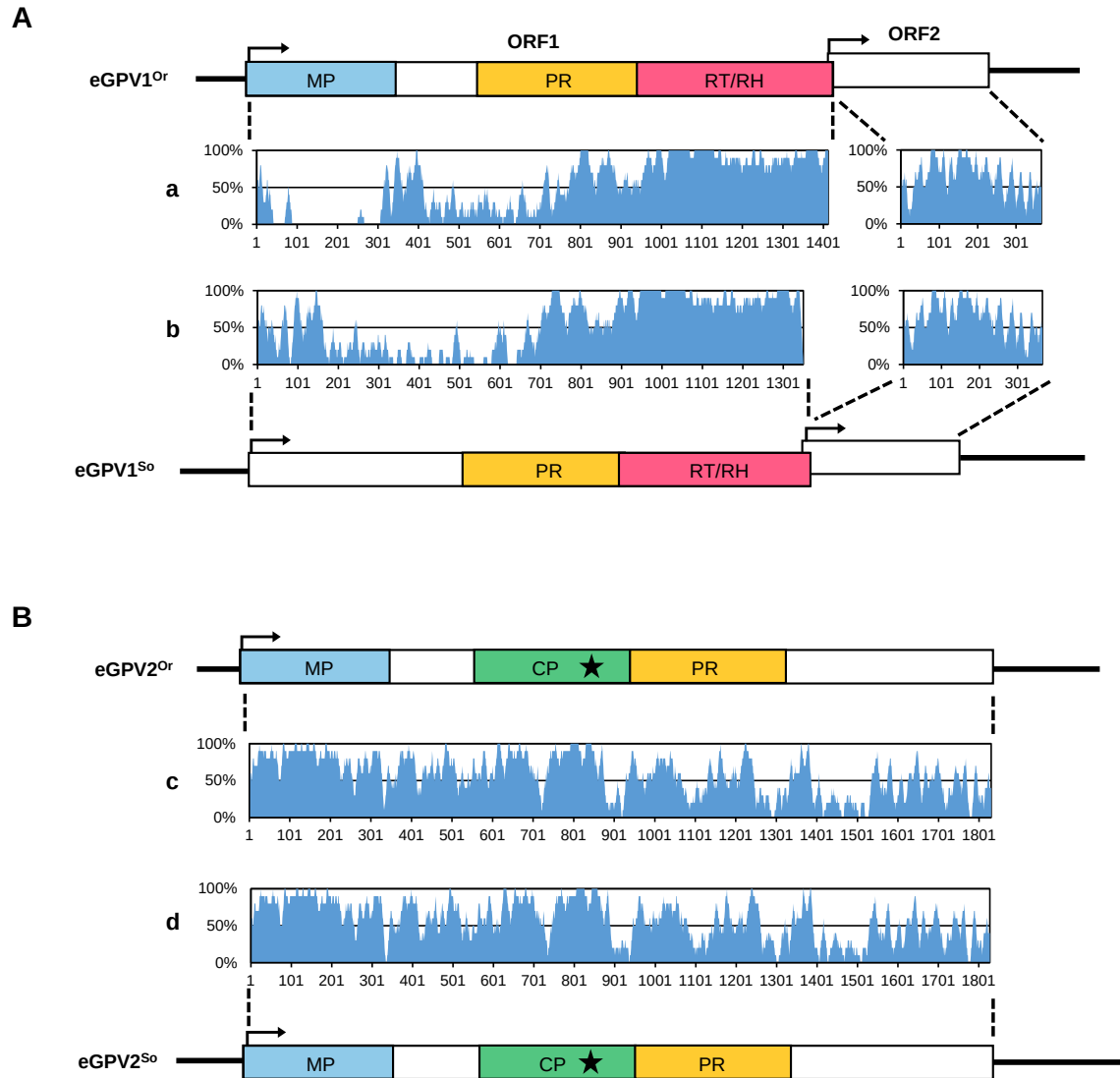


Figure 4.8 Pairwise alignments of protein sequences of eGPV1s and eGPV2s

The sequence comparisons were produced between eGPV1^{Or} and eGPV1^{So} using the eGPV1-OS and eGPV1-SB consensus sequences respectively (**A**), and between eGPV2^{Or} and eGPV2^{So} using the eGPV2-OS and eGPV2-SB consensus sequences respectively (**B**). The plots a and b show the relative amino acid identity (window size, 10) between the two eGPV1 sequences (a, eGPV1^{So} was aligned to eGPV1^{Or}; b, eGPV1^{Or} was aligned to eGPV1^{So}), and the plots c and d display that between the two eGPV2 sequences (c, eGPV2^{So} was aligned to eGPV2^{Or}; d, eGPV2^{Or} was aligned to eGPV2^{So}). Linear viral genomic structures of eGPV1 and eGPV2 are displayed, and their ORFs and IGRs are indicated by rectangles (arrows show the ORF transcription orientation) and solid lines respectively.

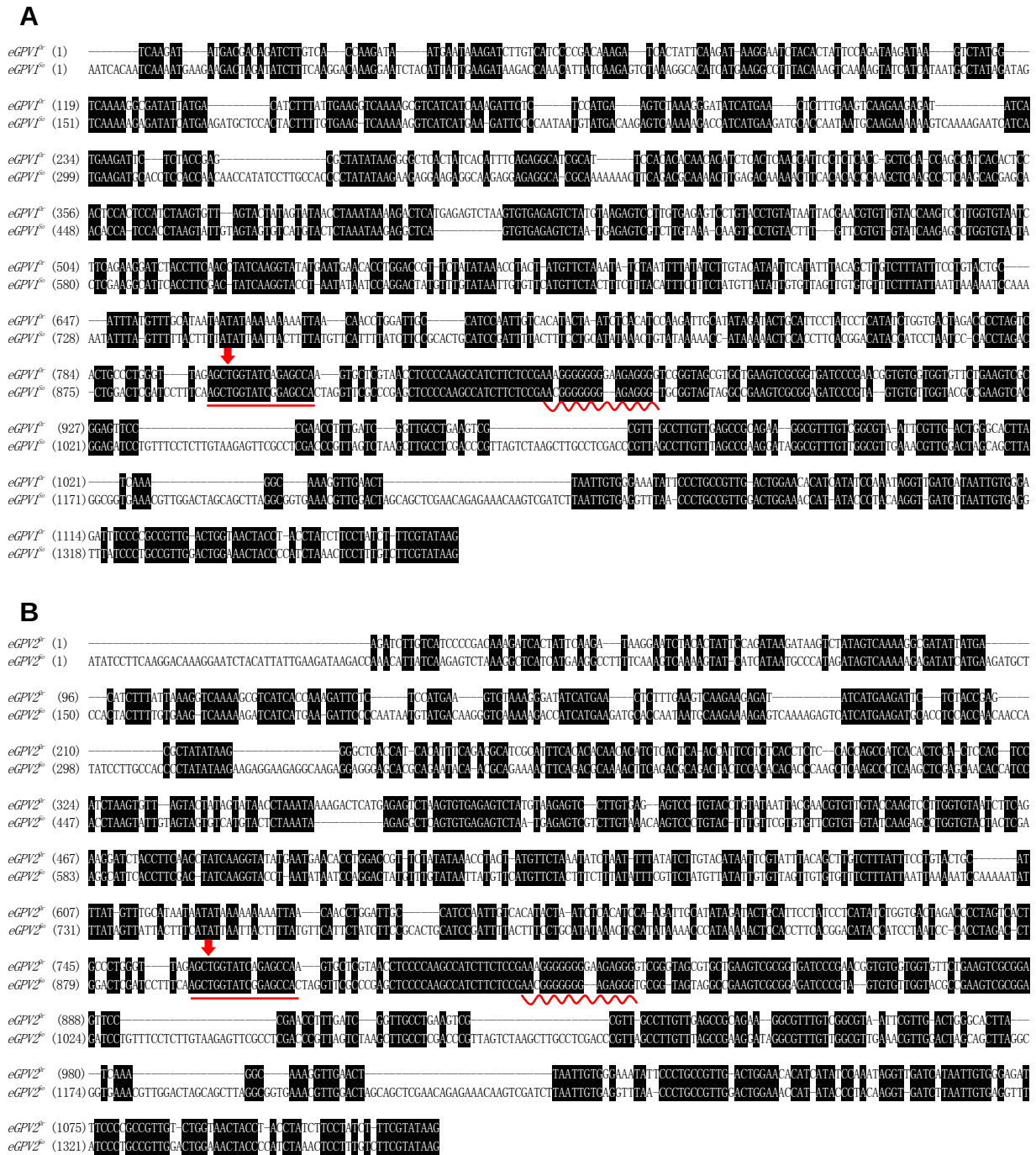


Figure 4.9 Pairwise alignments of the IGR sequences of eGPV1s and eGPV2s

The sequence comparisons are produced between eGPV1^{Or} and eGPV1^{So} using the eGPV1-OS and eGPV1-SB consensus sequences respectively **(A)**, and between eGPV2^{Or} and eGPV2^{So} using the eGPV2-OS and eGPV2-SB consensus sequences respectively **(B)**. The identical sites are marked by black background. The relative positions of the first nucleotides in each row of alignment are enclosed in brackets. The red arrows, lines and waves indicate the strand discontinuity (the start of 5'-terminal of minus strand), PBS and PPT sites respectively in the IGR of pararetroviral circle genomes.

4.3.5 IGR swapping establishes and maintains the mutualism between nonautonomous pararetroviruses

Nonautonomous LTR retrotransposons parasitize their autonomous partners through the conservation of LTR sequences containing *cis*-regulation elements. Natural selection and region-specific recombination are two mechanisms for the maintenance of LTR sequence conservation between them (Jiang *et al*, 2002; Du *et al*, 2010). Although IGR is a highly diverse noncoding region among pararetroviral genomes, the whole IGR and its downstream region between eGPV1^{Or} and eGPV2^{Or}, and between eGPV1^{So} and eGPV2^{So} showed an extraordinarily high sequence similarity, which suggested that region-specific recombination could contribute this high conservation mainly.

To test it, I built phylogenetic trees of the IGR sequences from both eGPV1 and eGPV2 segments. In the clades of the IGR sequences from *Oryza* AA-genome group species (Figure 4.10A), the orthologous sequences of oldest eGPV1/eGPV2 insertion locus (shared by three/four *Oryza* genomes) formed respective clusters; one *O. glaberrima*-specific-eGPV2 sequence appeared with several *O. glaberrima*-specific-eGPV1 sequences together in a subclade where the orthologous sequences of older eGPV1 insertion locus (shared by two *Oryza* genomes) is located; three *O.sativa*-specific-eGPV2 sequence appeared with three *O.sativa*-specific-eGPV1 sequence together in another subclade; two subclades comprised only species-specific eGPV2 sequences, and these eGPV2 sequences are from multiple *Oryza* genomes, suggesting a viral horizontal transmission. In the clades of the IGR sequences from *O. punctata* (Figure 4.10A), the eGPV1 sequences were clustered with eGPV2 sequences together. The phylogenetic relationship of *Oryza* species-shared and -specific IGR sequences support the results of orthology analysis above. I also observed that the IGR sequences of eGPV1-SB are phylogenetically clustered with those of eGPV2-SB in the sorghum genome (Figure 4.10B). When taking a close look, I found that the virus-derived small

indel (deletions and insertions) variations in the IGR are shared between some sequences of eGPV1-SB and of eGPV2-SB (Figure 4.11). Taken together, these data demonstrated that the IGR and neighbor regions were often swapped between the viral genomes of eGPV1 and eGPV2 sequences.

Further recombination analysis of the raw (not consensus) eGPV1/eGPV2 sequences in these grass genomes did not show any sequence credibly derived from a recombinant virus lineage which was generated by the non-homologous recombination between the viral genomes of eGPV1s and eGPV2s (data not shown), that is, the potential recombination breakpoints during the swapping of the viral genomes of eGPV1s and eGPV2s are confined in the IGR and neighbor regions which show a high sequence similarity between the viruses of eGPV1s and eGPV2s. I thus proposed that non-homologous recombination events or template switches (RT-mediated recombination; Du *et al*, 2010) between different ancient pararetroviruses could produce initial viruses of eGPV1s and eGPV2s, two defective but complementary pararetroviruses carrying the similar or identical IGR sequences that made them form a mutualistic relationship, and that during the long-term symbiosis, homologous recombination frequently occurred between the highly homologous sequences of the viruses of eGPV1s and eGPV2s to lead to the homogenization of *cis*-regulation elements in the noncoding region.

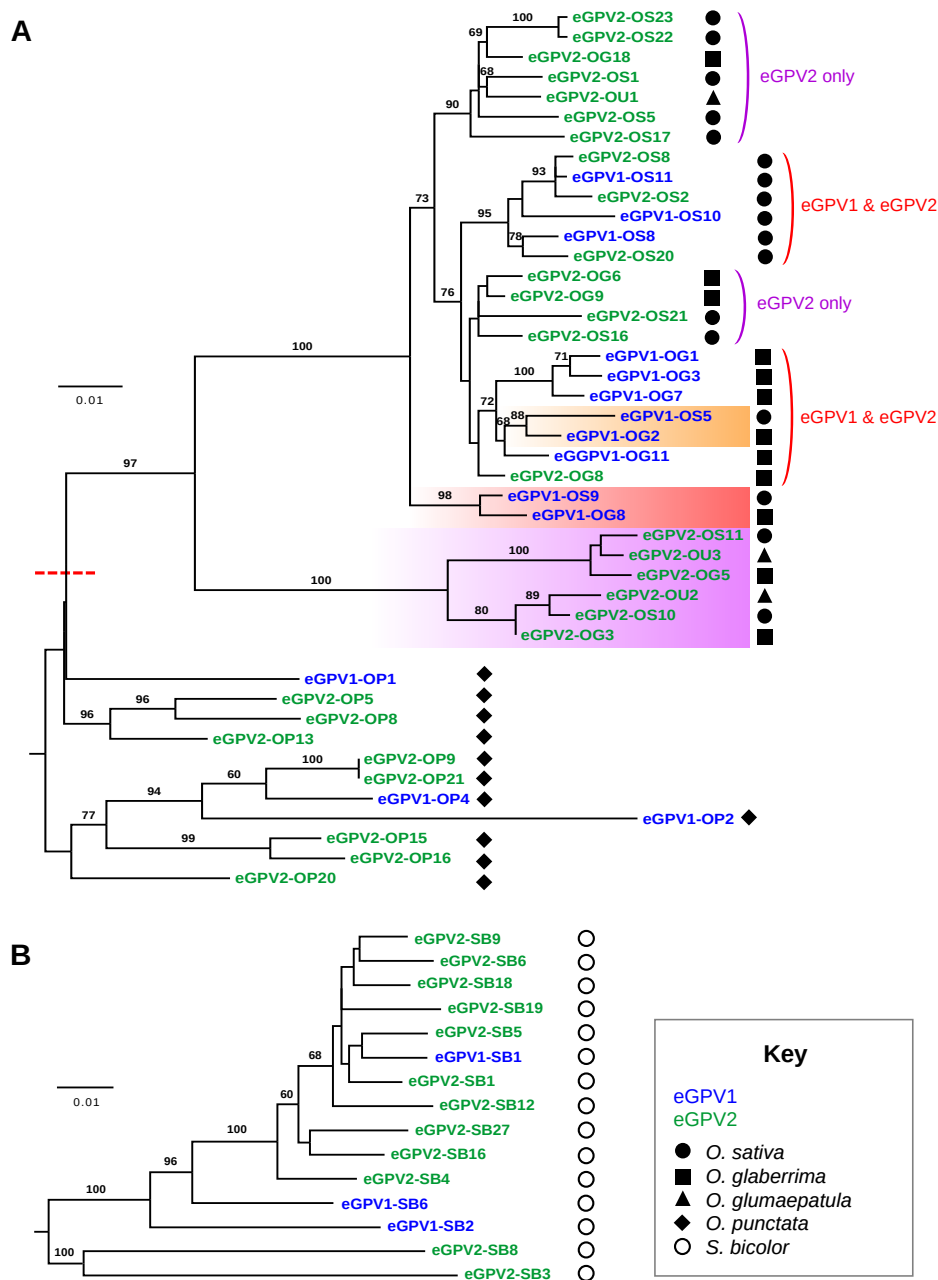


Figure 4.10 Phylogenetic relationships of the IGR sequences of eGPV1s and eGPV2s

Respective ML trees were developed for the IGR sequences of eGPV1^{Or} and eGPV2^{Or} groups (**A**), and for those of eGPV1^{So} and eGPV2^{So} groups (**B**). Highly truncated sequences are not included in the phylogenetic analysis. The trees are midpoint rooted for display purpose, and the bootstrap supports derived from 1000 replicates are showed in percentage at the branches (only values greater than 60% are shown). Scale bars represent evolutionary distance in nucleotide substitutions per site. The **taxon** names display the segment names of the eGPV1 (blue) and eGPV2 (green) sequences, and the location informations of eGPV1/eGPV2 sequences are available in Supplementary Table S7. The respective orthologous sequence sets of the oldest eGPV1 segment, the oldest eGPV2 segment, and the older eGPV1 segment (see the results of the orthology analysis of eGPV1/eGPV2 insertions) in *Oryza* genomes are highlighted separately in red, purple, and orange backgrounds. The sequences from a same grass genome are labeled by a same shape after the taxon name (solid circle, *O. sativa*; solid square, *O. glaberrima*; solid triangle, *O. glumaepatula*; solid diamond, *O. punctata*; hollow circle, *S. bicolor*).

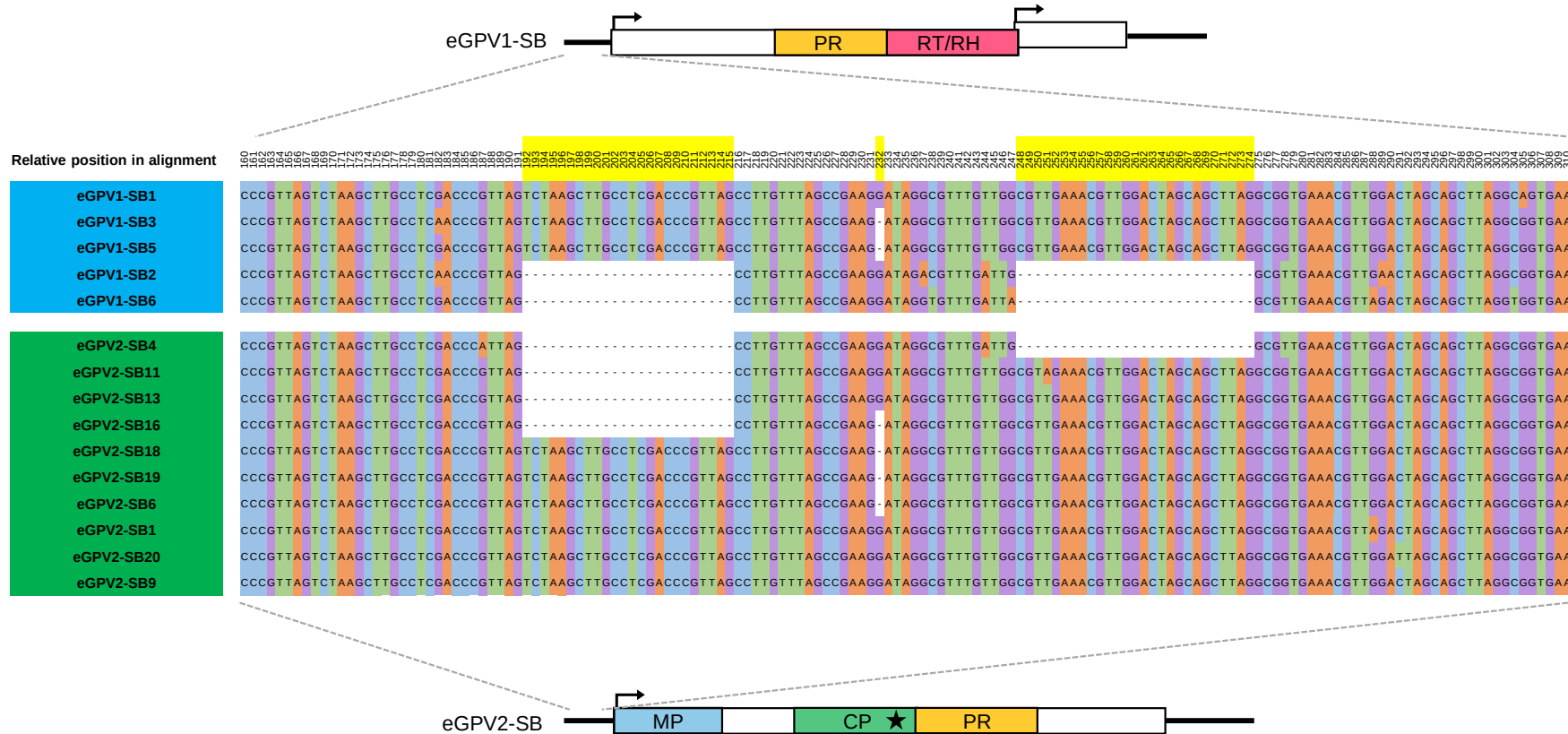


Figure 4.11 Virus-derived small indel (deletions and insertions) variations in the eGPV1-SB and eGPV2-SB IGR sequences

A part of the sequence alignment of eGPV1-SB and eGPV2-SB IGR sequences (corresponding to the 5'-terminal of the linear viral genomes) are shown in the center. The positions of small indel variations are highlighted by yellow background. The detailed information of eGPV1-SB and eGPV2-SB segments are available in Supplementary Table S7. Linear viral genomic structures of eGPV1-SB and eGPV2-SB are displayed, and their ORFs and IGRs are indicated by rectangles (arrows show the ORF transcription orientation) and solid lines respectively.

4.4 Discussion

Numerous endogenous viral elements have been identified, and most of them are defective due to the mutation accumulation after insertion (Holmes, 2011; Feschotte and Gilbert, 2012). Here, the eGPV1 and eGPV2 sequences I identified in a series of grass genomes were derived from two naturally defective ancient pararetroviruses, which are the first discovered nonautonomous species of pararetroviruses. All the eGPV1 and eGPV2 sequences are flanked by unrelated host genomic sequences and are often carrying mutation-derived interruptions, supporting that those eGPV1 and eGPV2 sequences are EPRVs of ancient pararetroviruses. The viral genome sizes of eGPV1s and eGPV2s (6544-6890 bp) are shorter than all the reference genomes of pararetroviruses (7140-9314 bp, data from NCBI genome databases). The two nonautonomous species have continually evolved and diverged during the millions years of grass host evolution. The long-term existence and development of the two species could be contributed by a mutualistic symbiosis between them, which were achieved through the coding capacity complementarity and noncoding *cis*-regulation sequence swapping. Interestingly, the noncoding homogenization did not result from mutation fixation but from region-specific recombination.

It was highly possible that in this symbiosis, the pregenomic RNAs of both the viruses of eGPV1s and eGPV2s could be harbored into the same viral cytoplasmic inclusion body (viroplasma), where reverse-transcription-mediated replication and assembly of pararetroviruses happened (Hull, 2002a); the RT/RH protein from the virus of eGPV1s reverse transcribed its own pregenomic RNA as well as that of the virus of eGPV2s, and the CP from the virus of eGPV2s assembled viral particles containing respective replicated genomes of the two viruses, separately or together. Both the viruses of eGPV1s and eGPV2s encoded a PR domain for the requirement of autocatalytic cleavage of their long ORF polyproteins (Torruella *et al*, 1989). The same *cis*-regulation elements of the viruses of eGPV1s and eGPV2s also allowed them

recruit identical host factors for the processes such as transcription and translation. The mutualistic entity of the two ancient viruses mimics the extant complex viruses possessing multiple polynucleotide sequences (such as influenza viruses), which thus suggests that functional complementarity and co-regulation could conduce the evolutionary origin of multipartite viruses.

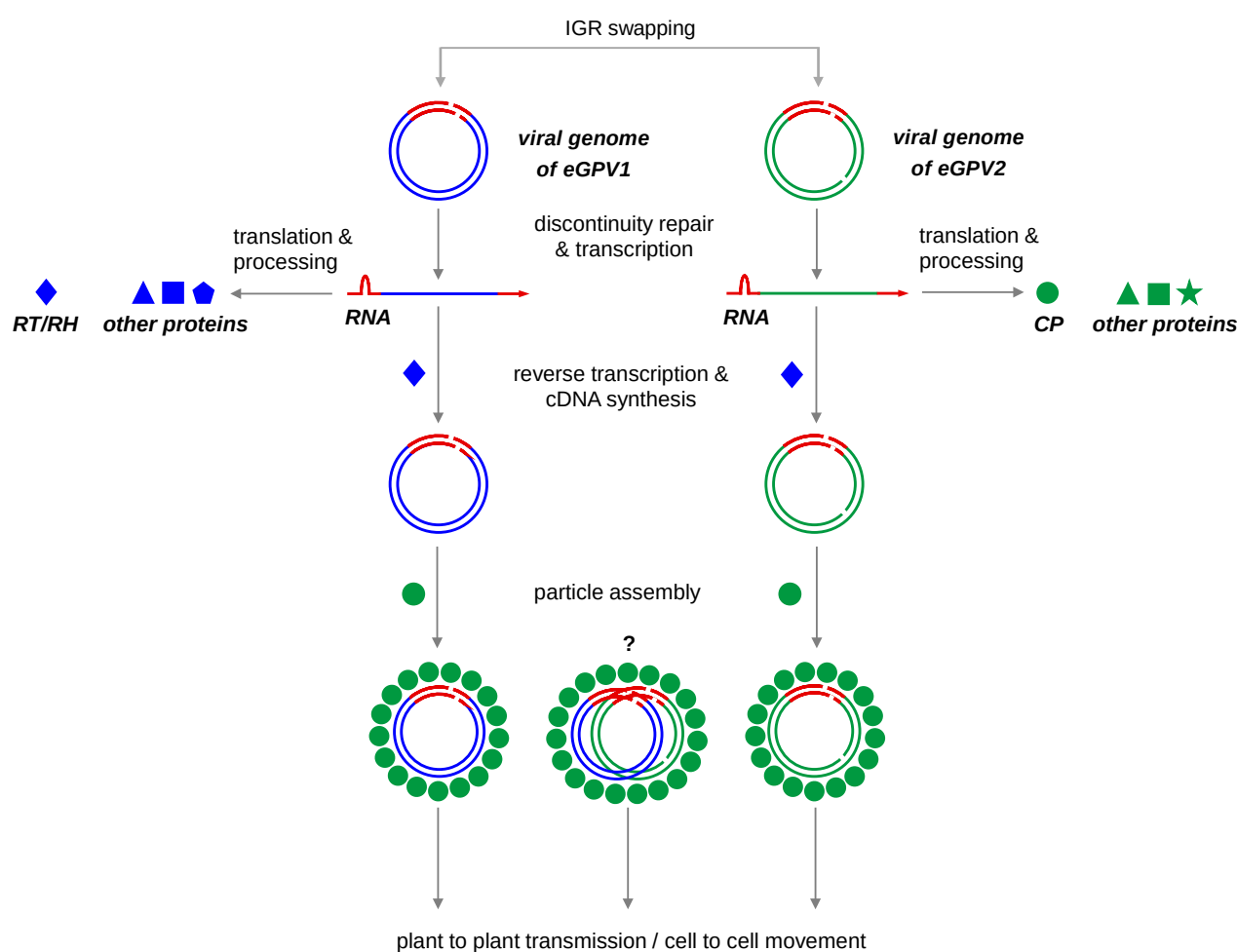


Figure 4.12 A proposed mutualism between two ancient nonautonomous pararetroviruses in the life circles in host cells

Red lines indicate the highly similar IGR sequences between the viral genomic and corresponding RNA sequences of eGPV1s and eGPV2s. Blue lines and shapes show the viral DNA, RNA and protein of eGPV1s, and green lines and shapes show those of eGPV2s.

Chapter 5 Recombinant pararetrovirus-derived eRPV sequences support an interspecies recombination-mediated viral modular evolution

5.1 Introduction

The theory of modular evolution was first proposed by David Botstein at 1980 initially for bacteriophage and now was considered to apply to all viruses (Botstein, 1980; Hull, 2009). This theory suggests that macroevolution of viruses is considerably progressed by reassortments of interchangeable modules, each of which performs a particular biological function. Although numerous examples of module reassortments in the genomic structure variation among different virus genus/family, our understanding of molecular mechanisms underlying viral modular evolution remains limited (Hull, 2009).

EVEs are genomic fossil records of past viral invasions and provide invaluable resources for the study of viral genomic macroevolution (Feschotte and Gilbert, 2012). EPRVs are a kind of EVEs derived from the incidental integration of pararetroviruses in the host germline genomes (Staginnus and Richert-Pöggeler, 2006). I and Kunii *et al* (2004) have identified three kinds of EPRVs from distinct pararetrovirus species in the genomes of rice and other grass species. The eRTBVL families were the first identified grass EPRVs which were derived from *Tungrovirus*-related ancient pararetroviruses (Kunii *et al*, 2004). The eGPV1 and eGPV2 families were subsequently identified grass EPRVs, both of which were derived from *Petuvirus*-related ancient pararetroviruses (see Chapter 4). These ancient grass pararetrovirus species displayed different genomic structures and each had unique functional units (modules). The in-depth study of these EPRVs can provide novel insights into the macroevolution of pararetroviruses. Here I identified a family of chimeric EPRVs (eRPVs) in rice genomes, the corresponding pararetrovirus of which was generated from the illegitimate recombination between the viruses of eRTBVL and eGPV2

sequences. The virus of eRPVs had developed into a pararetrovirus species during the divergence of host *O. sativa* subspecies and had a reassorted genomic structure, which provided genomic fossil evidence to support that interspecies illegitimate recombination is one of the mechanisms driving viral modular evolution.

5.2 Materials and methods

5.2.1 Genomic data mining

The genomic data informations of grass species were listed in Supplementary Table S2. The consensus sequence of eRPVs was constructed by the sequences collected using a BLASTN search in the BLAST+ 2.2.27 application (Camacho *et al*, 2009). The BLASTN search was performed against the grass genomes using the recombinant sequence from the eRTBVL-X cluster in the chromosome 8 of *O. sativa japonica* (see Chapter 2) as a query and the following parameters were used: word size, 11; gap open, 5; gap extend, 2; penalty, -3; reward, 2; and e-values < 1e-10. To further search the other possible recombinant sequences, BLASTN searches were performed using the eRTBVL (DDBJ accession numbers: BR001195-BR001199), eGPV1 and eGPV2 (see Chapter 4) consensus sequences as queries against the grass genomes with the same parameters, and the eRTBVL segments adjacent to eGPV1/eGPV2 segments were examined.

5.2.2 Phylogenetic analysis

ML phylogenetic analyses were performed using PhyML version 3.0 (Guindon *et al*, 2010). To analyzing the phylogenetic relationship between eRPV and eGPV2 sequences, the nucleotide sequences of the homologous region between eRPVs and eGPV2s were extracted from the *japonica* genome. The nucleotide sequences were aligned by ClustalW (Thompson *et al*, 1994) plus manual rechecking. The jModelTest version 2.1.4 (Darriba *et al*, 2012) was used to judge the best fitting substitution model based on AIC. The HKY+G model was chosen as the best model used in PhyML. The

branch supports of ML tree were calculated by 1,000 bootstrap replicates. The ML tree was visualized with FigTree version 1.4.1 (<http://tree.bio.ed.ac.uk/software/figtree>). Because the eRPV sequences described here were recognized as eRTBVL loci formerly (Liu *et al*, 2012), the phylogenetic analysis of the homologous regions between eRPVs and eRTBVLs has been performed in Chapter 2 (Supplementary Figures S1-S6).

5.2.3 Orthologous insertion analysis

The dataset of genome collinearity (genome-wide alignment) between *japonica* and *indica* was obtained from Gramene database (Monaco *et al*, 2014). The left and right 5 kb flanking sequences of each locus of eRPVs in the *japonica* genome were mapped into the *indica* genome using BLASTn algorithm in BLAST+ 2.2.27 utility, and subsequently the flanking sequences of *indica*-eRPV loci were mapped into the *japonica* genome in the same manner. Those eRPV loci each mapped to a unique genomic locus were retrieved, and ambiguous mapped results were rechecked using the genome collinearity data.

5.3 Results

5.3.1 The eRPV sequences derived from a recombinant virus recently

In the certain grass genomes, a small amount of eGPV1 segments neighbored eGPV2 segments (Supplementary Table S7), but no consensus pattern was observed among the link regions of eGPV1 and eGPV2 segments at such kinds of loci, which suggested that these loci could result from insertion/deletion or rearrangement events after viral integrations into grass genomes. Strikingly, some specific eGPV2-similar sequences (1.72 kb in length) were embedded within the eRTBVL sequences in the *japonica* genome (Figure 5.1A). The examination of the link regions between eGPV2-similar and eRTBVL sequences from these loci showed a consensus pattern (Figure 5.1A), which suggested that these chimeric sequences were derived from a recombinant virus, parents of which could be the viruses of eGPV2s and eRTBVLs, or derived from a

common mother virus of the viruses of eGPV2s and eRTBVLs. I named these chimeric ERPVs as endogenous recombinant pararetrovirus (eRPV). The eRPV sequences described in this study were improperly recognized as eRTBVL loci formerly (Liu *et al*, 2012) (details see Figure 5.1 A). The sequence alignment revealed possible recombination events (Figure 5.1B) and two recombination breakpoints that correspond to the borders of the link regions were shown in Figure 5.1A.

Except for *O. sativa*, I did not found eRPV or related chimeric sequences in any other *Oryza* genomes investigated in Chapter 4. In the phylogeny of the sequences of homologous part between eRPVs and eGPV2s (rooted by the known oldest eGPV2 sequence), the sequences from eRPVs were phylogenetically younger than the young *O. sativa*-specific eGPV2 sequences (Figure 5.2), which suggested that the virus of eGPV2s was a parent of the virus of eRPVs. *O. sativa* comprises two subspecies, *japonica* and *indica*, I thus examined the homologous sequences of eGPV2s and eRPVs in the *indica* genome. The orthologous locus analysis of these sequences revealed that most eGPV2 sequences in the *japonica* genome are also present in the *indica* genome and those eRPV sequences in the *japonica* genome in the tree are absent in the *indica* genome (Figure 5.2). Consequently, these suggested a hybrid origin of the virus of eRPVs from the recombination between the viruses of eGPV2s and eRTBVLs and their endogenization events after the divergence of the host rice subspecies. The emergence of the virus of eRPVs could also occur during or after the divergence of *japonica* and *indica*.

5.3.2 Interspecies illegitimate recombination led modular evolution and subsequently symbiosis switch

Examination of the recombination breakpoints showed no obvious sequence homology between the eGPV2 and eRTBVL consensus sequences, just a small microhomologous region at the left breakpoint (Figure 5.1 A), suggesting an

illegitimate recombination event between the viral genomes of eGPV2s and eRTBVLs to produce the viral genome of eRPVs. As shown in Figure 5.1B, this recombination between distinct pararetroviruses induced a genomic domain reassortment which accords with modular evolution. Multiple copies of eRPVs suggested that the corresponding virus of eRPVs should be able to replicate as the viruses of eGPV1s and eGPV2s, although it possessed a hybrid PR domain and lost an RT/RH domain after recombination (Figure 5.1B). Additionally, the eRPV sequences can be grouped into different subclades in the phylogenetic tree (Figure 5.2), suggesting a persistent evolution and divergence of the virus species of eRPVs after their generation.

Noteworthily, the viral genome of eRPVs was mainly derived from the viral genome of eRTBVLs and kept an IGR from the virus of eRTBVLs (Figure 5.1 A). I thus proposed the virus of eRPVs as another nonautonomous pararetrovirus which could hijack the molecular machinery of the virus of eRTBVLs through common *cis*-regulation elements in the IGR to complete its life circle in host cells. The viruses of eRTBVLs were originated from four major lineages (viruses of eRTBVL-A, -B, -C and -D families) and three of the lineages (viruses of eRTBVLs-A including -A1 and -A2 subfamilies, -B and -C families) emerged relatively recently (Figure 2.11). The detailed phylogenetic analyses of the sequences of homologous part between eRPVs and eRTBVLs I performed (Figure 2.1, and Supplementary Figures S1-S6) revealed that the sequences of eRPVs generally cluster with eRTBVL-A2 sequences together, and suggested that the lineage related to eRTBVL-A2 contributed to this recombinant virus. Accordingly, the virus related to eRTBVL-A2 could be the autonomous partner of the nonautonomous virus of eRPVs initially.

More than half segments of eRPVs in the *japonica* genome were clustered together in the chromosome 8, i.e. in the eRTBVL-X cluster (Figure 5.1 A) (Liu *et al*, 2012). Consistent with the phylogenetic positions of these eRPV sequences (Figure 5.2), the

eRTBVL-X cluster carrying the eRPV sequences were only existed in a part of *japonica* accessions by PCR detection at population level (Supplementary Table S4), which indicated that the virus of the eRPV sequences in the eRTBVL-X cluster evolved and integrated during the further divergence of *japonica*. The virus of eRTBVL-X sequences was the evolutionarily youngest lineage of viruses of eRTBVLs (Figure 2.11). The eRPV sequences identified here unveiled that the eRTBVL-X cluster in fact comprises the sequences of eRTBVL-X and eRPV families, and the eRTBVL-X cluster provided evidence of coinfection of the viruses of the eRTBVL-X and phylogenetically youngest eRPV sequences. Furthermore, the IGR sequences of eRTBVL-X are highly phylogenetically close to those of eRPVs (Figure S6). Taken together, the corresponding virus of the eRPV sequences branching later in the phylogenetic tree (Figure 5.2) could parasitize the virus of eRTBVL-X, which suggested a co-evolution between a pair of nonautonomous and autonomous pararetroviral partners during host rice evolution. The genomic fossil record of eRPVs revealed that the modular evolution mediated by interspecies illegitimate recombination conducted a symbiosis switch of nonautonomous viruses.

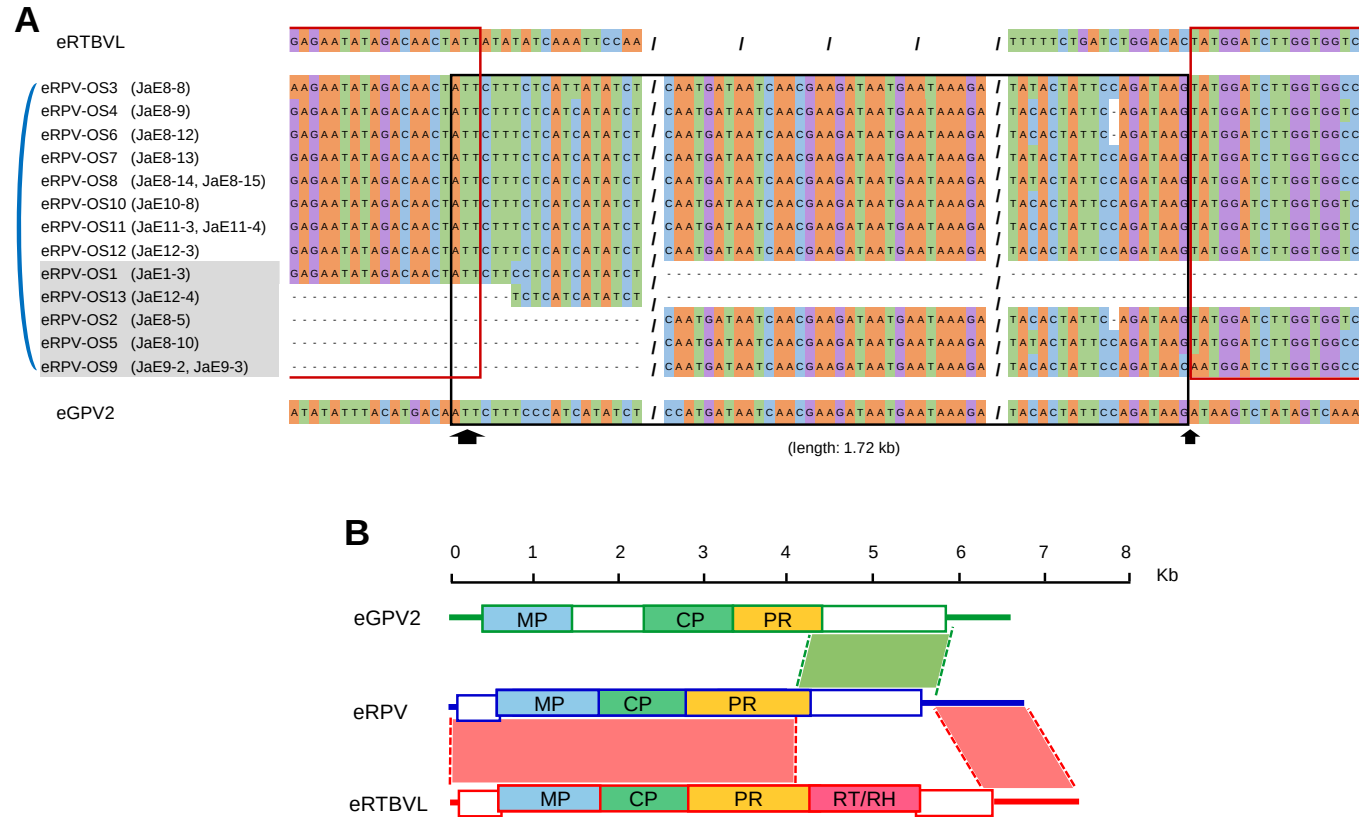


Figure 5.1 The eRPV sequences derived from a recombinant pararetrovirus

(A) A portion of the alignment of eRPV, eGPV2 and eRTBVL sequences. The eRPV segment names are exhibited at the left of panel (old names from Liu *et al* (2012) are enclosed in brackets), and part of eRPV sequences are shown and aligned to the corresponding parts of eGPV2 and eRTBVL sequences (eGPV2-OS and eRTBVL-OS consensus sequences were used). The eRPV segments having a deletion in their sequences are highlighted in grey background. The highly similar parts between eRPV and eGPV2 sequences are marked by a black box, and those between eRPV and eRTBVL sequences are marked by red lines. The estimated recombination breakpoints are indicated by arrows. Slashes represent sequence omissions, and hyphens show aligning gaps. **(B)** Presumed recombination event between the viruses of eGPV2s and eRTBVLs to produce the virus of eRPVs. The viral structures of eGPV2s is depicted in green solid lines (IGR part) and boxes (ORF part), and those of eRPVs and eRTBVLs are illustrated in red and blue ones, respectively. The homologous region between viruses of eGPV2s and eRPVs is indicated by green dash lines, and the homologous regions between viruses of eRTBVLs and eRPVs by red dash lines. All the element structures are drawn according to the length scale.

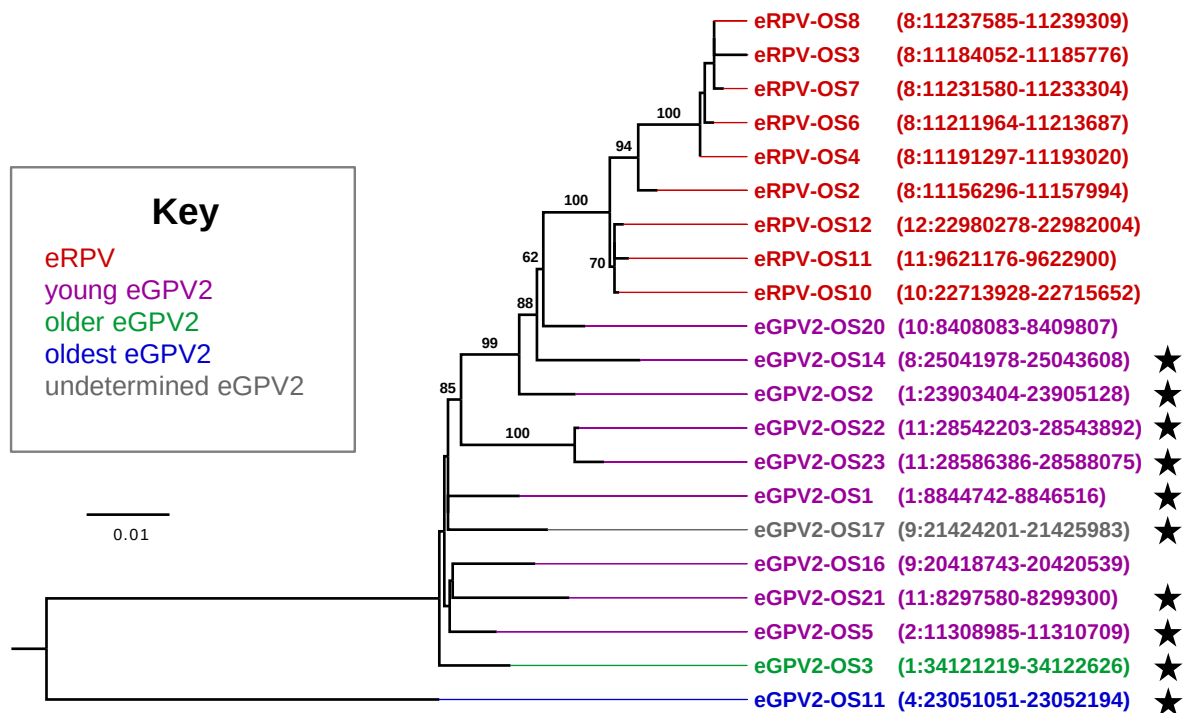


Figure 5.2 Phylogenetic relationship of the homologous parts between eRPV and eGPV2 sequences

The homologous sequences between eRPV and eGPV2 segments in the *O. sativa japonica* genome were used to build the ML tree (highly truncated sequences are not included). The tree is rooted using the known most ancient eGPV2 segment (see Chapter 4). The bootstrap supports derived from 1000 replicates are showed in percentage at the branches (only values greater than 60% are shown). The taxa in red, purple, green, blue and grey denote the eRPV, young eGPV2, older eGPV2, oldest eGPV2 and undetermined eGPV2 sequences, respectively (see Chapter 4, and Supplementary Table S8). The genomic locations of these sequences are provided in brackets. Black stars indicate the sequences shared by both subspecies of *O. sativa*. Scale bar represents evolutionary distance in nucleotide substitutions per site.

5.4 Discussion

Multiple paleogenomes of grass pararetroviruses identified in a series of grass genomes have provided the evidence to reveal the evolution of nonautonomous pararetroviruses. Besides a mutualistic relationship between the viruses of eGPV1s and eGPV2s, here I also observed a parasitic relationship between the viruses of eRPVs and eRTBVLs linking that between nonautonomous and autonomous retrotransposons (Jiang *et al*, 2002). These partnerships between pararetroviruses in the mutualistic or parasitic symbiosis should be achieved and maintained based on the sharing of *cis*-regulation elements in the noncoding region (IGR). Compared with nonautonomous transposons, nonautonomous viruses have rarely been identified and delineated. A well-known nonautonomous virus is adeno-associated virus (*Dependoparvovirus*, *Parvoviridae*), which is a small replication-defective ssDNA virus and parasitizes the autonomous viruses of *Parvoviridae* family (Zinn and Vandenberghe, 2014). My data suggested an ancient rice nonautonomous virus species produced by the modular evolution of an autonomous virus (the major viral genome of eRPVs was derived from the virus of eRTBVLs) and subsequently parasitized its mother autonomous virus persistively during evolution.

The case of modular evolution in this study should be driven by the illegitimate recombination between distinct pararetroviruses. The possible interspecies recombination events of viruses have been frequently reported (Smith *et al*, 2000; Magiorkinis *et al*, 2005; Maori *et al*, 2007; Vigne *et al*, 2008; Mekuria *et al*, 2009; Yozwiak *et al*, 2010; Schibler *et al*, 2012). Here I provided a clear molecular fossil record that interspecies recombinations drove the pararetrovirus diversity, and revealed that illegitimate recombination was the mechanism of this viral genomic exchange event. The *Petuvirus*-related virus of eGPV2s and the *Tungrovirus*-related virus of eRTBVLs belong to different genera of pararetroviruses and possess different genomic structures and very weak nucleotide sequence similarity. Although an abundance of

reports of viral interspecies recombination, the credible reports of intergenus recombination are relatively rare. A credible intergenus recombination event revealed here has led viral modular evolution. The illegitimate recombination between the distinct pararetrovirus species coinfections within a plant could have led various chimeric genomes of interspecies recombinants. Certain such hybrid viruses might survive and develop into new species of autonomous viruses which had different structural and functional modules compared to their parents; some defective hybrids losing some indispensable modules were functionally complementary with each other and thus evolved into a mutualism; while certain defective hybrids losing indispensable modules successfully hijacked the autonomous viruses appeared in a same host cell and evolved into a parasitism. Based my observations, I concluded that interspecies illegitimate recombination is one of the mechanisms driving viral modular evolution.

Chapter 6 General discussion

6.1 Outcome of ancient grass pararetrovirus and origin of extant one

Phylogenetic analyses suggested that the viruses corresponding to eRTBVLs could be a sister of the extant pararetrovirus RTBV (Figure 2.10). The absence of ORF2 and different ORF junctions also implied that these ancient viruses could not be the direct ancestors of RTBV (Geering *et al*, 2010). My data further suggested that the divergence between RTBV and the viruses of eRTBVL occurred at least before the immediate progenitor of the *Oryza* BB-genome species (Figures 2.9 and 2.11). Further identification of eRTBVL superfamily in widespread grass genomes suggested that the viruses of eRTBVL superfamily could be a sister genus of *Tungrovirus* (Figure 3.1). The viruses corresponding to eRTBVLs may be extinct now, or may still be circulating in certain grass populations but have remained uncharacterized because no obvious symptoms have been detected. It has been suggested that EPRVs could play a role in viral resistance by homology-dependent gene silencing (Mette *et al*, 2002; Staginnus and Richert-Pöggeler, 2006). If eRTBVLs were employed by their hosts to counter the corresponding exogenous viruses, no or weak symptoms of an exogenous viral infection as a result of this resistance might have led to them being overlooked. Host switch from grass to other plant species may be another evolutionary result of these ancient viruses.

Notably, no integrated RTBV sequence (highly similar) has been found in any of the available grass genome assemblies until now. One interpretation is that RTBVs either lack the property required to make them be prone to insertion or they possess a property that inhibits integration. It has been suggested that ORF2 in the RTBV genomes may play a role in suppressing the incorporation of viral segments into host genomes by decreasing the numbers of unpacked viral genomes that are suspended in the cells of their rice hosts (Liu and Kishima, 2014). Indeed, the protein encoded by ORF2 were

shown previously to participate in capsid assembly to complete virion packing (Herzog *et al*, 2000), while the counterparts of these proteins were reported to be absent in eRTBVLs (Kunii *et al*, 2004). Alternatively, RTBV might recently have undergone a switch of hosts, from asymptomatic natural hosts to a new rice species host (Hull, 2002b) without having left any integrated copies in any rice populations as yet. In the future, integrated counterparts of RTBV may be identified in other distinct plant genomes at which time this likelihood can be investigated further.

Interestingly, besides *Tungrovirus*-related EPRVs (eRTBVLs), I also identified two *Petuvirus*-related EPRVs (eGPV1s and eGPV2s) in grass species. The viruses of eRTBVLs have been endogenized in all the examined grass genomes, but those of eGPV1s and eGPV2s have only been endogenized in some of the grass genomes (Table 3.1 and Figure 4.5), which suggested that the host ranges of the viruses of eGPV1s and eGPV2s could be narrower than those of the viruses of eRTBVLs. Alternatively, an interaction between the viruses of eGPV1s/eGPV2s and specific host grass species led to the integrations of the virus sequences of eGPV1s/eGPV2s into those grass genomes. The viruses of eGPV1s/eGPV2s could be a sister of current virus PVCV, and strikingly, the latter can be induced from the corresponding EPRVs (Richert-Poggeler *et al*, 2003). I did not find the obvious evidence that the viruses of eGPV1s and eGPV2s have recombined into an autonomous monopartite virus, but found the evidence to support that the virus of eGPV2s have recombined with the virus of eRTBVLs to generate another nonautonomous virus, the virus of eRPVs. Liking the case of the virus of eRTBVLs, the nonautonomous viruses corresponding to eGPV1s, eGPV2s and eRPVs may be extinct now, or live but uncharacterized. The identification of these viruses of grass EPRVs in this study suggested that the diversity of pararetroviruses is greater than formerly thought, and that our knowledge of viruses remains limited.

6.2 Impact of ancient pararetroviruses on host immunity

EPRVs were generally non-infectious due to mutational decay and epigenetic regulation (Staginnus and Richert-Pöggeler, 2006). For grass EPRVs, I observed relatively intact EPRV sequences (inactive) and an eRTBVL long cluster, no EPRV reactivation has been reported in grass species and hybrids, although the induction of EPRV-derived infections have been observed in tobacco, petunia and banana hybrids (reviewed by Chabannes and Iskra-Caruana, 2013). Abundance of EPRVs have been recurrently fixed in grass and other various plant species (Staginnus and Richert-Pöggeler, 2006; and this study), which make us expect their potential roles in the interplay between viruses and hosts. EPRVs were assumed to play an antiviral function against exogenous pararetroviruses by homology-dependent gene silencing (Mette *et al*, 2002). In plants, the transcription of transposons and other repetitive elements can be controlled by RNA-directed DNA methylation pathway, where small interfering RNAs (siRNA) guide DNA methyltransferases to the target *in cis* or *in trans* (Lisch, 2009; Zhang and Zhu, 2011). The EPRVs could be also controlled at transcription level guided by EPRV-derived siRNAs *in cis*, which are able to target on similar exogenous pararetroviruses *in trans* to repress their transcription by promoter methylation (Staginnus and Richert-Pöggeler, 2006). The observation of high methylation and low transcription level of EPRVs offered a foundation to the hypothesis (Mette *et al*, 2002; Kunii *et al*, 2004; Noreen *et al*, 2007; Staginnus *et al*, 2007). EPRV-derived siRNAs have been also reported in some plant species (Staginnus and Richert-Pöggeler, 2006; Noreen *et al*, 2007; Staginnus *et al*, 2007; Becher *et al*, 2014). It was reported that the transgene under EPRV-derived promoter was methylated and silenced in tobacco, which possesses corresponding EPRVs, but not in *Arabidopsis*, which lacks these elements (Mette *et al*, 2002). In addition, the EPRV-derived siRNAs may also work in the RNAi/ post-transcriptional gene silencing against exogenous viruses (Staginnus and Richert-Pöggeler, 2006; Staginnus *et al*, 2007), and histone methylation pathway regulating EPRV and hence exogenous ones (Noreen *et al*, 2007). However, more direct

evidence about EPRV-derived viral resistance was absent until now. eRTBVL loci were also methylated in rice genomes (Kunii *et al*, 2004), and abundance rice siRNA reads could be mapped into eRTBVL sequences by searching in the rice small RNA database (data not shown, Nakano *et al*, 2006). Hence it is possible that eRTBVLs and the other grass EPRVs were employed by hosts to counter the corresponding exogenous viruses. As I discussed above, the viruses of grass EPRVs could be extinct or just unisolated. No or weak symptoms of exogenous viral infection led by EPRV-derived resistance might make people ignore these viruses.

Besides EPRV-derived viral resistance, ancient pararetroviruses could have also shaped the immunity system of plant hosts by driving the evolution of antiviral factors during their long-term invasions. Host antiviral proteins, especially those directly contacting with viral proteins need keep rapid evolution to counteract constantly evolving variants of viruses and therefor showed a positive selection signature of certain domains (Emerman and Malik, 2010; Patel *et al*, 2011). Positive selection possibly driven by ancient retroviruses have been observed in animal genomes such as an antiviral protein TRIM5 in primate genomes that specially counter only retroviruses (Sawyer *et al*, 2005; Goldschmidt *et al*, 2008). The identification and analysis of orthologous anti-pararetroviral factors of plants could test this impact of ancient pararetroviruses on plant immunity.

It is an interesting question that whether ancient rice pararetroviruses involved in a resistance to their extant sister RTBV by eRTBVL-derived silencing or shaped antiviral factors. Kunii *et al* (2004) supposed an association between eRTBVL profiles and the degree of RTBV susceptibility. The species with low-copy-number eRTBVLs tend to be vulnerable to RTBV infection, and the methylation level of eRTBVLs was linked with their copy number in host genomes (Kunii *et al*, 2004). Because of an abundance of microhomologous regions between eRTBVL and RTBV sequences, eRTBVLs might

mediate DNA methylation of RTBV genomes or histone methylation of formed viral minichromosomes (Hull, 2002a) via perfect or imperfect complementarity between multiple eRTBVL-derived siRNA and RTBV targets. Alternatively, the copy number of eRTBVLs mirrored the experience of ancient pararetrovirus invasions, and more viral invasion events during long-term co-evolution could shape the corresponding anti-pararetroviral factors which may also counter sister viruses to a certain extent.

6.3 Revisit the mechanism of integration of EPRVs into grass genomes

The integration of the viruses without integration ability into host genomes are achieved through different mechanisms for different virus types. An exogenous nonretroviral RNA virus (lymphocytic choriomeningitis virus, a minus ssRNA virus) was found to recombine with an endogenous retrotransposon (IAP), which led to reverse transcription of the recombinant viral RNA and then integration of the cDNA into the host genome with the activity of IAP elements (Geuking *et al*, 2009). This suggested that the integration of nonretroviral RNA viruses could be completed via retrotransposon/retrovirus-mediated cDNA integration. For animal and plant pararetroviruses (dsDNA virus), host DNA DSBs are their major integration targets and their integration could occur by NHEJ between pararetrovirus and host DNAs (Bill and Summers, 2004; Liu *et al*, 2012). For some other DNA viruses, common fragile sites are their preferential integration targets, such as human papillomavirus (dsDNA virus) and simian virus 40 (dsDNA virus) (Mishmar *et al*, 1998; Thorland *et al*, 2002). It seems that chromosome fragile sites are the preferential targets for dsDNA virus integration.

Chromosome fragile sites have a high recombination frequency and are associated with chromosome rearrangements (Matsuyama *et al*, 2003; Durkin and Glover, 2007; dela Paz *et al*, 2012). AT-rich sequences exist in chromosome fragile sites (Yu *et al*, 1997; Jackson *et al*, 2003; Zhang and Freudenreich, 2007; Ozeri-Galai *et al*, 2011; dela

Paz *et al*, 2012). These AT repeats form strong secondary structures that induce fork stalling and additional origin inactivation during DNA replication, especially when suffering a replication stress, and then cause a chromosome break (Zhang and Freudenreich, 2007; Ozeri-Galai *et al*, 2011). The breaks of chromosome fragile sites can be repaired by homologous recombination and NHEJ pathways (Schwartz *et al*, 2005). Grass EPRVs (eRTBVL, eGPV1 and eGPV2 superfamilies) are preferentially integrated in AT-rich sequences of rice and other grass genomes (This study; Kunii *et al*, 2004; Liu *et al*, 2012). A comprehensive comparative genomic study has suggested that these AT-rich region in the rice genome could be a hotspot of DSB and also of eRTBVL integrations (Liu *et al*, 2012). These AT-rich regions for pararetrovirus integration should be chromosome fragile sites in grass genomes. Accordingly, pararetroviral invasion and reproduction could induce a replication stress in the interphase duration for a germline cell and aggravate the AT-rich sequence breaks, and subsequently pararetroviral DNAs could be used for break repairs by a NHEJ pathway, which could lead to the endogenization of pararetrovirus sequences. The AT-rich sequences in chromosome fragile sites are also recognized as a scaffold/matrix attachment region (Jackson *et al*, 2003; dela Paz *et al*, 2012), which are involved in chromatin organization and gene expression regulation (Rudd *et al*, 2004). The AT repeats targeted by eRTBVL integration in the rice genome are also possibly related to scaffold/matrix attachment regions (Liu *et al*, 2012). The role of nuclear scaffold/matrix in pararetrovirus integration remains unclear.

Experimental evidence from animal pararetrovirus (hepatitis B virus) integration suggested that viral linear dsDNAs from aberrant replication are prone to integrate (Bill and Summers, 2004). In this study, I observed that some of grass EPRV sequences can be mapped to the viral circle genome more than one turn, and that some 3' terminal sequences of EPRVs do not correspond to the 3' end of viral linear genome but extend

to the corresponding 5' end. Consequently, aberrantly replicated linear dsDNAs of pararetroviruses could be preferential substrates for NHEJ-mediated integration.

Chapter 7 Summary

Pararetroviruses (hepadnaviruses and caulimoviruses) are one of the major health threats worldwide to humans, animals and plants. EPRV sequences in plant genomes, such as eRTBVLs in rice genomes, are viral genomic fossil records and can help us to understand ancient virus activities and host histories. In this study, I performed detailed analyses to reveal the organization, endogenization, distribution and evolution of eRTBVLs, as well as of newly identified eGPV1s, eGPV2s and eRPVs. My results provide new insights into pararetrovirus evolution and host grass divergence.

Firstly, I collected about 300 eRTBVL segments from the genomic databases of Nipponbare (*Oryza sativa japonica*), 93-11 (*O. sativa indica*), and W1943 (*O. rufipogon*, eastern China wild accession). Virus sequences were reconstructed from eRTBVL segments in the three rice genomes and grouped into four distinct major families. Three of the families, eRTBVL-A, -B and -C, began endogenization during differentiation of *O. rufipogon*. Different proportions of these eRTBVL families were detected in the three genomes, suggesting that they had independent geographical origins. Screening of orthologous eRTBVL loci in 65 cultivated and 30 wild accessions demonstrated that the flow of eRTBVLs reflected the processes proposed previously for rice domestication. The genealogy of the putative viruses corresponding to the eRTBVL families was inferred from the results. The virus lineage of the eRTBVL-B family could be a parent for viruses of the eRTBVL-A and -C families. Further divergence might have resulted in the virus lineages corresponding to the eRTBVL-A1, -A2, and -X subfamilies; however, no more subfamilies were detected in the three rice genomes. The step-by-step virus lineages can be explained, for the most part, by recombination between the viruses corresponding to the eRTBVL families. The rice genomes have recorded only a few recombination events that nevertheless have contributed to major evolutionary lineages of the ancient pararetroviruses during rice

speciation.

Secondly, orthologous eRTBVL screening in the cultivated and wild rice collection suggested that the oldest family, eRTBVL-D, had been endogenized before differentiation of the rice progenitor, *O. rufipogon*. Further orthologous screening in the genomic databases and DNA samples of other *Oryza* species revealed that the endogenization of eRTBVL-D sequences initially occurred about 5 MYA before the divergence of *Oryza* AA-genome and BB-genome species. Together with phylogenetic analysis, these results indicated that the divergence between RTBV and the viruses of eRTBVLs could occur much earlier than the differentiation of the current natural host of RTBV. Besides the vertical inheritance of eRTBVL-D family in a part of *Oryza* species, I also identified multiple divergent eRTBVL-similar sequences (eRTBVL superfamily) in all the other grass genomes I examined. The corresponding viruses of eRTBVL superfamily could be a sister genus of *Tungrovirus* and possessed distinct types of genomic structures. Further sequence analysis implied independent integration events of most sequences of eRTBVL superfamily in different grass genomes. My results uncovered the extremely widespread “horizontal genome transfers” of a previously uncharacterized genus of pararetroviruses into grass genomes. These independent viral integrations could have been shaping the genomic evolution of host grass species.

Thirdly, although all the known pararetroviruses are autonomous for the life circle in host cells, I identified unexpected multiple EPRVs derived from nonautonomous pararetrovirus species in grass genomes. The reconstructed viral genomes and endogenization profiles of these EPRVs, eGPV1s and eGPV2s, revealed that two distinct nonautonomous pararetroviruses losing part essential functions had horizontally transmitted to a wide range of grass hosts and developed into various groups and lineages during the long-term grass evolution. Comparative molecular

biology and comparative genomics revealed the coding capacity complementarity and noncoding *cis*-regulation sequence identity between the viruses of eGPV1s and eGPV2s. Strikingly, the paleogenomic evidence support that the two virus species kept a co-evolution to develop a mutualistic symbiosis through noncoding region swapping. I further demonstrated that region-specific recombination had driven the homogenization of the noncoding regions required for the mutualism between nonautonomous pararetroviruses. The previously unknown genomic and ecological diversities of pararetroviruses revealed here supply fossil record evidence to support that not Darwinian model but symbiosis model contributes the viral rapid evolutionary events. My study provides paleobiological insights into the evolutionary origin of complex viruses consisting of multiple sequences.

Finally, I identified a family of EPRVs, eRPVs, derived from a recombinant pararetrovirus in rice genomes. Comprehensive sequence *in silico* analyses and orthologous locus screening revealed that the illegitimate recombination between two distinct pararetrovirus species, the viruses of eRTBVLs and eGPV2s, generated the viruses of eRPVs possibly after the divergence of host rice subspecies. The viral genomic structures of eRPVs reassorted from recombination events are consistent with the modular evolution theory. The essential RT/RH domain for viral replication was replaced by an inessential domain in the modular reassortment, which led the virus of eRPVs to be a nonautonomous pararetrovirus. Sequence analysis supported that this nonautonomous virus was able to replicate and even have undergone a persistent evolution during host divergence. The evidence of IGR sequence homogenization and co-infection suggested the parasitism of the virus of eRPVs on the virus of eRTBVLs, which proposed a symbiosis switch between pararetroviruses induced by interspecies recombination-mediated modular reassortment. My results provide special genomic fossil evidence supporting that interspecies illegitimate recombination is one of the mechanisms underlying viral modular evolution.

Supplementary Figures

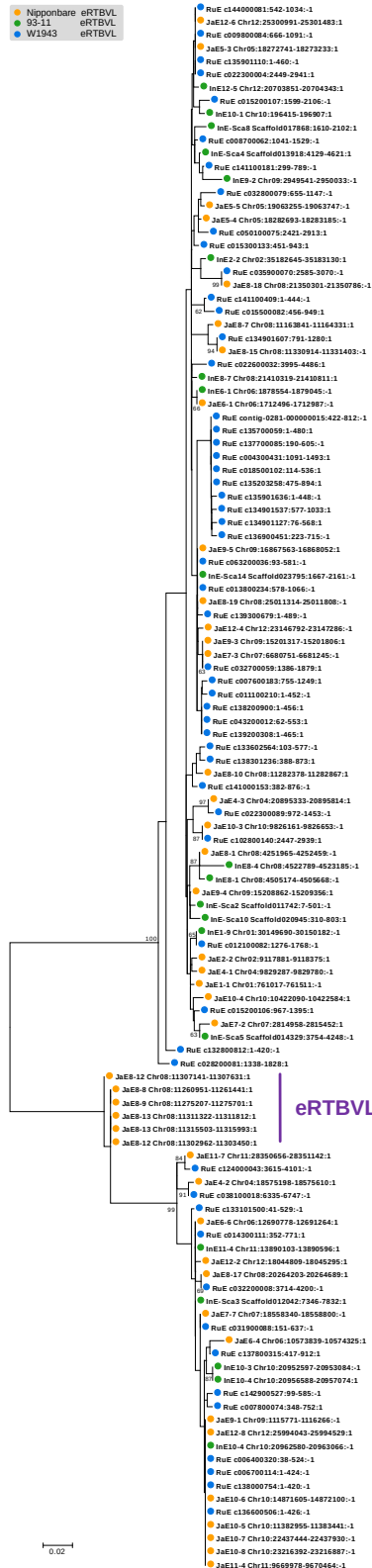


Figure S1 ML phylogenetic analysis of eRTBVL elements based on ORF_x

Nucleotide sequences of ORF_x (>80% of the length of consensus sequence) from three genomes were used. Genomic locations of the sequences are shown using their taxon names (ID, chromosome/contig, position, strand), and the rice genome to which they belong is indicated by the colored solid circle before the taxon name (orange for Nipponbare, green for 93-11, blue for W1943). Each eRTBVL family is indicated (the delimitation between eRTBVL-B and -C is ambiguous). The bootstrap consensus tree inferred from 1000 replicates, and bootstrap values greater than 60 are shown above the branches. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The eRTBVL structure is displayed above the phylogenetic tree and ORF_x is indicated.

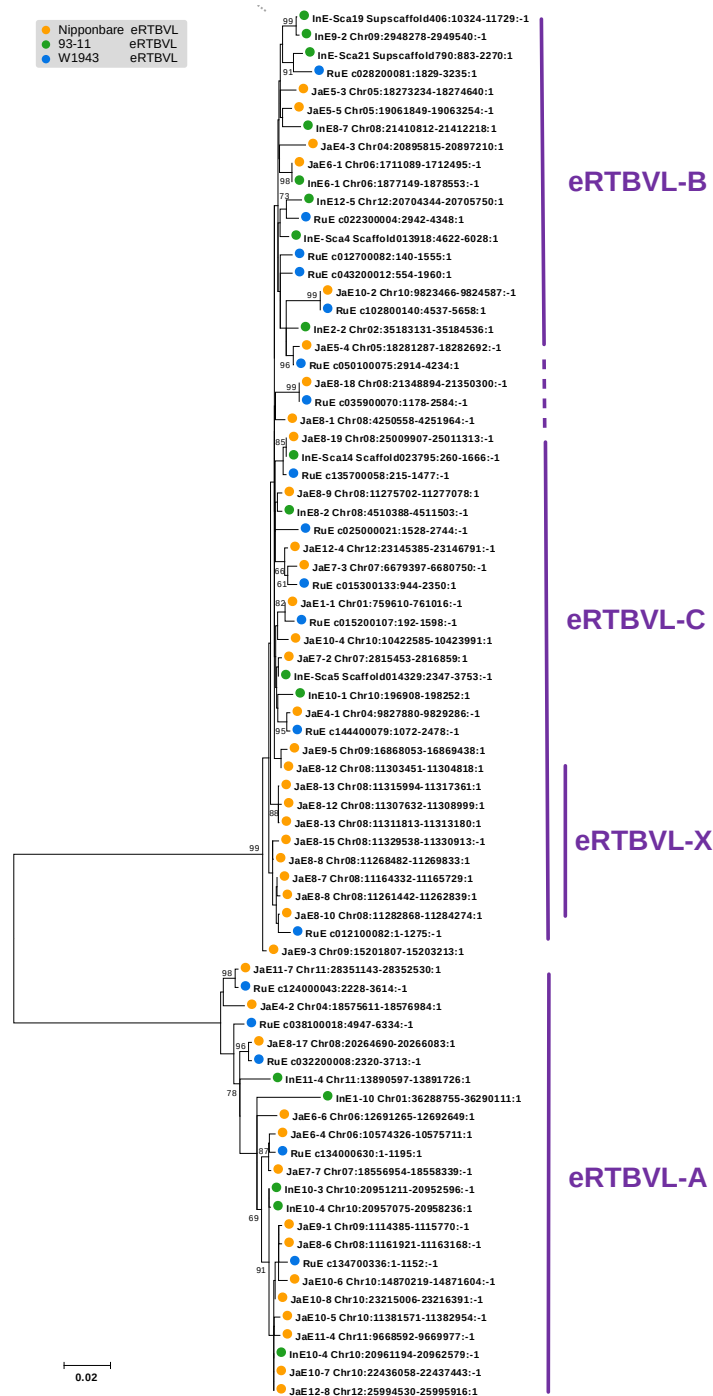


Figure S2 ML phylogenetic analysis of eRTBVL elements based on the MP region

Nucleotide sequences of MP (>80% of the length of consensus sequence) from three genomes were used. Genomic locations of the sequences are shown using their taxon names (ID, chromosome/ contig, position, strand), and the rice genome to which they belong is indicated by the colored solid circle before the taxon names (orange for Nipponbare, green for 93-11, blue for W1943). Each eRTBVL family is indicated (the delimitation between eRTBVL-B and -C is ambiguous). The bootstrap consensus tree inferred from 1000 replicates, and bootstrap values greater than 60 are shown above the branches. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The eRTBVL structure is displayed above the phylogenetic tree and MP is indicated.

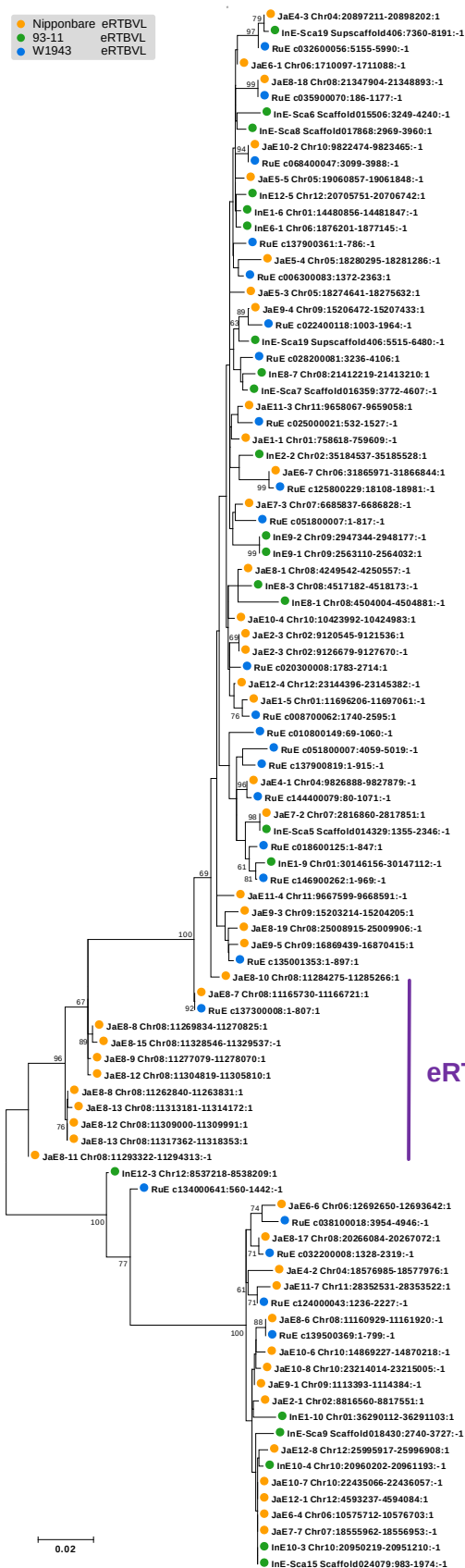


Figure S3 ML phylogenetic analysis of eRTBVL elements based on the CP region

Nucleotide sequences of CP (>80% of the length of consensus sequence) from three genomes were used. Genomic locations of the sequences are shown using their taxon names (ID, chromosome/contig, position, strand), and the rice genome to which they belong is indicated by the colored solid circle before the taxon names (orange for Nipponbare, green for 93-11, blue for W1943). Each eRTBVL family is indicated (the delimitation between eRTBVL-B and -C is ambiguous). The bootstrap consensus tree inferred from 1000 replicates, and bootstrap values greater than 50% are shown above the branches. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The eRTBVL structure is displayed above the phylogenetic tree and CP is indicated.

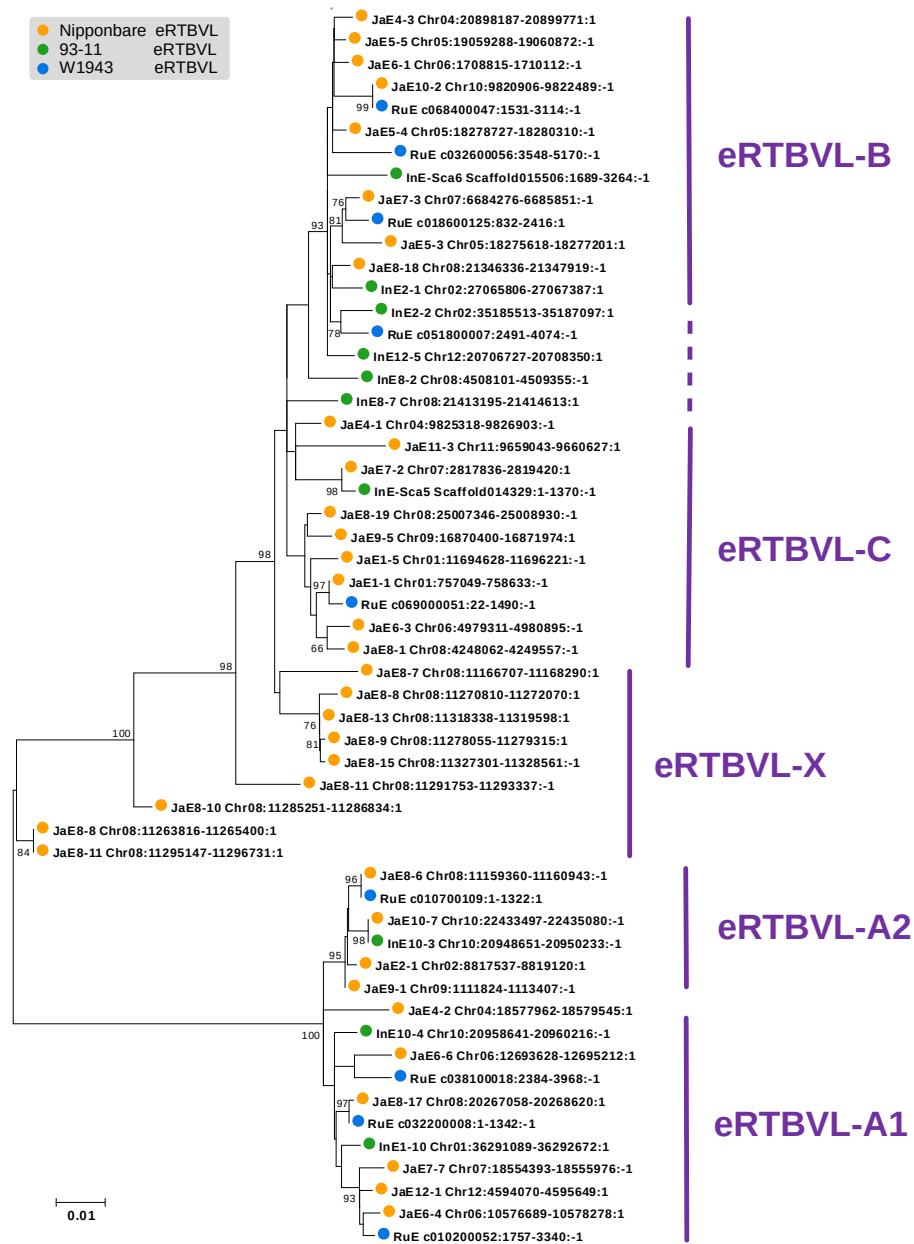


Figure S4 ML phylogenetic analysis of eRTBVL elements based on the PR region

Nucleotide sequences of PR (>80% of the length of consensus sequence) from three genomes were used. Genomic locations of sequences are shown using their taxon names (ID, chromosome/contig, position, strand), and the rice genome to which they belong is indicated by the colored solid circle before the taxon names (orange for Nipponbare, green for 93-11, blue for W1943). Each eRTBVL family is indicated (the delimitation between eRTBVL-B and -C is ambiguous). The bootstrap consensus tree inferred from 1000 replicates, and bootstrap values greater than 60 are shown above the branches. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The eRTBVL structure is displayed above the phylogenetic tree and PR is indicated.

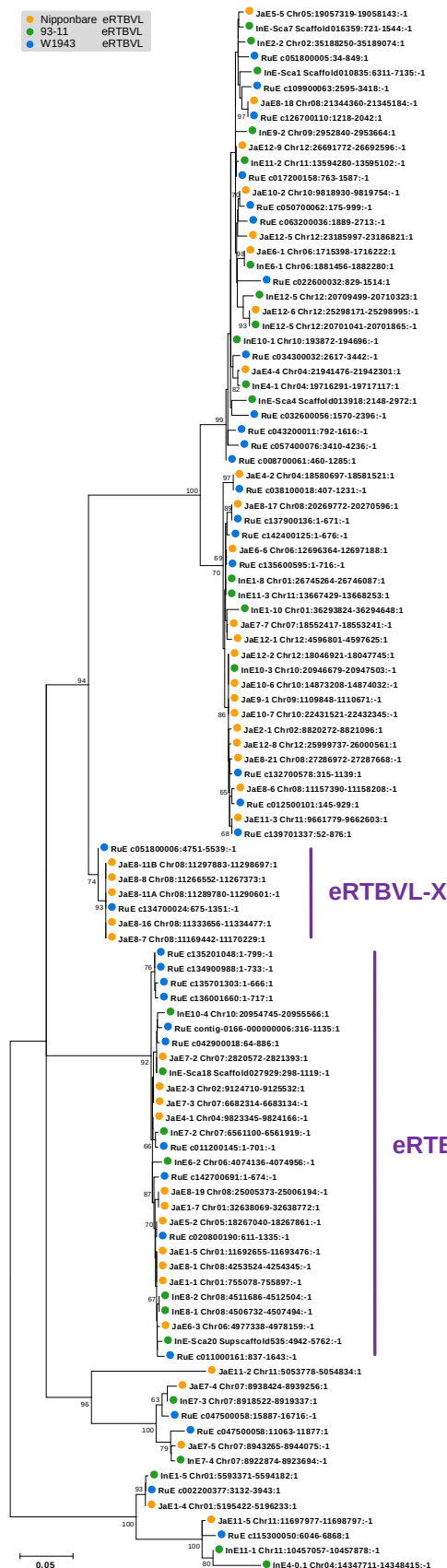


Figure S5 ML phylogenetic analysis of eRTBV elements based on ORFz

Nucleotide sequences of ORFz (>80% of the length of consensus sequence) from three genomes were used. Genomic locations of the sequences are shown using their taxon names (ID, chromosome/contig, position, strand), and the rice genome to which they belong is indicated by the colored solid circle before the taxon names (orange for Nipponbare, green for 93-11, blue for W1943). Each eRTBV family is indicated. The bootstrap consensus tree inferred from 1000 replicates, and bootstrap values greater than 60 are shown above the branches. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The eRTBV structure is displayed above the phylogenetic tree and ORFz is indicated.

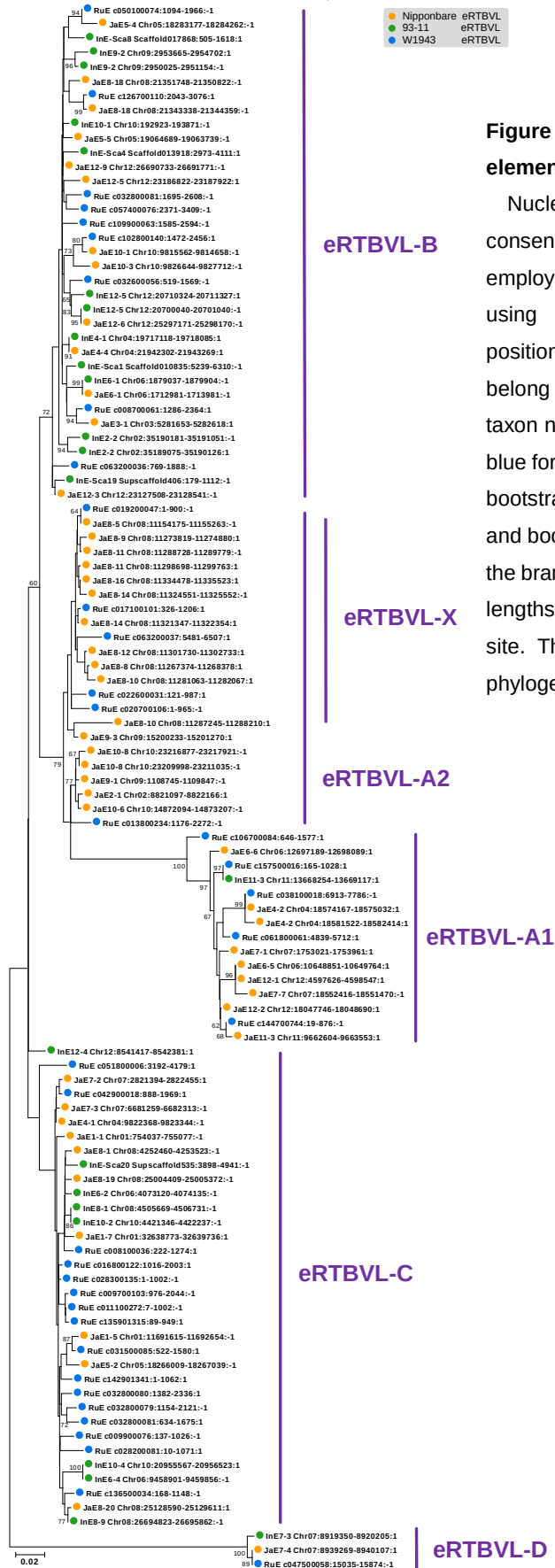


Figure S6 ML phylogenetic analysis of eRTBVL elements based on the IGR

Nucleotide sequences of IGR (>80% of the length of consensus sequence) from three genomes were employed. Genomic locations of sequences are shown using their taxon name (ID, chromosome/contig, position, strand), and the rice genome to which they belong is indicated by the colored solid circle before the taxon names (orange for Nipponbare, green for 93-11, blue for W1943). Each eRTBVL family is indicated. The bootstrap consensus tree inferred from 1000 replicates, and bootstrap values greater than 60 are shown above the branches. The tree was drawn to scale, with branch lengths measured in the number of substitutions per site. The eRTBVL structure is displayed above the phylogenetic tree and IGR is indicated.

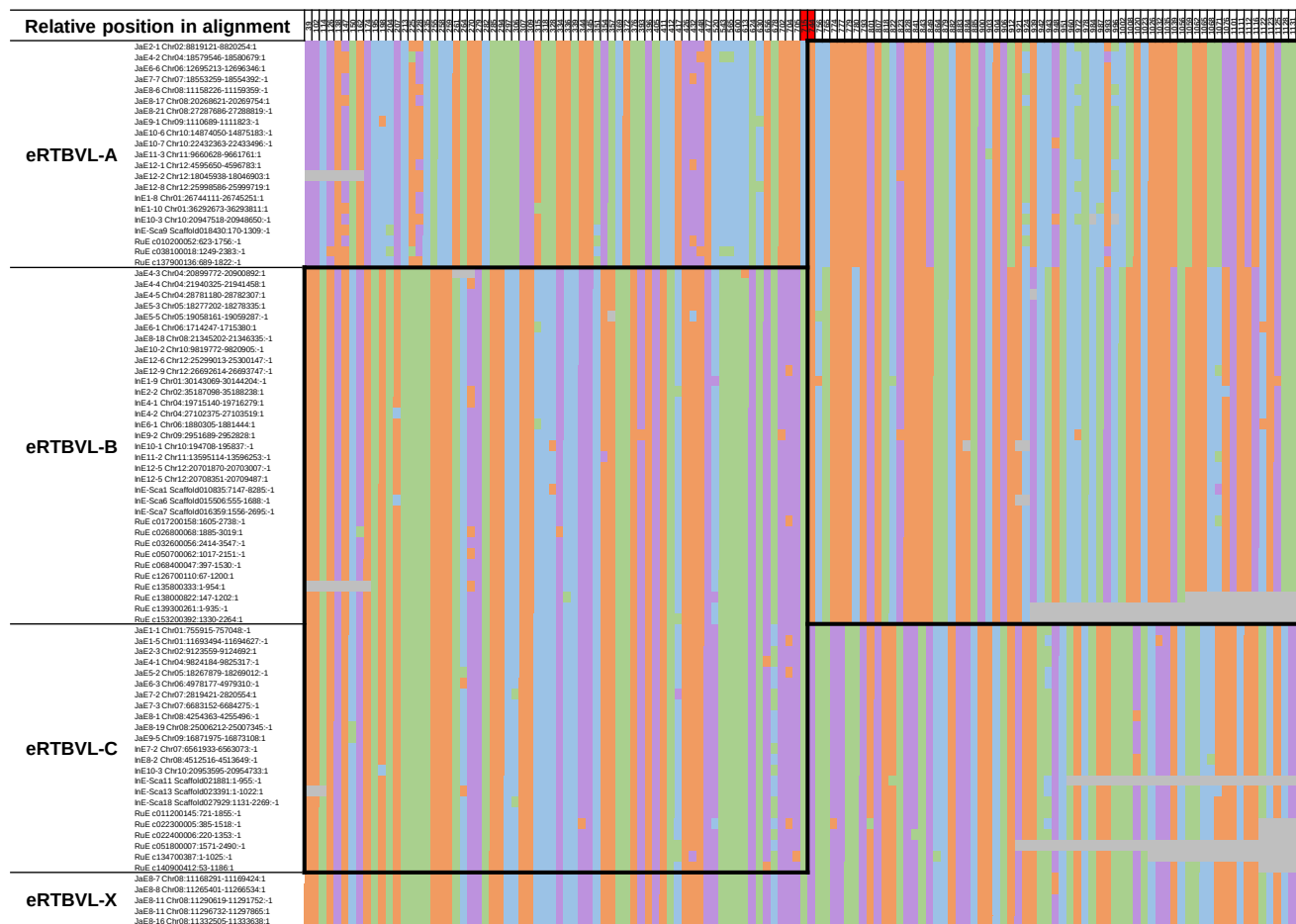


Figure S7 Informative SNP patterns in the RT/RH sequences of eRTBVL families and presumed breakpoint positions in the alignment

Orange, A; green T; blue, C; purple, G. The ranges (relative positions) of the presumed breakpoints in the alignment are indicated in red. Missing data are indicated by hyphens. Black boxes indicate similarity between two eRTBVL families.

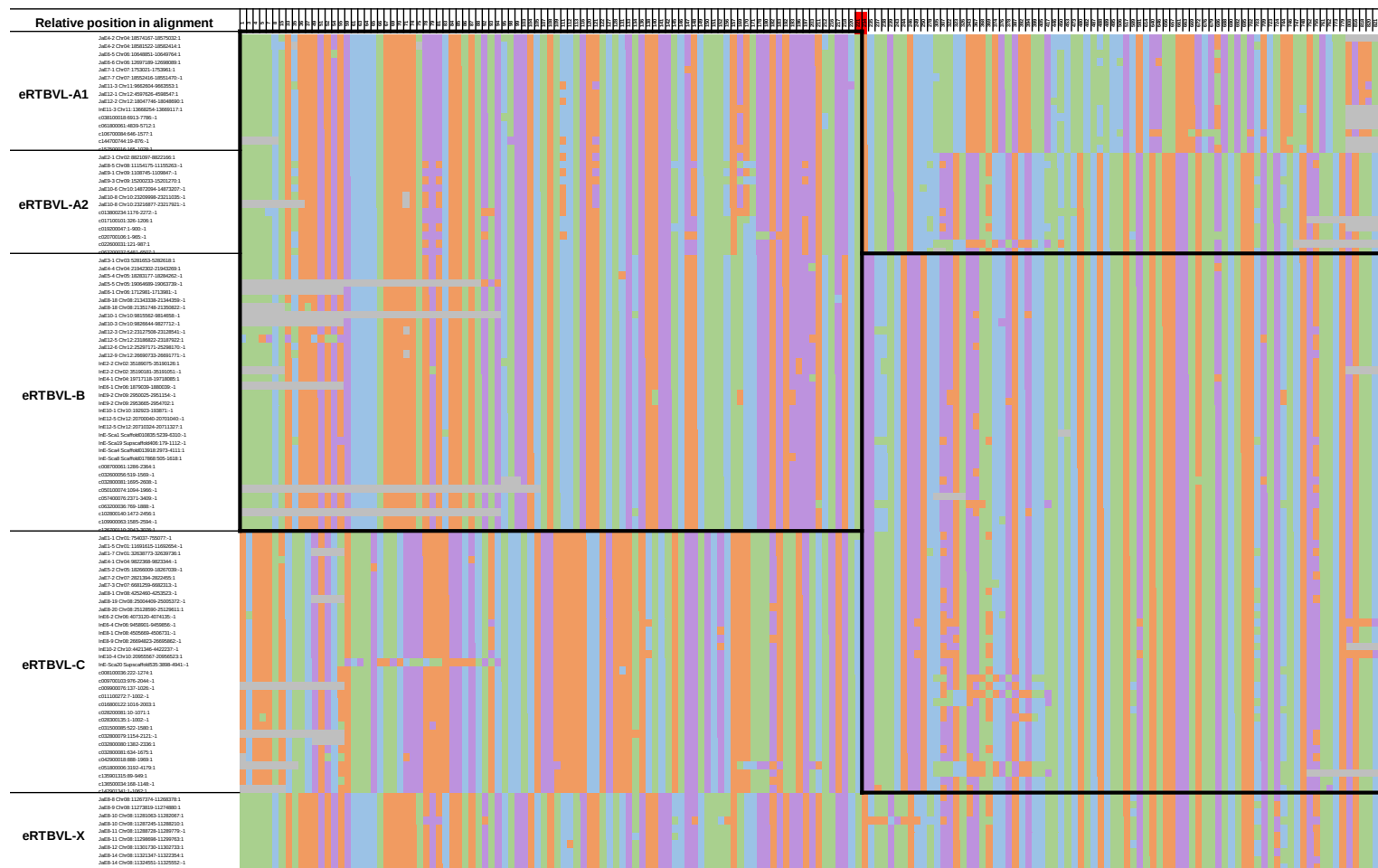


Figure S8 Informative SNP patterns in the IGR sequences of eRTBVL families and presumed breakpoint positions in the alignment (for Figure 2.8A)

Orange, A; green T; blue, C; purple, G. The ranges (relative positions) of the presumed breakpoints in the alignment are indicated in red. Missing data are indicated by hyphens. Black boxes indicate similarity between two eRTBVL families.

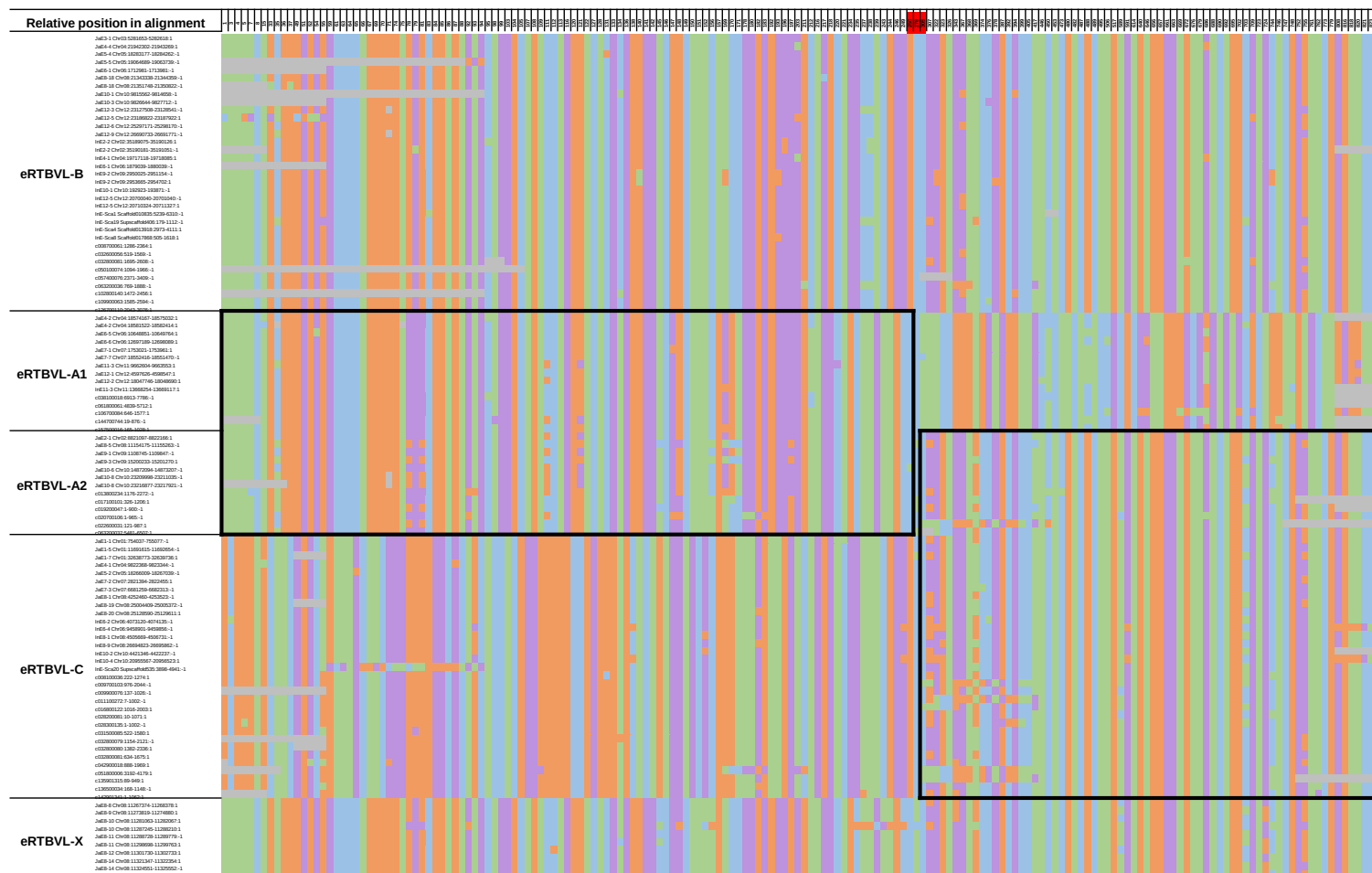


Figure S9 Informative SNP patterns in the IGR sequences of eRTBVL families and presumed breakpoint positions in the alignment (for Figure 2.8B)

Orange, A; green T; blue, C; purple, G. The ranges (relative positions) of the presumed breakpoints in the alignment are indicated in red; missing data are indicated by hyphens. Black boxes indicate similarity between two eRTBVL families.

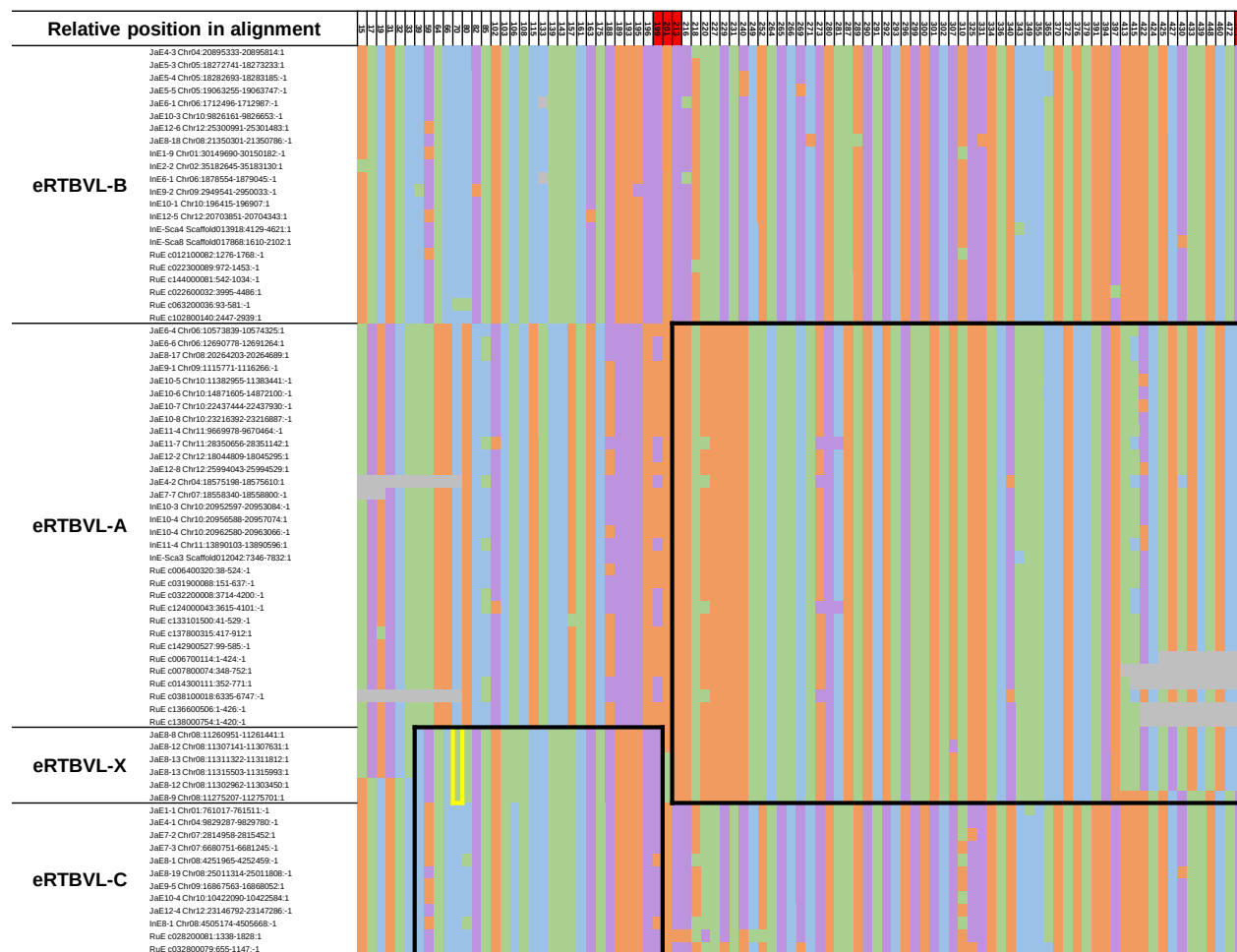


Figure S10 Informative SNP patterns in the ORFx sequences of eRTBVL families and presumed breakpoint positions in the alignment

Orange, A; green T; blue, C; purple, G. The ranges (relative positions) of the presumed breakpoints in the alignment are indicated in red. Missing data are indicated by hyphens. Black boxes indicate similarity between two eRTBVL families. Yellow boxes indicate eRTBVL-X-specific SNPs.

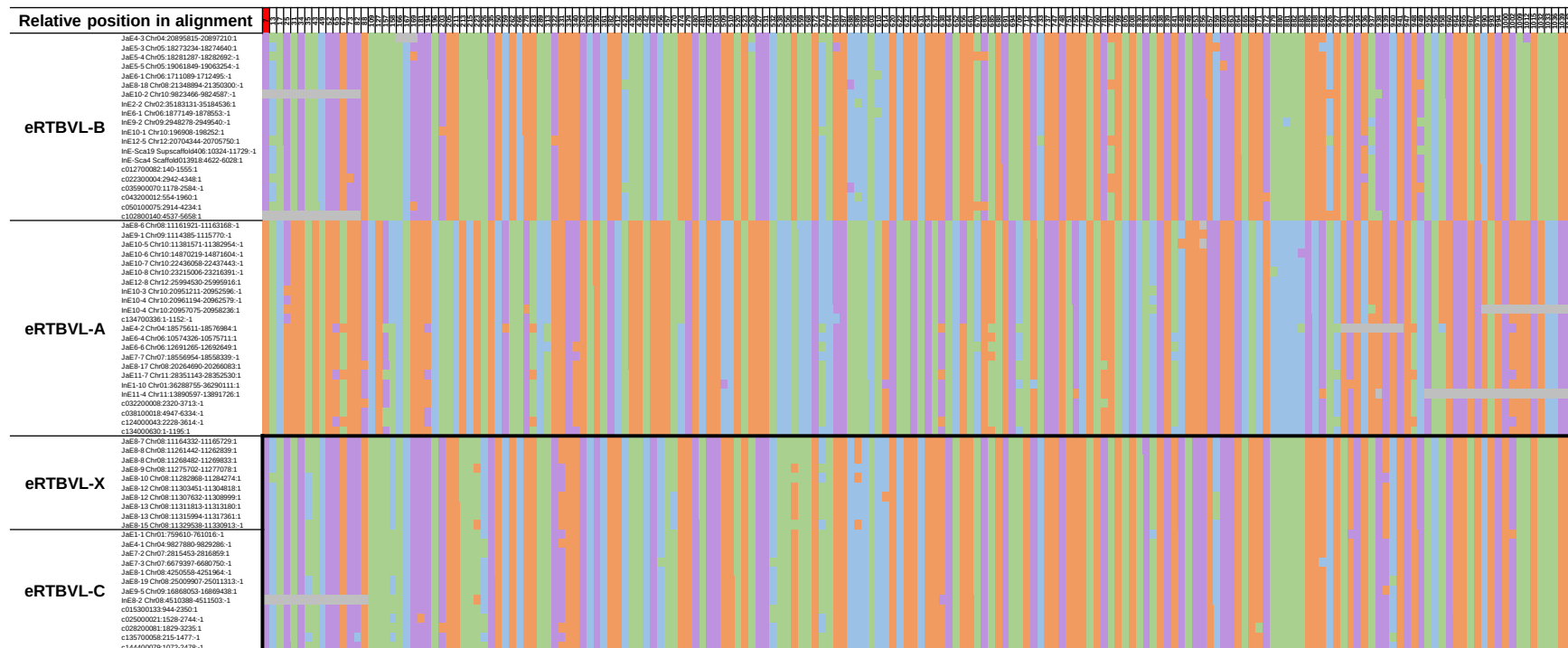


Figure S11 Informative SNP patterns in the MP sequences of eRTBVL families and presumed breakpoint positions in the alignment

Orange, A; green T; blue, C; purple, G. The ranges (relative positions) of the presumed breakpoints in the alignment are indicated in red. Missing data are indicated by hyphens. Black boxes indicate similarity between two eRTBVL families.

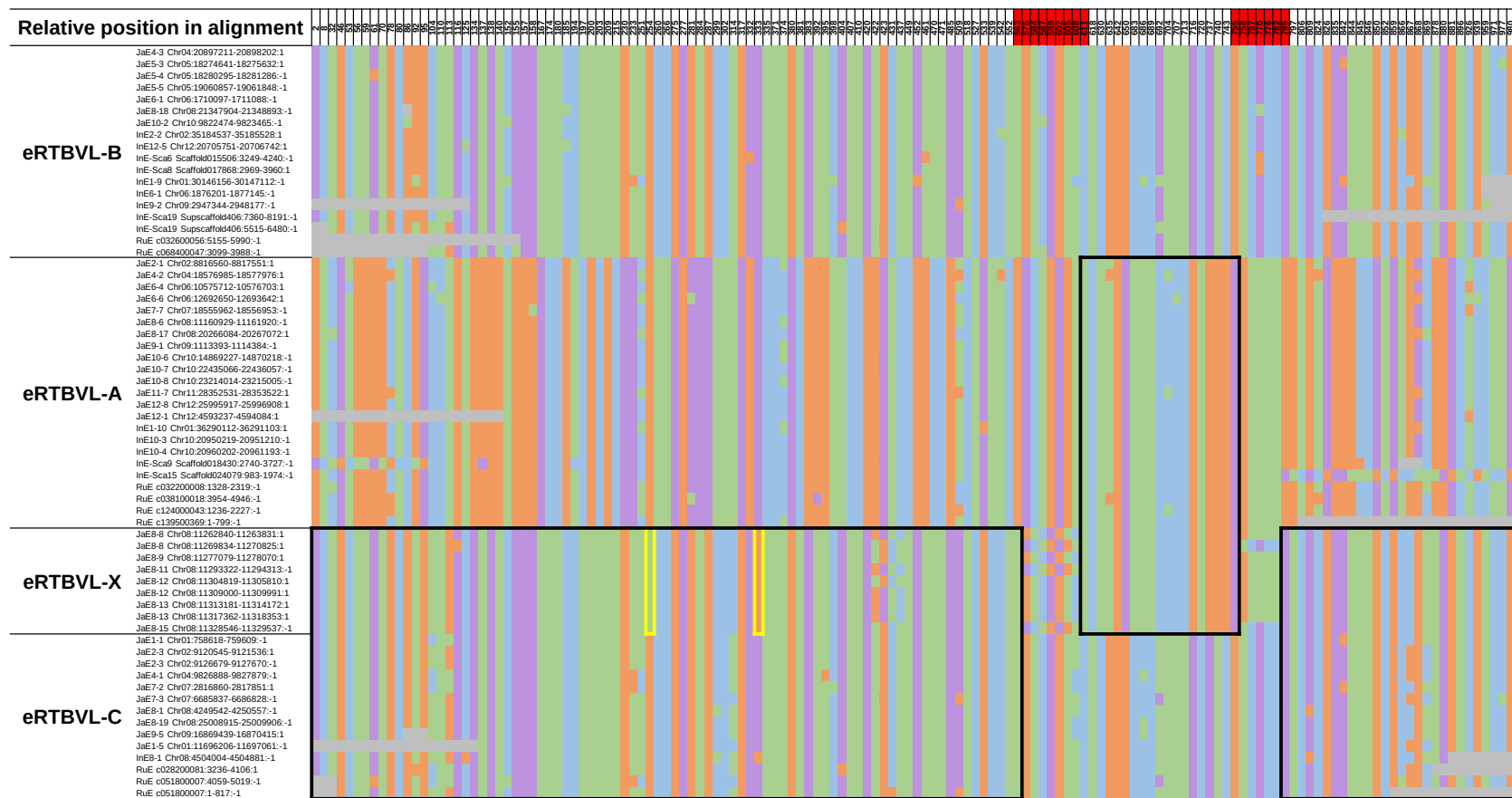


Figure S12 Informative SNP patterns in the CP sequences of eRTBVL families and presumed breakpoint positions in the alignment

Orange, A; green T; blue, C; purple, G. The ranges (relative positions) of the presumed breakpoints in the alignment are indicated in red. Missing data are indicated by hyphens. Black boxes indicate similarity between two eRTBVL families. Yellow boxes indicate eRTBVL-X-specific SNPs.



Figure S13 Informative SNP patterns in the ORFz sequences of eRTBVL families and presumed breakpoint positions in the alignment

Orange, A; green T; blue, C; purple, G. The ranges (relative positions) of the presumed breakpoints in the alignment are indicated in red. Missing data are indicated by hyphens. Black boxes indicate similarity between two eRTBVL families. Yellow boxes indicate eRTBVL-X-specific SNPs.



Figure S14 Motif alignments in the four domains of rice EPRVs and various exogenous pararetrovirus sequences

Portions of sequence alignments are shown for displaying the conserved motifs in the MP (A), PR (B), RT/RH (C) and CP (D) domains. The each conserved motif is indicated by red lines. The amino acid abbreviations in yellow (and red font), blue (and indigo font), and green backgrounds represent identical site (100%), conservative site (>50%, <100%), and block of similar, respectively. The abbreviations in green font show weakly similar residues (>20%, <50%). Slashes represent sequence omissions. The accession numbers can be found in Figures 4.2 and 4.3.

Supplementary Tables

Table S1 Plant materials used in this study

1. <i>Oryza sativa</i>				
Accession ID	Name	Original area	Subspecies	Remark
WRC 01	Nipponbare	Japan	<i>Japonica</i>	modern
WRC 02	Kasalath	India	<i>Indica</i>	landrace
WRC 03	Bei Khe	Cambodia	<i>Indica</i>	landrace
WRC 04	Jena 035	Nepal	<i>Indica</i>	landrace
WRC 05	Naba	India	<i>Indica</i>	landrace
WRC 07	Davao 1	Philippines	<i>Indica</i>	landrace
WRC 09	Ryou Suisan Koumai	China	<i>Indica</i>	landrace
WRC 11	Jinguoyin	China	<i>Indica</i>	landrace
WRC 12	Dahonggu	China	<i>Indica</i>	landrace
WRC 13	Asu	Bhutan	<i>Indica</i>	landrace
WRC 14	IR 58	Philippines	<i>Indica</i>	modern
WRC 15	Co 13	India	<i>Indica</i>	landrace
WRC 16	Vary Futsi	Madagascar	<i>Indica</i>	landrace
WRC 17	Keiboba	China	<i>Indica</i>	landrace
WRC 18	Qingyu	China	<i>Indica</i>	landrace
WRC 19	Deng Pao Zhai	China	<i>Indica</i>	modern
WRC 20	Tadukan	Philippines	<i>Indica</i>	landrace
WRC 21	Shwe Nang Gyi	Myanmar (Burma)	<i>Indica</i>	landrace
WRC 25	Muha	Indonesia	<i>Indica</i>	landrace
WRC 26	Jhona 2	India	<i>Indica</i>	landrace
WRC 29	Kalo Dhan	Nepal	<i>Indica</i>	landrace
WRC 30	Anjana Dhan	Nepal	<i>Indica</i>	landrace
WRC 31	Shoni	Bangladesh	<i>Indica</i>	landrace
WRC 32	Tupa 121-3	Bangladesh	<i>Indica</i>	landrace
WRC 33	Surjamukhi	India	<i>Indica</i>	modern
WRC 34	ARC 7291	India	<i>Indica</i>	landrace
WRC 35	ARC 5955	India	<i>Indica</i>	landrace
WRC 36	Ratul	India	<i>Indica</i>	landrace
WRC 38	ARC 11094	India	<i>Indica</i>	landrace
WRC 39	Badari Dhan	Nepal	<i>Indica</i>	landrace
WRC 40	Nepal 555	India	<i>Indica</i>	landrace
WRC 41	Kaluheenati	Sri Lanka	<i>Indica</i>	landrace
WRC 42	Local Basmati	India	<i>Indica</i>	landrace
WRC 43	Dianyu 1	China	<i>Japonica</i>	modern
WRC 44	Basilanon	Philippines	<i>Indica</i>	landrace
WRC 45	Ma sho	Myanmar (Burma)	<i>Japonica</i>	landrace
WRC 46	Khao Nok	Laos	<i>Japonica</i>	landrace
WRC 48	Khau Mac Kho	Vietnam	<i>Japonica</i>	landrace
WRC 49	Padi Perak	Indonesia	<i>Japonica</i>	landrace
WRC 50	Rexmont	America	<i>Japonica</i>	modern
WRC 51	Urasan 1	Japan	<i>Japonica</i>	landrace
WRC 52	Khau Tan Chiem	Vietnam	<i>Japonica</i>	landrace
WRC 53	Tima	Bhutan	<i>Japonica</i>	landrace
WRC 55	Tupa729	Bangladesh	<i>Japonica</i>	landrace
WRC 57	Milyang 23	South Korea	<i>Indica</i>	modern
WRC 58	Neang Menh	Cambodia	<i>Indica</i>	landrace
WRC 59	Neang Phtong	Cambodia	<i>Indica</i>	landrace
WRC 60	Hakphaynhay	Laos	<i>Indica</i>	landrace
WRC 61	Radin Goi Sesat	Malaysia	<i>Indica</i>	landrace
WRC 62	Kemasin	Malaysia	<i>Indica</i>	landrace
WRC 63	Bleiyo	Thailand	<i>Indica</i>	landrace
WRC 64	Padi Kuning	Indonesia	<i>Indica</i>	landrace
WRC 65	Rambhog	India	<i>Indica</i>	landrace
WRC 66	Bingala	Myanmar (Burma)	<i>Indica</i>	landrace
WRC 67	Phulba	India	<i>Japonica</i>	landrace
WRC 68	Khao Nam Jen	Laos	<i>Japonica</i>	landrace

WRC 97	Chin Galay	Myanmar(Burma)	<i>Indica</i>	landrace
WRC 98	Deejiaohualuo	China	<i>Indica</i>	landrace
WRC 100	Vandaran	Sri Lanka	<i>Indica</i>	landrace
NA	Taichung 65	China (Taiwan)	<i>Japonica</i>	modern
NA	Kokushokuto-2	Japan	<i>Japonica</i>	landrace
NA	Koshihikari	Japan	<i>Japonica</i>	modern
NA	IR36	Philippines	<i>Indica</i>	modern
NA	#108	China	<i>Indica</i>	NA
NA	93-11	China	<i>Indica</i>	modern
2. <i>Oryza rufipogon</i>				
Accession ID	Original area	Latitude	Longitude	
W0106	India	20.46	85.88	
W0107	India	20.95	85.10	
W0120	India	20.46	85.88	
W0137	India	16.91	81.82	
W0593	Malaysia	3.14	101.69	
W0610	Burma	16.80	96.15	
W0630	Myanmar (Burma)	20.46	94.56	
W1230	Dutch New Guinea	-4.63	138.93	
W1236	Australian New Guinea	-5.31	141.61	
W1294	Philippines	7.86	124.86	
W1551	Thailand	14.50	100.89	
W1681	India	20.09	84.45	
W1685	India	20.14	85.47	
W1690	Thailand	19.41	99.34	
W1715	China	NA	NA	
W1718	China	NA	NA	
W1807	Sri Lanka	6.93	79.95	
W1852	Thailand	20.28	100.09	
W1865	Thailand	14.56	100.98	
W1866	Thailand	14.57	100.99	
W1921	Thailand	14.44	100.90	
W1943	China	NA	NA	
W1944	China	NA	NA	
W1952	China	NA	NA	
W2003	India	15.30	73.50	
W2014	India	20.18	72.55	
W2051	Bangladesh	23.71	90.41	
W2078	Australia	-14.30	132.40	
W2263	Cambodia	11.33	104.50	
W2265	Laos	14.50	105.49	
3. Other <i>Oryza</i> species				
Accession ID	Species	<i>Oryza</i> group	Original area	
IRGC103777	<i>O. glaberrima</i>	AA	Mali	
IRGC96717	<i>O. glaberrima</i>	AA	Senegal	
W1588	<i>O. barthii</i>	AA	Cameroun	
W2112	<i>O. meridionalis</i>	AA	Australia	
W1625	<i>O. meridionalis</i>	AA	Australia	
W1169	<i>O. glumaepatula</i>	AA	Cuba	
W2199	<i>O. glumaepatula</i>	AA	Brazil	
W1514	<i>O. punctata</i>	BB	Kenya	
W1361	<i>O. officinalis</i>	CC	Malaysia	
W0065	<i>O. officinalis</i>	CC	Thailand	
W0008	<i>O. australiensis</i>	EE	Australia	
W1402	<i>O. brachyantha</i>	FF	Sierra Leone	
W0067(B)	<i>O. granulata</i>	GG	Thailand	

Latitude and longitude data are from Huang *et al* (2012). NA, not available.

Table S2 Information of investigated grass genome databases

Genome	Common name	Data source	Data version	References
<i>Oryza sativa (japonica)</i>	Rice (AA-genome)	RAP-DB	IRGSP build 5/ IRGSP-1.0	Tanaka <i>et al</i> (2008)
<i>Oryza sativa (indica)</i>	Rice (AA-genome)	BGI-RISV2	ASM465v1	Zhao <i>et al</i> (2004)
<i>Oryza rufipogon</i>	Common wild rice (AA-genome)	RiceHap3	W1943_v2	Huang <i>et al</i> (2012)
<i>Oryza glaberrima</i>	African rice (AA-genome)	Gramene	AGI1.1	Monaco <i>et al</i> (2014)
<i>Oryza barthii</i>	Wild rice (AA-genome)	Gramene	OGE.2012Jul	Monaco <i>et al</i> (2014)
<i>Oryza glumaepatula</i>	Wild rice (AA-genome)	Gramene	ALNU02000000	Monaco <i>et al</i> (2014)
<i>Oryza meridionalis</i>	Wild rice (AA-genome)	Gramene	ALNW00000000	Monaco <i>et al</i> (2014)
<i>Oryza punctata</i>	Wild rice (BB-genome)	Gramene	AVCL00000000	Monaco <i>et al</i> (2014)
<i>Oryza brachyantha</i>	Wild rice (FF-genome)	Gramene	V1.4b	Monaco <i>et al</i> (2014)
<i>Hordeum vulgare</i>	Barley	Gramene	IBSC_1.0	Monaco <i>et al</i> (2014)
<i>Triticum aestivum</i>	Bread wheat	Gramene	IWGSP1	Monaco <i>et al</i> (2014)
<i>Brachypodium distachyon</i>	Purple false brome	Phytozome	JGI v1.0	Goodstein <i>et al</i> (2012)
<i>Phyllostachys heterocycla</i>	Moso bamboo	NCGR	V1.0	Peng <i>et al</i> (2013)
<i>Zea mays</i>	Maize	Gramene	AGPv3	Monaco <i>et al</i> (2014)
<i>Sorghum bicolor</i>	Sorghum	Phytozome	JGI v2.1	Goodstein <i>et al</i> (2012)
<i>Setaria italica</i>	Foxtail millet	Phytozome	JGI v2.0	Goodstein <i>et al</i> (2012)
<i>Panicum virgatum</i>	Switchgrass	Phytozome	V1.1	Goodstein <i>et al</i> (2012)

The version IRGSP build 5 of the genome assembly of *Oryza sativa japonica* was used in Chapter 2, and the version IRGSP-1.0 was used in Chapters 4 and 5.

Table S3 Mapping of W1943-eRTBVLs into the Nipponbare and 93-11 genomes based on flanking sequence alignments

W1943 eRTBVLs			Counterparts in Nipponbare		Counterparts in 93-11	
ID ^a	Location ^b	Length ^c	ID ^a	Location ^b	ID ^a	Location ^b
<i>Mapping of Nipponbare and 93-11-shared eRTBVLs</i>						
RuE1	c002200377:627-4879	4253/9524	JaE1-4	Chr1:5192920-5197168	InE1-5	Chr1:5590870-5595117
RuE2	c085300093:14836-15696	861/22793	JaE2-3.1	Chr2:12149381-12150241	InE2-0.1	Chr2:12927855-12928715
RuE3	c095500041:6752-7013	262/10631	JaE4-3.1	Chr4:21320347-21320609	InE4-0.1	Chr4:18989244-18989506
RuE4	c052900044:8701-9031	331/17468	JaE5-0.1	Chr5:6506730-6507060	InE5-0.1	Chr5:6768278-6768608
RuE5	c086800018:3190-3840	651/13046	JaE5-1	Chr5:13117697-13118346	InE5-1	Chr5:14091105-14091754
RuE6	c015200107:1-3467	3467/4984	JaE6-2	Chr6:4974982-4976200	InE6-3	Chr6:5576679-5577999
RuE7	c047500058:13617-16774	3158/28845	JaE7-4	Chr7:8938366-8941525	InE7-3	Chr7:8918467-8921134
RuE8	c047500058:10909-11796	888/28845	JaE7-5	Chr7:8943333-8944229	InE7-4	Chr7:8922955-8923849
RuE9	c047500058:2731-3343	613/28845	JaE7-6	Chr7:8951794-8952406	InE7-5	Chr7:8931205-8931817
RuE10	c133601507:422-698	277/699	JaE8-2	Chr8:4256114-4256290	InE8-3	Chr8:4514252-4514528
RuE11	c006100100:162-657	496/849	JaE8-3	Chr8:4258857-4259350	InE8-5	Chr8:4523301-4523790
RuE12	c100500038:235-1033	799/9223	JaE8-4	Chr8:9707851-9708649	InE8-6	Chr8:10824714-10825512
RuE13	c002000018:1-454	454/8572	JaE8-19	Chr8:25004409-25004862	InE8-8	Chr8:26561718-26562130
RuE14	c076000081:11164-11790	627/11790	JaE8-20	Chr8:25129069-25129611	InE8-9	Chr8:26694836-26695383
RuE15	c012500363:4326-4927	602/8985	JaE11-2	Chr11:5051782-5052373	InE9-3	Chr9:11810021-11810622
RuE16	c115300050:4241-6876	2636/11499	JaE11-5	Chr11:11697969-11700596	InE11-1	Chr11:10457049-10459678
RuE17	c032600056:519-5990	5472/5990	JaE12-6	Chr12:25297171-25300975	InE12-5	Chr12:20705907-20711312
<i>Mapping of Nipponbare-shared eRTBVLs</i>						
RuE18	c135700032:719-1443	725/1443	JaE4-1	Chr4:9822314-9823020	-	-
RuE19	c106700084:1-1577	1577/29400	JaE4-2	Chr4:18580877-18582415	-	-
RuE20	c137600035:1-521	521/9892	JaE4-3	Chr4:20900760-20901280	-	-
	c022300089:1-1453	1453/23973		Chr4:20895333-20896775		
RuE21	c144000081:1-1523	1523/7527	JaE5-3	Chr5:18272252-18273774	-	-
RuE22	c106700079:12845-13092	248/13092	JaE6-6	Chr6:12697841-12698089	-	-
RuE23	c138500632:275-379	105/379	JaE8-6	Chr8:11163064-11163168	-	-
RuE24	c133602563:1-121	121/325	JaE8-9	Chr8:11274760-11274880	-	-
	c134000811:1-248	248/339		Chr8:11274633-11274880		
RuE25	c029700105:711-1498	788/1498	JaE8-11	Chr8:11299195-11299982	-	-
RuE26	c142700765:1-585	585/955	JaE8-14	Chr8:11324551-11325135	-	-
RuE27	c143200273:1-540	540/787	JaE8-16	Chr8:11334984-11335523	-	-
RuE28	c133900080:53-1089	1037/1089	JaE8-17	Chr8:20269986-20271022	-	-
RuE29	c126700110:1-3076	3076/3353	JaE8-18	Chr8:21343338-21346401	-	-

RuE30	c008300053:374-1202	829/1637	JaE9-2	Chr9:15198845-15199673	-	-
RuE31	c000600075:89-657	569/657	JaE10-5	Chr10:11381448-11382013	-	-
RuE32	c019600008:1-468	468/3415	JaE10-6	Chr10:14868133-14868600	-	-
RuE33	c149100216:1-705	705/1035	JaE11-3	Chr11:9663953-9664657	-	-
RuE34	c006700114:1-424	424/9613	JaE11-4	Chr11:9670041-9670464	-	-
RuE35	c162200035:616-845	230/20119	JaE11-6.1	Chr11:28346180-28346390	-	-
RuE36	c124000043:57-4957	4901/5258	JaE11-7	Chr11:28354701-28349748	-	-
<i>Mapping of 93-11-shared eRTBVLs</i>						
RuE37	c012100082:1-1768	1768/1958	-	-	InE1-9	Chr1:30148632-30150182
RuE38	c080800102:1-450	450/17856	-	-	InE7-2	Chr7:6563107-6563556
RuE39	c142900526:1-233	233/299	-	-	InE10-4	Chr10:20963461-20963693
RuE40	c157500016:1-1028	1028/29044	-	-	InE11-3	Chr11:13668090-13669117
<i>Mapping of W1943-specific eRTBVLs</i>						
RuE41	c141100409:1-444	444/1177	-	Chr1:2089963-2089964	-	Chr1:2459347-2459348
RuE42	c050700062:4-2553	2550/6461	-	Chr1:22068739-22068783	-	Chr1:22793617-22793644
RuE43	c109900063:1-4203	4203/18514	-	Chr1:35517224-35517243	-	Chr1:37196766-37196787
RuE44	c125800229:17476-19028	1553/29016	-	Chr3:5786307-5786328	-	Chr3:6212685-6212744
RuE45	c005200381:1-558	558/7091	-	Chr4:19084783-19084784	-	Chr4:16434200-16434201
RuE46	c102400019:7133-7925	793/7925	-	Chr5:28255725-28255726	-	Chr5:29470779-29470780
RuE47	c061800061:4819-5854	1036/5957	-	Chr7:12252718-12252780	-	Chr1:11518644-11518700
RuE48	c023700009:22606-23283	678/23283	-	Chr8:20209752-20209784	-	Chr8:21520867-21520939
	c072900063:1-463	463/13707	-			
RuE49	c156000004:5-709	705/13642	-	Chr9:4925537-4925793	-	Chr9:4058473-4058820
RuE50	c139500660:973-1424	452/1424	-	Chr9:8704450-8704451	-	Chr9:7295057-7295058
RuE51	c046700160:1-515	515/26631	-	Chr10:20857886-20857887	-	Chr10:19348967-19348968
RuE52	c102400018:1-1144	1144/3017	-	Chr3:6554347-6554355	-	-
RuE53	c141300135:1-542	542/1326	-	Chr5:20518315-20519062	-	-
RuE54	c031900088:1-637	637/9173	-	Chr6:2440575-2440608	-	-
RuE55	c138000180:1-873	873/2222	-	Chr12:25931122-25931653	-	-
RuE56	c117000040:1-710	710/27923	-	-	-	Chr3:5304305-5304373
RuE57	c077800067:1-1255	1255/10891	-	-	-	Chr3:18288062-18288208

^aNumber of eRTBVLs in W1943, Nipponbare and 93-11 genomes used by this study and Liu *et al* (2012).

^bLocations of shared eRTBVLs in contigs or chromosomes, and the corresponding empty donor sites in the *japonica* and *indica* genomes for W1943- specific eRTBVLs.

^cThe eRTBVL length in contigs/the length of contigs.

Table S4 Detection of the eRTBVL-X cluster in cultivated and wild rice accessions by genomic PCR

Material information			PCR detection of the eRTBVL-X cluster sequences											Conclusion
Accession ID	Original area	Species	Part 1	Part 2	Part 3	Part 4	Part 5	Part 6	Part 7	Part 8	Part 9	Part 10	Whole	
Nipponbare	Japan	<i>Japonica</i>	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
Taichung 65	China (Taiwan)	<i>Japonica</i>	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
Kokushokuto-2	Japan	<i>Japonica</i>	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
Koshihikari	Japan	<i>Japonica</i>	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
WRC 68	Laos	<i>Japonica</i>	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	-	Y
WRC 43	China	<i>Japonica</i>	Y	Y	Y	Y	Y	-	Y	Y	Y	Y	-	Y
WRC 45	Myanmar(Burma)	<i>Japonica</i>	-	Y	-	-	-	-	-	-	-	-	-	?
WRC 49	Indonesia	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	-	-
WRC 46	Laos	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 48	Vietnam	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 50	America	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 51	Japan	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 52	Vietnam	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 53	Bhutan	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 55	Bangladesh	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 67	India	<i>Japonica</i>	-	-	-	-	-	-	-	-	-	-	E	-
IR36	Philippines	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
#108	China	<i>Indica</i>	-	?	-	-	-	-	-	-	-	-	E	-
93-11	China	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 02	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 03	Cambodia	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 04	Nepal	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 05	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 07	Philippines	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-
WRC 09	China	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 11	China	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 12	China	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 13	Bhutan	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 14	Philippines	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 15	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-
WRC 16	Madagascar	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-
WRC 17	China	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	E	-

WRC 18	China(Taiwan)	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 19	China	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 20	Philippines	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-
WRC 21	Myanmar(Burma)	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-
WRC 25	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 26	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 29	Nepal	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 30	Nepal	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-
WRC 31	Bangladesh	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 32	Bangladesh	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 33	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 34	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 35	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-
WRC 36	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 38	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 39	Nepal	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-
WRC 40	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 41	Sri Lanka	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-
WRC 42	India	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 44	Philippines	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 57	South Korea	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 58	Cambodia	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 59	Cambodia	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 60	Laos	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 61	Malaysia	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 62	Malaysia	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 63	Thailand	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 64	Indonesia	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 65	Indonesia	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 66	Myanmar(Burma)	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	-	-
WRC 97	Myanmar(Burma)	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 98	China	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
WRC 100	Sri Lanka	<i>Indica</i>	-	-	-	-	-	-	-	-	-	-	-	E	-
W1715	China	<i>O. rufipogon</i>	Y	Y	Y	-	-	-	-	-	-	-	-	-	Y
W1718	China	<i>O. rufipogon</i>	Y	Y	Y	-	-	-	-	-	-	-	-	-	Y
W1943	China	<i>O. rufipogon</i>	Y	Y	Y	Y	-	-	Y	Y	Y	Y	-	-	Y
W1944	China	<i>O. rufipogon</i>	Y	Y	-	-	-	-	-	-	-	-	?	-	Y

W1952	China	<i>O. rufipogon</i>	Y	Y	-	-	-	Y	-	-	-	?	-	Y
W0106	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W0107	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W0120	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W0137	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W0593	Malaysia	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W0610	Burma	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W0630	Myanmar (Burma)	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W1230	Dutch New Guinea	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W1236	Australian New	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W1294	Philippines	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W1551	Thailand	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W1681	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W1685	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W1690	Thailand	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W1807	Sri Lanka	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W1852	Thailand	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W1865	Thailand	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W1866	Thailand	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W1921	Thailand	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W2003	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W2014	India	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W2051	Bangladesh	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W2078	Australia	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-
W2263	Cambodia	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	E	-
W2265	Laos	<i>O. rufipogon</i>	-	-	-	-	-	-	-	-	-	-	-	-

Ten regions of the eRTBVL-X cluster sequences were examined using different primer sets. Primer sets: part 1, JaE8-5-1-2 & JaE8-5-4; part 2, JaE8-8-1 & JaE8-8-2; part 3, JaE8-8-3 & JaE8-8-4; part 4, JaE8-9-1 & JaE8-9-2; part 5, JaE8-10-1 & JaE8-10-2; part 6, JaE8-10-3 & JaE8-10-4; part 7, JaE8-11-3 & JaE8-11-4; part 8, JaE8-13-1 & JaE8-13-2; part 9, JaE8-13-3 & JaE8-13-4; part 10, JaE8-15-3 & JaE8-15-4; whole, JaE8-5-L-1 & JaE8-16-4. The absence of the whole cluster was detected using one of the primer pairs; a predicted small band for the empty donor site would be observed if the cluster was absent. Target bands are indicated by the letter Y; no band by hyphens; bands for empty donor site by the letter E. Ambiguous cases are indicated by a question mark. Detailed information of materials and primers is available in Supplementary Tables S1 and S10.

Table S5 Genomic locations of orthologous eRTBVL-D loci in *Oryza* genomes

ID	<i>O. sativa japonica</i>	<i>O. sativa indica</i>	<i>O. rufipogonn</i>	<i>O. barthii</i>	<i>O. glaberrima</i>	<i>O. glumaepatula</i>	<i>O. meridionalis</i>	<i>O. punctata</i>
JaE1-4	1:5194409-5197055	1:5592358-5595003	c002200377:2115-4765	1:4385376-4388048	1:4154706-4158078	1:5279651-5303960	1:5261426-5264514	Absence
JaE2-3.1	2:12149381-12150241	2:12927855-12928715	c085300093:14836-15696	2:11025428-11026288	By PCR	2:12220876-12221520	By PCR	2:12163171-12165278
JaE4-2.1	4:18711125-18711492	4:15946417-15946784	c175900023:6560-6928	4:12208515-12208883	4:11186433-11186799	4:15674937-15675110	4:14453108-14453480	Absence
JaE4-3.1	4:21320364-21320606	4:18989261-18989503	c095500041:6755-6996	4:14694704-14694946	4:13690402-13690644	By PCR	4:17045107-17045242	Absence
JaE5-0.1	5:6506730-6507060	5:6768278-6768608	c052900044:8701-9031	5:5433872-5434189	By PCR	5:6588049-6588342	Unknown	Absence
JaE7-4	7:8938366-8941514	7:8918464-8920735	c047500058:13628-16774	7:7802381-7805527	7:7646229-7649634	1:39938892-39942155	Unknown	Absence
JaE7-5	7:8942869-8944249	7:8922478-8923869	c047500058:10889-12273	7:7819843-7821235	7:7664965-7665431	1:39920345-39921717	By PCR	Absence
JaE7-6	7:8951794-8952406	7:8931205-8931817	c047500058:2731-3343	7:7828580-7829192	7:7672776-7673382	1:39912404-39913017	By PCR	Absence
JaE9-5.1	9:17432919-17433026	9:15279153-15279260	c079500005:10397-10504	9:13332228-13332335	9:12218973-12219080	Unknown	9:13757744-13757848	Absence
JaE10-0.1	10:9436733-9436862	10:7478152-7478281	c161100022:8644-8773	10:7045450-7045579	PCR	10:8287110-8287216	Unknown	Absence
JaE11-2	11:5051782-5054834	9:11810621-11810040	c012500363:4327-4908	11:4551560-4554632	Oglab11_unplaced045:205-3284	11:4748693-4749306	11:4424168-4424276	Absence
JaE11-5	11:11697969-11700596	11:10457049-10459678	c115300050:4241-6876	11:9891200-9893829	11:9061409-9063961	11:11473828-11476492	Unknown	Absence

The genomic locations of eRTBVL-D loci in each genome are shown as chromosome/contig: start position- end position.

Table S6 Sequence distribution of eRTBVL superfamily in grass genomes.

Genome	ID	Chromosome/scaffold/contig	Start position	End position	Length
<i>Oryza punctata</i>	eRTBVL-OP-1	1	2991637	2993653	2017
	eRTBVL-OP-2	1	14284720	14284870	151
	eRTBVL-OP-3	1	14286952	14287154	203
	eRTBVL-OP-4	1	16694154	16697681	3528
	eRTBVL-OP-5	1	18381973	18382421	449
	eRTBVL-OP-6	1	28854385	28854627	243
	eRTBVL-OP-7	1	29553237	29562967	9731
	eRTBVL-OP-8	1	42782091	42782245	155
	eRTBVL-OP-9	2	7778983	7779281	299
	eRTBVL-OP-10	2	7791389	7792202	814
	eRTBVL-OP-11	2	7801700	7809490	7791
	eRTBVL-OP-12	2	8232845	8234447	1603
	eRTBVL-OP-13	2	10670239	10670566	328
	eRTBVL-OP-14	2	12163171	12165278	2108
	eRTBVL-OP-15	2	19866001	19873049	7049
	eRTBVL-OP-16	2	22540024	22541566	1543
	eRTBVL-OP-17	2	34848236	34849992	1757
	eRTBVL-OP-18	2	35449771	35449953	183
	eRTBVL-OP-19	2	39302009	39302143	135
	eRTBVL-OP-20	3	810511	810755	245
	eRTBVL-OP-21	3	5298607	5299004	398
	eRTBVL-OP-22	3	5920652	5920931	280
	eRTBVL-OP-23	3	8840551	8840940	390
	eRTBVL-OP-24	3	13165779	13173649	7871
	eRTBVL-OP-25	3	15812170	15819655	7486
	eRTBVL-OP-26	3	19772487	19772627	141
	eRTBVL-OP-27	3	31523706	31526434	2729
	eRTBVL-OP-28	3	31529741	31531553	1813
	eRTBVL-OP-29	4	510481	511241	761
	eRTBVL-OP-30	4	12590574	12591164	591
	eRTBVL-OP-31	4	18118065	18118308	244
	eRTBVL-OP-32	4	18611842	18612627	786
	eRTBVL-OP-33	4	18619123	18619908	786
	eRTBVL-OP-34	4	25197779	25198108	330
	eRTBVL-OP-35	4	28516348	28524687	8340
	eRTBVL-OP-36	4	29384246	29385660	1415
	eRTBVL-OP-37	5	2883966	2884285	320
	eRTBVL-OP-38	5	4091107	4095649	4543
	eRTBVL-OP-39	5	10820908	10822626	1719
	eRTBVL-OP-40	5	11646575	11648846	2272
	eRTBVL-OP-41	5	12370834	12371598	765
	eRTBVL-OP-42	5	12875475	12877970	2496
	eRTBVL-OP-43	5	14632837	14633457	621
	eRTBVL-OP-44	6	9592294	9599270	6977
	eRTBVL-OP-45	6	15792888	15795171	2284
	eRTBVL-OP-46	6	15807217	15812905	5689
	eRTBVL-OP-47	6	23051417	23057746	6330
	eRTBVL-OP-48	6	29928010	29928233	224
	eRTBVL-OP-49	7	8490686	8491010	325
	eRTBVL-OP-50	7	29852137	29855717	3581
	eRTBVL-OP-51	8	5187570	5189979	2410
	eRTBVL-OP-52	8	6744104	6746367	2264
	eRTBVL-OP-53	8	8872479	8873150	672
	eRTBVL-OP-54	8	13071716	13077767	6052
	eRTBVL-OP-55	8	17473499	17477290	3792
	eRTBVL-OP-56	9	2808065	2813526	5462
	eRTBVL-OP-57	9	5882503	5882651	149
	eRTBVL-OP-58	9	14957195	14964303	7109
	eRTBVL-OP-59	9	15026831	15027049	219
	eRTBVL-OP-60	10	4307255	4309355	2101
	eRTBVL-OP-61	10	5909677	5910093	417
	eRTBVL-OP-62	10	19170135	19175758	5624
	eRTBVL-OP-63	11	2036323	2036512	190
	eRTBVL-OP-64	11	5637985	5639591	1607
	eRTBVL-OP-65	11	5839046	5839566	521
	eRTBVL-OP-66	11	9318460	9318663	204
	eRTBVL-OP-67	11	11283297	11292028	8732
	eRTBVL-OP-68	11	11295688	11296393	706
	eRTBVL-OP-69	11	22158726	22167385	8660
	eRTBVL-OP-70	11	22592997	22599450	6454
	eRTBVL-OP-71	12	9467202	9468962	1761
	eRTBVL-OP-72	12	10696900	10697936	1037
	eRTBVL-OP-73	12	11677313	11685240	7928
	eRTBVL-OP-74	12	17250761	17251498	738
	eRTBVL-OP-75	12	26770028	26770398	371
<i>Oryza brachyantha</i>	eRTBVL-OB-1	1	18058773	18059171	399

	eRTBVL-OB-2	2	11578717	11580679	1963
	eRTBVL-OB-3	2	12007128	12007263	136
	eRTBVL-OB-4	4	6003897	6005362	1466
	eRTBVL-OB-5	4	6009909	6011199	1291
	eRTBVL-OB-6	5	5113734	5114590	857
	eRTBVL-OB-7	5	20008565	20008713	149
	eRTBVL-OB-8	6	15923677	15924943	1267
	eRTBVL-OB-9	7	593226	593414	189
	eRTBVL-OB-10	8	5658525	5661140	2616
	eRTBVL-OB-11	8	7725024	7725207	184
	eRTBVL-OB-12	8	9789438	9789659	222
	eRTBVL-OB-13	8	10434712	10435326	615
	eRTBVL-OB-14	8	14614685	14615071	387
	eRTBVL-OB-15	9	7056665	7057504	840
	eRTBVL-OB-16	10	13424929	13426129	1201
	eRTBVL-OB-17	11	9627649	9627806	158
	eRTBVL-OB-18	12	1820803	1821506	704
	eRTBVL-OB-19	super0063	36123	36334	212
<i>Hordeum vulgare</i>	eRTBVL-HV-1	1	288613654	288613940	287
	eRTBVL-HV-2	2	196375796	196376510	715
	eRTBVL-HV-3	2	506578530	506578729	200
	eRTBVL-HV-4	2	574669483	574669634	152
	eRTBVL-HV-5	2	618204915	618205237	323
	eRTBVL-HV-6	2	618373007	618373159	153
	eRTBVL-HV-7	3	209184571	209189150	4580
	eRTBVL-HV-8	3	230891776	230892131	356
	eRTBVL-HV-9	3	230894316	230894420	105
	eRTBVL-HV-10	3	541724153	541724313	161
	eRTBVL-HV-11	4	6812465	6812618	154
	eRTBVL-HV-12	4	72507158	72508970	1813
	eRTBVL-HV-13	4	137946106	137946455	350
	eRTBVL-HV-14	4	353846103	353847922	1820
	eRTBVL-HV-15	4	528426593	528426710	118
	eRTBVL-HV-16	5	296625422	296625642	221
	eRTBVL-HV-17	5	326071671	326071797	127
	eRTBVL-HV-18	5	335404650	335405092	443
	eRTBVL-HV-19	5	371195981	371196276	296
	eRTBVL-HV-20	5	424544851	424544981	131
	eRTBVL-HV-21	5	523451370	523452228	859
	eRTBVL-HV-22	6	119242864	119243568	705
	eRTBVL-HV-23	6	161452812	161465210	12399
	eRTBVL-HV-24	6	454062304	454063403	1100
	eRTBVL-HV-25	7	81925054	81925459	406
	eRTBVL-HV-26	7	112418024	112418919	896
	eRTBVL-HV-27	7	557419687	557420815	1129
	eRTBVL-HV-28	7	579850475	579850777	303
	eRTBVL-HV-29	7	583897969	583898246	278
	eRTBVL-HV-30	7	598152365	598152488	124
	eRTBVL-HV-31	1H	38722531	38722643	113
	eRTBVL-HV-32	1H	47039980	47040328	349
	eRTBVL-HV-33	1H	84474831	84475135	305
	eRTBVL-HV-34	1H	84496702	84497005	304
	eRTBVL-HV-35	1H	84585366	84585666	301
	eRTBVL-HV-36	1H	85715286	85715551	266
	eRTBVL-HV-37	1H	86987381	86987480	100
	eRTBVL-HV-38	1H	87430673	87430877	205
	eRTBVL-HV-39	2HL	42745751	42745987	237
	eRTBVL-HV-40	2HL	47632761	47633113	353
	eRTBVL-HV-41	2HL	47889066	47889393	328
	eRTBVL-HV-42	2HL	48025326	48025640	315
	eRTBVL-HV-43	2HL	48122362	48122668	307
	eRTBVL-HV-44	2HL	48391436	48391718	283
	eRTBVL-HV-45	2HL	48846822	48847075	254
	eRTBVL-HV-46	2HL	48911211	48911460	250
	eRTBVL-HV-47	2HL	49453480	49453694	215
	eRTBVL-HV-48	2HL	49494538	49494640	103
	eRTBVL-HV-49	2HS	2579554	2584204	4651
	eRTBVL-HV-50	2HS	23799685	23801096	1412
	eRTBVL-HV-51	2HS	36525293	36525617	325
	eRTBVL-HV-52	3HL	26137333	26138351	1019
	eRTBVL-HV-53	3HS	34699995	34700188	194
	eRTBVL-HV-54	3HS	37385737	37385942	206
	eRTBVL-HV-55	4HL	28519012	28519125	114
	eRTBVL-HV-56	4HS	3657600	3660998	3399
	eRTBVL-HV-57	4HS	12185574	12185767	194
	eRTBVL-HV-58	4HS	16961161	16962196	1036
	eRTBVL-HV-59	4HS	24352732	24355011	2280
	eRTBVL-HV-60	4HS	35738301	35738647	347
	eRTBVL-HV-61	4HS	36527786	36528097	312
	eRTBVL-HV-62	4HS	39070116	39070314	199
	eRTBVL-HV-63	4HS	40964306	40964702	397
	eRTBVL-HV-64	4HS	41652550	41652865	316
	eRTBVL-HV-65	5HL	23856620	23856739	120

	eRTBVL-HV-66	5HL	42022035	42022193	159
	eRTBVL-HV-67	5HL	45179903	45180703	801
	eRTBVL-HV-68	5HL	50648038	50648169	132
	eRTBVL-HV-69	6HL	1654218	1657153	2936
	eRTBVL-HV-70	6HL	5601605	5604313	2709
	eRTBVL-HV-71	6HS	6703378	6708537	5160
	eRTBVL-HV-72	6HS	17125133	17125249	117
	eRTBVL-HV-73	6HS	19242528	19244633	2106
	eRTBVL-HV-74	6HS	24291407	24292299	893
	eRTBVL-HV-75	6HS	30485113	30485312	200
	eRTBVL-HV-76	6HS	32501557	32501741	185
	eRTBVL-HV-77	6HS	34551581	34551844	264
	eRTBVL-HV-78	7HL	7825866	7826309	444
	eRTBVL-HV-79	7HL	18389192	18389417	226
	eRTBVL-HV-80	7HL	28666836	28667946	1111
	eRTBVL-HV-81	7HL	29755461	29757060	1600
	eRTBVL-HV-82	7HL	33392098	33393019	922
	eRTBVL-HV-83	7HL	33823433	33824502	1070
	eRTBVL-HV-84	7HL	34395531	34395638	108
	eRTBVL-HV-85	7HL	38681339	38681650	312
	eRTBVL-HV-86	7HL	38799331	38799896	566
	eRTBVL-HV-87	7HL	39565276	39565390	115
	eRTBVL-HV-88	7HL	39752296	39752579	284
	eRTBVL-HV-89	7HL	39845327	39845594	268
	eRTBVL-HV-90	7HL	39945316	39945629	314
	eRTBVL-HV-91	7HL	40168133	40168448	316
	eRTBVL-HV-92	7HL	40311763	40311870	108
	eRTBVL-HV-93	7HL	41058118	41058439	322
	eRTBVL-HV-94	7HL	41117215	41117530	316
	eRTBVL-HV-95	7HL	41189073	41189380	308
	eRTBVL-HV-96	7HL	41316179	41316478	300
	eRTBVL-HV-97	7HL	41318975	41319273	299
	eRTBVL-HV-98	7HL	41421730	41423975	2246
	eRTBVL-HV-99	7HL	41471565	41471852	288
	eRTBVL-HV-100	7HL	41490544	41490784	241
	eRTBVL-HV-101	7HL	41505785	41505884	100
	eRTBVL-HV-102	7HL	41536383	41536660	278
	eRTBVL-HV-103	7HL	41543276	41543556	281
	eRTBVL-HV-104	7HL	41628429	41628693	265
	eRTBVL-HV-105	7HL	41757159	41757336	178
	eRTBVL-HV-106	7HL	41975119	41975266	148
	eRTBVL-HV-107	7HL	42223458	42223620	163
	eRTBVL-HV-108	7HL	42402468	42402690	223
	eRTBVL-HV-109	7HL	42708006	42708119	114
	eRTBVL-HV-110	7HL	42730590	42730778	189
	eRTBVL-HV-111	7HS	13649763	13649875	113
	eRTBVL-HV-112	7HS	42964192	42964421	230
	eRTBVL-HV-113	morex_contig_45375	135	624	490
	eRTBVL-HV-114	morex_contig_80901	1	1368	1368
	eRTBVL-HV-115	morex_contig_130970	1	1201	1201
	eRTBVL-HV-116	morex_contig_147484	1410	3598	2189
	eRTBVL-HV-117	morex_contig_156386	1	1742	1742
	eRTBVL-HV-118	morex_contig_166134	61	2054	1994
	eRTBVL-HV-119	morex_contig_167750	1	1309	1309
	eRTBVL-HV-120	morex_contig_203784	912	1259	348
	eRTBVL-HV-121	morex_contig_292117	1	1232	1232
	eRTBVL-HV-122	morex_contig_464113	1	1294	1294
	eRTBVL-HV-123	morex_contig_1604379	338	1099	762
	eRTBVL-HV-124	morex_contig_1606461	266	2357	2092
<i>Triticum aestivum</i>	eRTBVL-TA-1	1AL_scaff_3973781	1	2043	2043
	eRTBVL-TA-2	1AL_scaff_3976197	2	4073	4072
	eRTBVL-TA-3	1AL_scaff_3980494	1	3905	3905
	eRTBVL-TA-4	1BL_scaff_3826911	293	1576	1284
	eRTBVL-TA-5	1BL_scaff_3889833	1	1934	1934
	eRTBVL-TA-6	1BL_scaff_3915807	3	6698	6696
	eRTBVL-TA-7	1BL_scaff_3916290	90	1246	1157
	eRTBVL-TA-8	1BS_scaff_3418523	2192	2560	369
	eRTBVL-TA-9	1BS_scaff_3425270	221	830	610
	eRTBVL-TA-10	1BS_scaff_3454473	1182	1297	116
	eRTBVL-TA-11	1BS_scaff_3472882	7269	9870	2602
	eRTBVL-TA-12	1DL_scaff_1180207	4847	5263	417
	eRTBVL-TA-13	1DL_scaff_2198867	3	7648	7646
	eRTBVL-TA-14	1DL_scaff_2224992	5112	9663	4552
	eRTBVL-TA-15	1DL_scaff_2271768	4986	5443	458
	eRTBVL-TA-16	1DL_scaff_2277483	3380	12035	8656
	eRTBVL-TA-17	1DS_scaff_726870	10496	10777	282
	eRTBVL-TA-18	1DS_scaff_1877689	2175	4090	1916
	eRTBVL-TA-19	1DS_scaff_1878683	1	1614	1614
	eRTBVL-TA-20	1DS_scaff_1894091	19650	21806	2157
	eRTBVL-TA-21	1DS_scaff_1895581	1	6383	6383
	eRTBVL-TA-22	1DS_scaff_1898152	275	2418	2144
	eRTBVL-TA-23	1DS_scaff_1901613	1	502	502
	eRTBVL-TA-24	1DS_scaff_1902238	3001	7206	4206

eRTBVL-TA-25	1DS_scaff_1906885	4137	4483	347
eRTBVL-TA-26	1DS_scaff_1912672	1	5806	5806
eRTBVL-TA-27	1DS_scaff_1914170	1	5535	5535
eRTBVL-TA-28	1DS_scaff_1914990	1	4696	4696
eRTBVL-TA-29	1DS_scaff_1915292	4156	6488	2333
eRTBVL-TA-30	2AL_scaff_6374805	5755	7911	2157
eRTBVL-TA-31	2AL_scaff_6374805	11030	11678	649
eRTBVL-TA-32	2AL_scaff_6411617	84	7755	7672
eRTBVL-TA-33	2AL_scaff_6437445	26577	30876	4300
eRTBVL-TA-34	2AS_scaff_4883773	900	1101	202
eRTBVL-TA-35	2AS_scaff_5200869	5019	5143	125
eRTBVL-TA-36	2AS_scaff_5227641	586	4030	3445
eRTBVL-TA-37	2AS_scaff_5256876	406	6105	5700
eRTBVL-TA-38	2AS_scaff_5258431	1930	5577	3648
eRTBVL-TA-39	2BL_scaff_8022475	343	3883	3541
eRTBVL-TA-40	2BL_scaff_8050356	606	1140	535
eRTBVL-TA-41	2BL_scaff_8093120	488	4332	3845
eRTBVL-TA-42	2BS_scaff_5157340	4060	5184	1125
eRTBVL-TA-43	2BS_scaff_5160216	764	3090	2327
eRTBVL-TA-44	2BS_scaff_5166077	56	2382	2327
eRTBVL-TA-45	2BS_scaff_5178888	41	6946	6906
eRTBVL-TA-46	2BS_scaff_5181976	79	3289	3211
eRTBVL-TA-47	2BS_scaff_5225781	7718	10928	3211
eRTBVL-TA-48	2BS_scaff_5237591	9203	13297	4095
eRTBVL-TA-49	2DL_scaff_9894750	9077	10627	1551
eRTBVL-TA-50	2DL_scaff_9910878	8753	8981	229
eRTBVL-TA-51	2DS_scaff_5326131	84	1251	1168
eRTBVL-TA-52	2DS_scaff_5330128	1	588	588
eRTBVL-TA-53	2DS_scaff_5350818	160	3125	2966
eRTBVL-TA-54	2DS_scaff_5388381	11152	11356	205
eRTBVL-TA-55	2DS_scaff_5389858	1	2181	2181
eRTBVL-TA-56	2DS_scaff_5390929	1	2538	2538
eRTBVL-TA-57	3B_scaff_10395021	1	603	603
eRTBVL-TA-58	3B_scaff_10395799	1072	4229	3158
eRTBVL-TA-59	3B_scaff_10427489	463	3518	3056
eRTBVL-TA-60	3B_scaff_10474566	1	1845	1845
eRTBVL-TA-61	3B_scaff_10526135	21911	22326	416
eRTBVL-TA-62	3B_scaff_10553629	1	1382	1382
eRTBVL-TA-63	3B_scaff_10555832	1	3720	3720
eRTBVL-TA-64	3B_scaff_10555833	1	4411	4411
eRTBVL-TA-65	3B_scaff_10571546	2162	4940	2779
eRTBVL-TA-66	3B_scaff_10630485	1	471	471
eRTBVL-TA-67	3B_scaff_10645606	156	6343	6188
eRTBVL-TA-68	3B_scaff_10677798	594	3649	3056
eRTBVL-TA-69	3B_scaff_10697179	13474	15617	2144
eRTBVL-TA-70	3B_scaff_10717329	337	7445	7109
eRTBVL-TA-71	3B_scaff_10746420	1	1254	1254
eRTBVL-TA-72	3B_scaff_10746783	2	1167	1166
eRTBVL-TA-73	3B_scaff_10763533	1	9224	9224
eRTBVL-TA-74	3B_scaff_10766284	203	708	506
eRTBVL-TA-75	3DL_scaff_6870230	161	2790	2630
eRTBVL-TA-76	3DL_scaff_6906354	284	2444	2161
eRTBVL-TA-77	3DL_scaff_6921789	7797	9780	1984
eRTBVL-TA-78	3DL_scaff_6932306	1	3470	3470
eRTBVL-TA-79	3DL_scaff_6932483	1	4038	4038
eRTBVL-TA-80	3DS_scaff_2569391	105	1742	1638
eRTBVL-TA-81	3DS_scaff_2603727	1	5013	5013
eRTBVL-TA-82	4AL_scaff_4AL_5666118	84	416	333
eRTBVL-TA-83	4AL_scaff_4AL_7100162	8097	8785	689
eRTBVL-TA-84	4AL_scaff_4AL_7120684	1	1489	1489
eRTBVL-TA-85	4AL_scaff_4AL_7131610	1	3450	3450
eRTBVL-TA-86	4AL_scaff_4AL_7176891	3693	6994	3302
eRTBVL-TA-87	4AS_scaff_5946822	3162	3657	496
eRTBVL-TA-88	4AS_scaff_5953036	1	1130	1130
eRTBVL-TA-89	4AS_scaff_5956990	1073	2413	1341
eRTBVL-TA-90	4AS_scaff_5965925	3491	4150	660
eRTBVL-TA-91	4AS_scaff_5969875	36	3076	3041
eRTBVL-TA-92	4AS_scaff_5975120	4076	4606	531
eRTBVL-TA-93	4AS_scaff_5998147	3667	4147	481
eRTBVL-TA-94	4BL_scaff_6918377	32	6722	6691
eRTBVL-TA-95	4BL_scaff_6974146	1	4646	4646
eRTBVL-TA-96	4BL_scaff_7039881	1	1512	1512
eRTBVL-TA-97	4BS_scaff_4886175	9722	10238	517
eRTBVL-TA-98	4BS_scaff_4886175	16698	19836	3139
eRTBVL-TA-99	4BS_scaff_4915102	47	1212	1166
eRTBVL-TA-100	4BS_scaff_4925033	3991	5212	1222
eRTBVL-TA-101	4BS_scaff_4934774	514	3705	3192
eRTBVL-TA-102	4BS_scaff_4940770	19	3709	3691
eRTBVL-TA-103	4BS_scaff_4946293	577	714	138
eRTBVL-TA-104	4BS_scaff_4946293	7145	9131	1987
eRTBVL-TA-105	4DL_scaff_4DL_14341313	2947	3285	339
eRTBVL-TA-106	4DL_scaff_4DL_14376464	3480	9153	5674
eRTBVL-TA-107	4DL_scaff_4DL_14405899	4586	5067	482

eRTBVL-TA-108	4DS_scaff_2282630	295	7433	7139
eRTBVL-TA-109	4DS_scaff_2284023	2	5297	5296
eRTBVL-TA-110	5AL_scaff_2736241	1	1473	1473
eRTBVL-TA-111	5AL_scaff_2740021	2889	5305	2417
eRTBVL-TA-112	5AL_scaff_2740085	5359	5960	602
eRTBVL-TA-113	5AL_scaff_2759664	1	2685	2685
eRTBVL-TA-114	5AL_scaff_2769653	1744	2166	423
eRTBVL-TA-115	5AL_scaff_2781274	1	2644	2644
eRTBVL-TA-116	5AS_scaff_1530548	1	1681	1681
eRTBVL-TA-117	5AS_scaff_1537869	1	1652	1652
eRTBVL-TA-118	5AS_scaff_1538070	1	1652	1652
eRTBVL-TA-119	5BL_scaff_10708607	4319	4481	163
eRTBVL-TA-120	5BL_scaff_10799490	2889	9220	6332
eRTBVL-TA-121	5BL_scaff_10920697	15587	15766	180
eRTBVL-TA-122	5DL_scaff_4474632	303	832	530
eRTBVL-TA-123	5DL_scaff_4516533	8831	10813	1983
eRTBVL-TA-124	5DL_scaff_4516533	13949	15931	1983
eRTBVL-TA-125	5DL_scaff_4541692	1348	4692	3345
eRTBVL-TA-126	5DL_scaff_4575265	176	480	305
eRTBVL-TA-127	5DL_scaff_4605071	1	8718	8718
eRTBVL-TA-128	5DL_scaff_4605259	2257	3234	978
eRTBVL-TA-129	6AL_scaff_5779628	8204	9324	1121
eRTBVL-TA-130	6AL_scaff_5784649	751	2473	1723
eRTBVL-TA-131	6AL_scaff_5807957	1	1123	1123
eRTBVL-TA-132	6AL_scaff_5807958	1	6662	6662
eRTBVL-TA-133	6AL_scaff_5824356	2204	5652	3449
eRTBVL-TA-134	6AL_scaff_5828918	6434	10370	3937
eRTBVL-TA-135	6AS_scaff_4423820	1967	2700	734
eRTBVL-TA-136	6BL_scaff_4213837	2	2011	2010
eRTBVL-TA-137	6BL_scaff_4267701	547	891	345
eRTBVL-TA-138	6BL_scaff_4325176	2862	2984	123
eRTBVL-TA-139	6BL_scaff_4357386	1617	14733	13117
eRTBVL-TA-140	6BS_scaff_2953457	2272	5744	3473
eRTBVL-TA-141	6BS_scaff_2976559	358	1531	1174
eRTBVL-TA-142	6BS_scaff_3002818	812	1217	406
eRTBVL-TA-143	6BS_scaff_3046709	123	398	276
eRTBVL-TA-144	6DL_scaff_3231390	3236	3377	142
eRTBVL-TA-145	6DL_scaff_3258205	1	3000	3000
eRTBVL-TA-146	6DL_scaff_3268251	3732	10329	6598
eRTBVL-TA-147	6DL_scaff_3275891	2212	5273	3062
eRTBVL-TA-148	6DL_scaff_3277197	884	3819	2936
eRTBVL-TA-149	6DL_scaff_3307942	3482	4386	905
eRTBVL-TA-150	6DL_scaff_3324873	1	2016	2016
eRTBVL-TA-151	6DL_scaff_3327379	4876	5209	334
eRTBVL-TA-152	6DL_scaff_3329177	1335	4172	2838
eRTBVL-TA-153	6DS_scaff_2077618	1	3060	3060
eRTBVL-TA-154	6DS_scaff_2107914	50	3342	3293
eRTBVL-TA-155	6DS_scaff_2115750	1	3048	3048
eRTBVL-TA-156	6DS_scaff_2118132	1	2296	2296
eRTBVL-TA-157	6DS_scaff_2122153	1	4794	4794
eRTBVL-TA-158	6DS_scaff_2122196	1	3143	3143
eRTBVL-TA-159	7AL_scaff_4400098	364	703	340
eRTBVL-TA-160	7AL_scaff_4460486	498	3062	2565
eRTBVL-TA-161	7AL_scaff_4483041	1029	1420	392
eRTBVL-TA-162	7AL_scaff_4494924	4303	5451	1149
eRTBVL-TA-163	7AL_scaff_4503354	528	2001	1474
eRTBVL-TA-164	7AL_scaff_4530934	5065	5456	392
eRTBVL-TA-165	7AL_scaff_4535207	2209	2600	392
eRTBVL-TA-166	7AL_scaff_4541956	533	2310	1778
eRTBVL-TA-167	7AS_scaff_4049612	1	1167	1167
eRTBVL-TA-168	7AS_scaff_4171573	1556	2290	735
eRTBVL-TA-169	7AS_scaff_4171789	1	422	422
eRTBVL-TA-170	7AS_scaff_4240839	799	1533	735
eRTBVL-TA-171	7AS_scaff_4246685	6930	8718	1789
eRTBVL-TA-172	7AS_scaff_4254600	866	970	105
eRTBVL-TA-173	7AS_scaff_4254600	5047	5466	420
eRTBVL-TA-174	7AS_scaff_4255547	1	1732	1732
eRTBVL-TA-175	7BL_scaff_6642097	7	2936	2930
eRTBVL-TA-176	7BL_scaff_6643087	1	801	801
eRTBVL-TA-177	7BL_scaff_6645611	1376	4362	2987
eRTBVL-TA-178	7BL_scaff_6664637	1	1112	1112
eRTBVL-TA-179	7BL_scaff_6664638	1	1112	1112
eRTBVL-TA-180	7BL_scaff_6670602	14	1615	1602
eRTBVL-TA-181	7BL_scaff_6670603	14	1615	1602
eRTBVL-TA-182	7BL_scaff_6674241	1252	3745	2494
eRTBVL-TA-183	7BL_scaff_6689826	1	1747	1747
eRTBVL-TA-184	7BL_scaff_6695292	3060	3182	123
eRTBVL-TA-185	7BL_scaff_6716206	1038	1373	336
eRTBVL-TA-186	7BL_scaff_6742726	13	1781	1769
eRTBVL-TA-187	7BL_scaff_6743065	603	2868	2266
eRTBVL-TA-188	7BL_scaff_6749115	867	3360	2494
eRTBVL-TA-189	7BL_scaff_6751184	537	705	169
eRTBVL-TA-190	7BS_scaff_3069253	2	2235	2234

	eRTBVL-TA-191	7BS_scaff_3069255	1	2940	2940
	eRTBVL-TA-192	7BS_scaff_3072851	1	2516	2516
	eRTBVL-TA-193	7BS_scaff_3114835	265	4874	4610
	eRTBVL-TA-194	7BS_scaff_3126977	7	3083	3077
	eRTBVL-TA-195	7BS_scaff_3139984	26	2160	2135
	eRTBVL-TA-196	7BS_scaff_3163673	1	3174	3174
	eRTBVL-TA-197	7BS_scaff_3163676	1	1389	1389
	eRTBVL-TA-198	7DL_scaff_3330575	35	3143	3109
	eRTBVL-TA-199	7DL_scaff_3335015	1671	4029	2359
	eRTBVL-TA-200	7DL_scaff_3337743	1417	2603	1187
	eRTBVL-TA-201	7DL_scaff_3337743	4623	6638	2016
	eRTBVL-TA-202	7DL_scaff_3392298	1746	3526	1781
	eRTBVL-TA-203	7DL_scaff_3395450	6163	6472	310
	eRTBVL-TA-204	7DL_scaff_3395711	1	1224	1224
	eRTBVL-TA-205	7DS_scaff_3812497	4270	7337	3068
	eRTBVL-TA-206	7DS_scaff_3862876	12528	13636	1109
	eRTBVL-TA-207	7DS_scaff_3904759	5209	5436	228
	eRTBVL-TA-208	7DS_scaff_3970029	1800	2340	541
<i>Brachypodium distachyon</i>	eRTBVL-BD-1	1	2291143	2296479	5337
	eRTBVL-BD-2	1	4059052	4060332	1281
	eRTBVL-BD-3	1	6124432	6124807	376
	eRTBVL-BD-4	1	7106447	7106677	231
	eRTBVL-BD-5	1	8092663	8093911	1249
	eRTBVL-BD-6	1	11259829	11261072	1244
	eRTBVL-BD-7	1	17277872	17283990	6119
	eRTBVL-BD-8	1	27603495	27612790	9296
	eRTBVL-BD-9	1	29773788	29775246	1459
	eRTBVL-BD-10	1	34667131	34668304	1174
	eRTBVL-BD-11	1	41501865	41503061	1197
	eRTBVL-BD-12	1	43205209	43206461	1253
	eRTBVL-BD-13	1	47190454	47199714	9261
	eRTBVL-BD-14	1	47289582	47301653	12072
	eRTBVL-BD-15	1	50066390	50066990	601
	eRTBVL-BD-16	1	52982798	52982940	143
	eRTBVL-BD-17	1	54968210	54977377	9168
	eRTBVL-BD-18	1	55054752	55063789	9038
	eRTBVL-BD-19	1	55542126	55543301	1176
	eRTBVL-BD-20	1	56369439	56369845	407
	eRTBVL-BD-21	1	57216538	57218173	1636
	eRTBVL-BD-22	1	68423862	68434499	10638
	eRTBVL-BD-23	1	69635047	69642691	7645
	eRTBVL-BD-24	1	74117578	74118705	1128
	eRTBVL-BD-25	2	900071	908766	8696
	eRTBVL-BD-26	2	2627435	2628575	1141
	eRTBVL-BD-27	2	10668790	10669460	671
	eRTBVL-BD-28	2	14287219	14288410	1192
	eRTBVL-BD-29	2	18289756	18290981	1226
	eRTBVL-BD-30	2	21391818	21393632	1815
	eRTBVL-BD-31	2	21636201	21638844	2644
	eRTBVL-BD-32	2	23364027	23365043	1017
	eRTBVL-BD-33	2	32734627	32736317	1691
	eRTBVL-BD-34	2	32763102	32764955	1854
	eRTBVL-BD-35	2	32780845	32785814	4970
	eRTBVL-BD-36	2	56929049	56929458	410
	eRTBVL-BD-37	3	11769868	11770561	694
	eRTBVL-BD-38	3	15277493	15281519	4027
	eRTBVL-BD-39	3	15289734	15292850	3117
	eRTBVL-BD-40	3	27042169	27043279	1111
	eRTBVL-BD-41	3	29375585	29376076	492
	eRTBVL-BD-42	3	30035566	30036594	1029
	eRTBVL-BD-43	3	44673886	44674800	915
	eRTBVL-BD-44	3	46249384	46249641	258
	eRTBVL-BD-45	3	54914757	54920742	5986
	eRTBVL-BD-46	3	55995033	55996209	1177
	eRTBVL-BD-47	4	265673	265920	248
	eRTBVL-BD-48	4	1658524	1665177	6654
	eRTBVL-BD-49	4	7270233	7271567	1335
	eRTBVL-BD-50	4	8718145	8719320	1176
	eRTBVL-BD-51	4	11725660	11728393	2734
	eRTBVL-BD-52	4	12921146	12921314	169
	eRTBVL-BD-53	4	19630830	19639381	8552
	eRTBVL-BD-54	4	20239277	20242153	2877
	eRTBVL-BD-55	4	20279412	20281993	2582
	eRTBVL-BD-56	4	21278370	21287378	9009
	eRTBVL-BD-57	4	23919949	23921702	1754
	eRTBVL-BD-58	4	26950205	26959518	9314
	eRTBVL-BD-59	4	27280663	27283471	2809
	eRTBVL-BD-60	4	29747894	29749975	2082
	eRTBVL-BD-61	4	30069477	30069650	174
	eRTBVL-BD-62	4	31068028	31070381	2354
	eRTBVL-BD-63	4	31127810	31133788	5979
	eRTBVL-BD-64	4	31138414	31143428	5015
	eRTBVL-BD-65	4	31175791	31183855	8065

	eRTBVL-BD-66	4	32522911	32531454	8544
	eRTBVL-BD-67	4	33507260	33508494	1235
	eRTBVL-BD-68	4	39610430	39618419	7990
	eRTBVL-BD-69	4	42300821	42302171	1351
	eRTBVL-BD-70	5	11204459	11217376	12918
	eRTBVL-BD-71	5	11252114	11259597	7484
	eRTBVL-BD-72	5	19086453	19087026	574
	eRTBVL-BD-73	5	19089496	19095702	6207
	eRTBVL-BD-74	scaffold_26	14913	15140	228
<i>Phyllostachys heterocycla</i>	eRTBVL-PH-1	PH01000002	27338	27642	305
	eRTBVL-PH-2	PH01000011	1271596	1275153	3558
	eRTBVL-PH-3	PH01000012	987428	987851	424
	eRTBVL-PH-4	PH01000012	990010	994966	4957
	eRTBVL-PH-5	PH01000012	1001323	1001809	487
	eRTBVL-PH-6	PH01000023	312299	318406	6108
	eRTBVL-PH-7	PH01000026	1127989	1131857	3869
	eRTBVL-PH-8	PH01000045	674987	676978	1992
	eRTBVL-PH-9	PH01000070	1073963	1079325	5363
	eRTBVL-PH-10	PH01000088	267331	268218	888
	eRTBVL-PH-11	PH01000097	541664	543742	2079
	eRTBVL-PH-12	PH01000097	565483	569315	3833
	eRTBVL-PH-13	PH01000121	286351	287614	1264
	eRTBVL-PH-14	PH01000134	556291	561037	4747
	eRTBVL-PH-15	PH01000154	92379	93298	920
	eRTBVL-PH-16	PH01000154	96239	102065	5827
	eRTBVL-PH-17	PH01000154	109839	113103	3265
	eRTBVL-PH-18	PH01000162	534311	536252	1942
	eRTBVL-PH-19	PH01000162	543643	544575	933
	eRTBVL-PH-20	PH01000162	553223	558247	5025
	eRTBVL-PH-21	PH01000173	774646	781811	7166
	eRTBVL-PH-22	PH01000183	213366	220257	6892
	eRTBVL-PH-23	PH01000196	253975	254477	503
	eRTBVL-PH-24	PH01000196	265063	265581	519
	eRTBVL-PH-25	PH01000207	705423	705944	522
	eRTBVL-PH-26	PH01000207	718301	720477	2177
	eRTBVL-PH-27	PH01000207	722846	724189	1344
	eRTBVL-PH-28	PH01000207	726759	728450	1692
	eRTBVL-PH-29	PH01000212	172095	172383	289
	eRTBVL-PH-30	PH01000212	215838	216110	273
	eRTBVL-PH-31	PH01000246	477164	483499	6336
	eRTBVL-PH-32	PH01000261	411413	413294	1882
	eRTBVL-PH-33	PH01000275	409051	409224	174
	eRTBVL-PH-34	PH01000282	252706	253257	552
	eRTBVL-PH-35	PH01000282	278731	291779	13049
	eRTBVL-PH-36	PH01000282	302101	302589	489
	eRTBVL-PH-37	PH01000282	319012	321203	2192
	eRTBVL-PH-38	PH01000290	634820	635109	290
	eRTBVL-PH-39	PH01000306	74631	80653	6023
	eRTBVL-PH-40	PH01000328	197866	201764	3899
	eRTBVL-PH-41	PH01000328	210930	212973	2044
	eRTBVL-PH-42	PH01000344	701139	707940	6802
	eRTBVL-PH-43	PH01000358	465333	471932	6600
	eRTBVL-PH-44	PH01000407	589261	594716	5456
	eRTBVL-PH-45	PH01000436	396014	398875	2862
	eRTBVL-PH-46	PH01000445	611639	615528	3890
	eRTBVL-PH-47	PH01000486	625916	627579	1664
	eRTBVL-PH-48	PH01000514	364755	368248	3494
	eRTBVL-PH-49	PH01000520	417342	418499	1158
	eRTBVL-PH-50	PH01000520	420885	423532	2648
	eRTBVL-PH-51	PH01000520	437019	438106	1088
	eRTBVL-PH-52	PH01000524	140432	140660	229
	eRTBVL-PH-53	PH01000568	441435	443768	2334
	eRTBVL-PH-54	PH01000568	447071	452004	4934
	eRTBVL-PH-55	PH01000572	270090	271806	1717
	eRTBVL-PH-56	PH01000572	292734	299871	7138
	eRTBVL-PH-57	PH01000578	279877	287077	7201
	eRTBVL-PH-58	PH01000602	76297	77536	1240
	eRTBVL-PH-59	PH01000606	969	4677	3709
	eRTBVL-PH-60	PH01000606	6872	10123	3252
	eRTBVL-PH-61	PH01000609	210923	214080	3158
	eRTBVL-PH-62	PH01000609	362758	373631	10874
	eRTBVL-PH-63	PH01000610	207642	213409	5768
	eRTBVL-PH-64	PH01000610	215945	218549	2605
	eRTBVL-PH-65	PH01000625	276642	278977	2336
	eRTBVL-PH-66	PH01000636	363674	364725	1052
	eRTBVL-PH-67	PH01000636	366873	372060	5188
	eRTBVL-PH-68	PH01000660	533697	535972	2276
	eRTBVL-PH-69	PH01000724	2551	3133	583
	eRTBVL-PH-70	PH01000828	313158	313698	541
	eRTBVL-PH-71	PH01000828	325177	332337	7161
	eRTBVL-PH-72	PH01000840	135357	140952	5596
	eRTBVL-PH-73	PH01000904	404497	404644	148
	eRTBVL-PH-74	PH01000924	429802	431139	1338

eRTBVL-PH-75	PH01000958	61128	67891	6764
eRTBVL-PH-76	PH01000970	276414	282299	5886
eRTBVL-PH-77	PH01001008	71361	72442	1082
eRTBVL-PH-78	PH01001008	94550	97672	3123
eRTBVL-PH-79	PH01001008	100691	101506	816
eRTBVL-PH-80	PH01001008	106227	106707	481
eRTBVL-PH-81	PH01001054	386652	387681	1030
eRTBVL-PH-82	PH01001084	226646	234193	7548
eRTBVL-PH-83	PH01001117	375238	375350	113
eRTBVL-PH-84	PH01001124	207800	214829	7030
eRTBVL-PH-85	PH01001212	130994	135851	4858
eRTBVL-PH-86	PH01001268	27847	28063	217
eRTBVL-PH-87	PH01001273	63574	64173	600
eRTBVL-PH-88	PH01001411	282438	283693	1256
eRTBVL-PH-89	PH01001411	290083	292267	2185
eRTBVL-PH-90	PH01001411	301774	301921	148
eRTBVL-PH-91	PH01001424	216380	223202	6823
eRTBVL-PH-92	PH01001431	219670	221216	1547
eRTBVL-PH-93	PH01001431	225045	234638	9594
eRTBVL-PH-94	PH01001431	239311	244403	5093
eRTBVL-PH-95	PH01001465	341821	343954	2134
eRTBVL-PH-96	PH01001478	255744	262042	6299
eRTBVL-PH-97	PH01001479	106353	108524	2172
eRTBVL-PH-98	PH01001479	111326	112059	734
eRTBVL-PH-99	PH01001512	68893	74835	5943
eRTBVL-PH-100	PH01001589	156134	156256	123
eRTBVL-PH-101	PH01001589	170502	177360	6859
eRTBVL-PH-102	PH01001662	224801	225302	502
eRTBVL-PH-103	PH01001663	319050	322903	3854
eRTBVL-PH-104	PH01001712	302758	303004	247
eRTBVL-PH-105	PH01002027	255078	259397	4320
eRTBVL-PH-106	PH01002050	199502	200176	675
eRTBVL-PH-107	PH01002091	126484	129979	3496
eRTBVL-PH-108	PH01002191	41789	43985	2197
eRTBVL-PH-109	PH01002191	48037	49273	1237
eRTBVL-PH-110	PH01002199	174224	174699	476
eRTBVL-PH-111	PH01002246	91809	94736	2928
eRTBVL-PH-112	PH01002246	116645	122396	5752
eRTBVL-PH-113	PH01002319	140378	146156	5779
eRTBVL-PH-114	PH01002319	159080	163989	4910
eRTBVL-PH-115	PH01002339	104813	107363	2551
eRTBVL-PH-116	PH01002339	109622	110073	452
eRTBVL-PH-117	PH01002356	83086	85497	2412
eRTBVL-PH-118	PH01002356	90771	91002	232
eRTBVL-PH-119	PH01002356	99455	100655	1201
eRTBVL-PH-120	PH01002513	76837	80040	3204
eRTBVL-PH-121	PH01002617	25547	29001	3455
eRTBVL-PH-122	PH01002633	160753	162609	1857
eRTBVL-PH-123	PH01002633	167091	167999	909
eRTBVL-PH-124	PH01002633	176882	178484	1603
eRTBVL-PH-125	PH01002647	533	1795	1263
eRTBVL-PH-126	PH01002647	15440	15657	218
eRTBVL-PH-127	PH01002655	190749	194627	3879
eRTBVL-PH-128	PH01003017	119217	119899	683
eRTBVL-PH-129	PH01003024	63776	65599	1824
eRTBVL-PH-130	PH01003032	59418	60013	596
eRTBVL-PH-131	PH01003032	62813	66594	3782
eRTBVL-PH-132	PH01003032	79133	80192	1060
eRTBVL-PH-133	PH01003063	71972	82135	10164
eRTBVL-PH-134	PH01003067	146035	148845	2811
eRTBVL-PH-135	PH01003159	101853	105453	3601
eRTBVL-PH-136	PH01003284	127585	134060	6476
eRTBVL-PH-137	PH01003327	100565	102896	2332
eRTBVL-PH-138	PH01003327	110055	110569	515
eRTBVL-PH-139	PH01003359	34215	39491	5277
eRTBVL-PH-140	PH01003359	42260	42736	477
eRTBVL-PH-141	PH01003479	100340	103408	3069
eRTBVL-PH-142	PH01003553	115828	117840	2013
eRTBVL-PH-143	PH01003553	120453	121323	871
eRTBVL-PH-144	PH01003553	129773	130409	637
eRTBVL-PH-145	PH01003813	48971	54813	5843
eRTBVL-PH-146	PH01003844	82551	87484	4934
eRTBVL-PH-147	PH01003857	53393	54914	1522
eRTBVL-PH-148	PH01004062	303	1535	1233
eRTBVL-PH-149	PH01004323	41572	41694	123
eRTBVL-PH-150	PH01004359	811	1081	271
eRTBVL-PH-151	PH01004359	7935	8942	1008
eRTBVL-PH-152	PH01004359	17756	19143	1388
eRTBVL-PH-153	PH01004359	26162	26272	111
eRTBVL-PH-154	PH01004620	65818	66598	781
eRTBVL-PH-155	PH01004707	6701	10526	3826
eRTBVL-PH-156	PH01004867	75661	79732	4072
eRTBVL-PH-157	PH01005366	61147	65488	4342

eRTBVL-PH-158	PH01005432	6740	7724	985
eRTBVL-PH-159	PH01005672	20142	21650	1509
eRTBVL-PH-160	PH01005672	24648	25596	949
eRTBVL-PH-161	PH01005672	40889	46559	5671
eRTBVL-PH-162	PH01005890	45244	46674	1431
eRTBVL-PH-163	PH01006418	33212	40087	6876
eRTBVL-PH-164	PH01006744	9453	16021	6569
eRTBVL-PH-165	PH01007888	1061	1804	744
eRTBVL-PH-166	PH01008208	15	12939	12925
eRTBVL-PH-167	PH01008268	1467	7093	5627
eRTBVL-PH-168	PH01008268	9972	11043	1072
eRTBVL-PH-169	PH01008982	10672	11468	797
eRTBVL-PH-170	PH01009591	1	1541	1541
eRTBVL-PH-171	PH01009851	17	575	559
eRTBVL-PH-172	PH01009851	7307	7843	537
eRTBVL-PH-173	PH01011055	3131	4146	1016
eRTBVL-PH-174	PH01011098	86	1312	1227
eRTBVL-PH-175	PH01011187	449	564	116
eRTBVL-PH-176	PH01011187	3411	4500	1090
eRTBVL-PH-177	PH01012023	2308	2957	650
eRTBVL-PH-178	PH01012529	1	3924	3924
eRTBVL-PH-179	PH01014443	1	3685	3685
eRTBVL-PH-180	PH01016676	1567	3179	1613
eRTBVL-PH-181	PH01016837	2	3095	3094
eRTBVL-PH-182	PH01020365	1815	2713	899
eRTBVL-PH-183	PH01021061	1	2647	2647
eRTBVL-PH-184	PH01022893	48	2501	2454
eRTBVL-PH-185	PH01022993	1	2503	2503
eRTBVL-PH-186	PH01025308	1	2351	2351
eRTBVL-PH-187	PH01025842	33	2320	2288
eRTBVL-PH-188	PH01025855	512	2318	1807
eRTBVL-PH-189	PH01027997	1228	2221	994
eRTBVL-PH-190	PH01030556	444	2113	1670
eRTBVL-PH-191	PH01032164	1402	2048	647
eRTBVL-PH-192	PH01032850	3	2023	2021
eRTBVL-PH-193	PH01034952	1	1948	1948
eRTBVL-PH-194	PH01035614	1	1927	1927
eRTBVL-PH-195	PH01036292	439	1910	1472
eRTBVL-PH-196	PH01037176	1	1881	1881
eRTBVL-PH-197	PH01037495	1	1864	1864
eRTBVL-PH-198	PH01037497	387	1879	1493
eRTBVL-PH-199	PH01038909	735	1849	1115
eRTBVL-PH-200	PH01039094	11	613	603
eRTBVL-PH-201	PH01039277	1	1840	1840
eRTBVL-PH-202	PH01040483	1459	1566	108
eRTBVL-PH-203	PH01044085	170	1740	1571
eRTBVL-PH-204	PH01046005	1	1114	1114
eRTBVL-PH-205	PH01046520	1	1672	1672
eRTBVL-PH-206	PH01047955	2	1617	1616
eRTBVL-PH-207	PH01049990	154	1597	1444
eRTBVL-PH-208	PH01053522	1	1527	1527
eRTBVL-PH-209	PH01055665	1	1501	1501
eRTBVL-PH-210	PH01058434	1	1471	1471
eRTBVL-PH-211	PH01059098	1	1465	1465
eRTBVL-PH-212	PH01060754	1	1411	1411
eRTBVL-PH-213	PH01064782	1	1419	1419
eRTBVL-PH-214	PH01067026	5	1402	1398
eRTBVL-PH-215	PH01068763	10	1387	1378
eRTBVL-PH-216	PH01069113	1	1384	1384
eRTBVL-PH-217	PH01072969	6	1343	1338
eRTBVL-PH-218	PH01073923	557	1204	648
eRTBVL-PH-219	PH01074201	975	1285	311
eRTBVL-PH-220	PH01075074	3	1324	1322
eRTBVL-PH-221	PH01075078	1	1324	1324
eRTBVL-PH-222	PH01076729	2	1304	1303
eRTBVL-PH-223	PH01078782	1	1275	1275
eRTBVL-PH-224	PH01083945	397	1156	760
eRTBVL-PH-225	PH01084393	1	1127	1127
eRTBVL-PH-226	PH01087974	3	1138	1136
eRTBVL-PH-227	PH01095844	57	1077	1021
eRTBVL-PH-228	PH01098023	4	1066	1063
eRTBVL-PH-229	PH01098264	3	1064	1062
eRTBVL-PH-230	PH01098561	22	1061	1040
eRTBVL-PH-231	PH01099698	5	1060	1056
eRTBVL-PH-232	PH01104081	1	1028	1028
eRTBVL-PH-233	PH01107520	4	1038	1035
eRTBVL-PH-234	PH01109105	2	1032	1031
eRTBVL-PH-235	PH01110721	1	1031	1031
eRTBVL-PH-236	PH01111224	441	1030	590
eRTBVL-PH-237	PH01111672	412	1029	618
eRTBVL-PH-238	PH01115979	1	1022	1022
eRTBVL-PH-239	PH01116125	1	1019	1019
eRTBVL-PH-240	PH01117741	1	1019	1019

eRTBVL-PH-241	PH01117822	1	1018	1018
eRTBVL-PH-242	PH01119246	1	597	597
eRTBVL-PH-243	PH01120086	1	1015	1015
eRTBVL-PH-244	PH01120327	1	1009	1009
eRTBVL-PH-245	PH01122503	1	1010	1010
eRTBVL-PH-246	PH01123016	3	1010	1008
eRTBVL-PH-247	PH01123962	3	1003	1001
eRTBVL-PH-248	PH01123984	769	1008	240
eRTBVL-PH-249	PH01124562	1	537	537
eRTBVL-PH-250	PH01124857	8	239	232
eRTBVL-PH-251	PH01125022	1	570	570
eRTBVL-PH-252	PH01125434	763	1006	244
eRTBVL-PH-253	PH01127284	1	1003	1003
eRTBVL-PH-254	PH01128567	1	1000	1000
eRTBVL-PH-255	PH01128576	1	594	594
eRTBVL-PH-256	PH01129195	1	912	912
eRTBVL-PH-257	PH01129679	2	960	959
eRTBVL-PH-258	PH01130052	1	998	998
eRTBVL-PH-259	PH01130338	1	997	997
eRTBVL-PH-260	PH01130390	1	997	997
eRTBVL-PH-261	PH01133537	1	991	991
eRTBVL-PH-262	PH01134483	1	215	215
eRTBVL-PH-263	PH01134714	1	969	969
eRTBVL-PH-264	PH01136353	36	585	550
eRTBVL-PH-265	PH01137158	1	984	984
eRTBVL-PH-266	PH01137307	2	983	982
eRTBVL-PH-267	PH01138356	27	167	141
eRTBVL-PH-268	PH01138552	1	980	980
eRTBVL-PH-269	PH01140473	1	974	974
eRTBVL-PH-270	PH01140524	1	973	973
eRTBVL-PH-271	PH01140592	3	222	220
eRTBVL-PH-272	PH01141167	5	567	563
eRTBVL-PH-273	PH01141233	1	613	613
eRTBVL-PH-274	PH01141836	9	967	959
eRTBVL-PH-275	PH01152060	97	894	798
eRTBVL-PH-276	PH01152396	6	884	879
eRTBVL-PH-277	PH01152629	665	890	226
eRTBVL-PH-278	PH01152977	3	884	882
eRTBVL-PH-279	PH01154003	53	593	541
eRTBVL-PH-280	PH01157348	2	240	239
eRTBVL-PH-281	PH01161604	1	722	722
eRTBVL-PH-282	PH01162014	6	710	705
eRTBVL-PH-283	PH01162922	595	773	179
eRTBVL-PH-284	PH01164259	622	767	146
eRTBVL-PH-285	PH01166492	81	701	621
eRTBVL-PH-286	PH01167364	115	742	628
eRTBVL-PH-287	PH01169247	1	733	733
eRTBVL-PH-288	PH01172167	1	714	714
eRTBVL-PH-289	PH01173072	1	211	211
eRTBVL-PH-290	PH01173190	1	712	712
eRTBVL-PH-291	PH01173552	1	710	710
eRTBVL-PH-292	PH01175245	109	487	379
eRTBVL-PH-293	PH01180130	289	682	394
eRTBVL-PH-294	PH01182108	1	671	671
eRTBVL-PH-295	PH01182711	3	671	669
eRTBVL-PH-296	PH01183612	3	668	666
eRTBVL-PH-297	PH01184968	1	665	665
eRTBVL-PH-298	PH01186141	12	217	206
eRTBVL-PH-299	PH01186569	1	526	526
eRTBVL-PH-300	PH01187149	1	449	449
eRTBVL-PH-301	PH01188078	1	656	656
eRTBVL-PH-302	PH01188328	1	641	641
eRTBVL-PH-303	PH01189464	1	649	649
eRTBVL-PH-304	PH01192898	14	608	595
eRTBVL-PH-305	PH01194053	2	639	638
eRTBVL-PH-306	PH01198510	1	631	631
eRTBVL-PH-307	PH01200266	1	628	628
eRTBVL-PH-308	PH01207086	1	606	606
eRTBVL-PH-309	PH01208532	364	614	251
eRTBVL-PH-310	PH01209166	1	611	611
eRTBVL-PH-311	PH01209337	41	613	573
eRTBVL-PH-312	PH01209577	1	613	613
eRTBVL-PH-313	PH01211032	1	609	609
eRTBVL-PH-314	PH01212187	357	594	238
eRTBVL-PH-315	PH01212497	1	390	390
eRTBVL-PH-316	PH01213907	22	605	584
eRTBVL-PH-317	PH01217621	355	601	247
eRTBVL-PH-318	PH01223030	1	597	597
eRTBVL-PH-319	PH01223107	383	586	204
eRTBVL-PH-320	PH01224781	1	174	174
eRTBVL-PH-321	PH01226182	3	593	591
eRTBVL-PH-322	PH01226227	3	593	591
eRTBVL-PH-323	PH01227677	1	591	591

	eRTBVL-PH-324	PH01227685	1	592	592
	eRTBVL-PH-325	PH01228232	1	590	590
	eRTBVL-PH-326	PH01229637	403	531	129
	eRTBVL-PH-327	PH01233913	373	585	213
	eRTBVL-PH-328	PH01234881	1	584	584
	eRTBVL-PH-329	PH01236073	1	582	582
	eRTBVL-PH-330	PH01236106	1	583	583
	eRTBVL-PH-331	PH01240976	1	578	578
	eRTBVL-PH-332	PH01242923	6	575	570
	eRTBVL-PH-333	PH01246584	1	338	338
	eRTBVL-PH-334	PH01247682	1	317	317
	eRTBVL-PH-335	PH01248509	1	568	568
	eRTBVL-PH-336	PH01250349	396	499	104
	eRTBVL-PH-337	PH01253428	1	157	157
	eRTBVL-PH-338	PH01254016	1	560	560
	eRTBVL-PH-339	PH01262038	1	543	543
	eRTBVL-PH-340	PH01265747	1	536	536
	eRTBVL-PH-341	PH01267151	123	527	405
	eRTBVL-PH-342	PH01268347	1	529	529
	eRTBVL-PH-343	PH01269032	54	528	475
	eRTBVL-PH-344	PH01269825	1	517	517
	eRTBVL-PH-345	PH01269973	1	390	390
	eRTBVL-PH-346	PH01271778	1	504	504
	eRTBVL-PH-347	PH01272766	4	518	515
	eRTBVL-PH-348	PH01272950	3	511	509
	eRTBVL-PH-349	PH01276772	53	505	453
	eRTBVL-PH-350	PH01276781	1	504	504
	eRTBVL-PH-351	PH01277571	73	503	431
	eRTBVL-PH-352	PH01278426	194	500	307
Zea mays	eRTBVL-ZM-1	3	132047391	132047897	507
	eRTBVL-ZM-2	4	75906075	75909183	3109
Sorghum bicolor	eRTBVL-SB-1	6	46266307	46267577	1271
	eRTBVL-SB-2	9	15103838	15104153	316
	eRTBVL-SB-3	9	52347708	52348436	729
	eRTBVL-SB-4	10	51014534	51015562	1029
	eRTBVL-SB-5	10	51038010	51038142	133
Setaria italica	eRTBVL-SI-1	scaffold_1	1488389	1489274	886
	eRTBVL-SI-2	scaffold_1	11931343	11937067	5725
	eRTBVL-SI-3	scaffold_1	11945187	11952952	7766
	eRTBVL-SI-4	scaffold_1	18881845	18889101	7257
	eRTBVL-SI-5	scaffold_1	21748554	21755671	7118
	eRTBVL-SI-6	scaffold_1	21867588	21867937	350
	eRTBVL-SI-7	scaffold_1	27753587	27758663	5077
	eRTBVL-SI-8	scaffold_2	4146551	4149142	2592
	eRTBVL-SI-9	scaffold_2	6757570	6761421	3852
	eRTBVL-SI-10	scaffold_2	12898466	12903594	5129
	eRTBVL-SI-11	scaffold_2	21042144	21050009	7866
	eRTBVL-SI-12	scaffold_2	21055760	21057985	2226
	eRTBVL-SI-13	scaffold_2	21071309	21082739	11431
	eRTBVL-SI-14	scaffold_2	21100228	21103961	3734
	eRTBVL-SI-15	scaffold_2	23257005	23257923	919
	eRTBVL-SI-16	scaffold_2	25418503	25419669	1167
	eRTBVL-SI-17	scaffold_2	28317586	28322793	5208
	eRTBVL-SI-18	scaffold_2	32131590	32132741	1152
	eRTBVL-SI-19	scaffold_3	9045962	9047735	1774
	eRTBVL-SI-20	scaffold_3	24275034	24276207	1174
	eRTBVL-SI-21	scaffold_3	26839819	26840554	736
	eRTBVL-SI-22	scaffold_3	28091697	28093554	1858
	eRTBVL-SI-23	scaffold_3	34213643	34214837	1195
	eRTBVL-SI-24	scaffold_3	35343502	35349100	5599
	eRTBVL-SI-25	scaffold_3	36017967	36018966	1000
	eRTBVL-SI-26	scaffold_3	36022056	36025980	3925
	eRTBVL-SI-27	scaffold_3	36425880	36439102	13223
	eRTBVL-SI-28	scaffold_3	38345942	38347411	1470
	eRTBVL-SI-29	scaffold_3	42175067	42182710	7644
	eRTBVL-SI-30	scaffold_3	46075260	46075365	106
	eRTBVL-SI-31	scaffold_3	48833387	48835752	2366
	eRTBVL-SI-32	scaffold_4	5633544	5637819	4276
	eRTBVL-SI-33	scaffold_4	29209855	29210462	608
	eRTBVL-SI-34	scaffold_4	30740829	30741790	962
	eRTBVL-SI-35	scaffold_4	34797608	34804006	6399
	eRTBVL-SI-36	scaffold_4	34991318	34991777	460
	eRTBVL-SI-37	scaffold_4	34996396	35015957	19562
	eRTBVL-SI-38	scaffold_5	10067305	10068497	1193
	eRTBVL-SI-39	scaffold_5	14168950	14174654	5705
	eRTBVL-SI-40	scaffold_5	14608048	14610067	2020
	eRTBVL-SI-41	scaffold_5	15458608	15463014	4407
	eRTBVL-SI-42	scaffold_5	15487943	15489143	1201
	eRTBVL-SI-43	scaffold_5	17546226	17552616	6391
	eRTBVL-SI-44	scaffold_5	17555778	17561010	5233
	eRTBVL-SI-45	scaffold_5	23753306	23758118	4813
	eRTBVL-SI-46	scaffold_5	23762620	23762963	344
	eRTBVL-SI-47	scaffold_5	24047648	24048752	1105

	eRTBVL-SI-48	scaffold_5	26779896	26781567	1672
	eRTBVL-SI-49	scaffold_5	37695159	37700942	5784
	eRTBVL-SI-50	scaffold_5	37704117	37704506	390
	eRTBVL-SI-51	scaffold_5	47075068	47077321	2254
	eRTBVL-SI-52	scaffold_5	47095676	47097240	1565
	eRTBVL-SI-53	scaffold_6	9459339	9460462	1124
	eRTBVL-SI-54	scaffold_6	15362581	15384265	21685
	eRTBVL-SI-55	scaffold_6	22663873	22664276	404
	eRTBVL-SI-56	scaffold_6	27977796	27982856	5061
	eRTBVL-SI-57	scaffold_7	7683966	7686646	2681
	eRTBVL-SI-58	scaffold_7	16297902	16301569	3668
	eRTBVL-SI-59	scaffold_7	16677633	16686742	9110
	eRTBVL-SI-60	scaffold_7	27019336	27023871	4536
	eRTBVL-SI-61	scaffold_8	10475357	10497838	22482
	eRTBVL-SI-62	scaffold_8	15650290	15651568	1279
	eRTBVL-SI-63	scaffold_8	17705138	17706546	1409
	eRTBVL-SI-64	scaffold_8	19806344	19806964	621
	eRTBVL-SI-65	scaffold_8	21528036	21534070	6035
	eRTBVL-SI-66	scaffold_8	24675260	24675656	397
	eRTBVL-SI-67	scaffold_8	36309541	36309649	109
	eRTBVL-SI-68	scaffold_8	37669309	37669805	497
	eRTBVL-SI-69	scaffold_8	38219448	38224506	5059
	eRTBVL-SI-70	scaffold_9	12901998	12902126	129
	eRTBVL-SI-71	scaffold_9	15218263	15219852	1590
	eRTBVL-SI-72	scaffold_9	15227118	15229087	1970
	eRTBVL-SI-73	scaffold_9	22490237	22490378	142
	eRTBVL-SI-74	scaffold_9	25287890	25293124	5235
	eRTBVL-SI-75	scaffold_9	41378026	41379227	1202
	eRTBVL-SI-76	scaffold_9	43637279	43638063	785
	eRTBVL-SI-77	scaffold_154	84	6194	6111
	eRTBVL-SI-78	scaffold_175	6629	7346	718
	eRTBVL-SI-79	scaffold_404	88	3403	3316
	eRTBVL-SI-80	scaffold_475	33	2787	2755
<i>Panicum virgatum</i>	eRTBVL-PV-1	1a	22221883	22230114	8232
	eRTBVL-PV-2	1a	37808837	37809239	403
	eRTBVL-PV-3	1a	51578362	51582862	4501
	eRTBVL-PV-4	1a	58069448	58071965	2518
	eRTBVL-PV-5	1a	58281848	58284593	2746
	eRTBVL-PV-6	1b	176940	178078	1139
	eRTBVL-PV-7	1b	304259	307035	2777
	eRTBVL-PV-8	1b	396698	399952	3255
	eRTBVL-PV-9	1b	1031348	1040435	9088
	eRTBVL-PV-10	1b	1186327	1187756	1430
	eRTBVL-PV-11	1b	1208867	1210395	1529
	eRTBVL-PV-12	1b	1232049	1233381	1333
	eRTBVL-PV-13	1b	4875779	4878719	2941
	eRTBVL-PV-14	1b	6577117	6577695	579
	eRTBVL-PV-15	1b	6898398	6898853	456
	eRTBVL-PV-16	1b	12887115	12889724	2610
	eRTBVL-PV-17	1b	13374053	13380577	6525
	eRTBVL-PV-18	1b	16983460	16985284	1825
	eRTBVL-PV-19	1b	21959684	21959971	288
	eRTBVL-PV-20	1b	23217565	23220023	2459
	eRTBVL-PV-21	1b	24345297	24345550	254
	eRTBVL-PV-22	1b	39733559	39733807	249
	eRTBVL-PV-23	1b	39738814	39741665	2852
	eRTBVL-PV-24	1b	49954522	49955647	1126
	eRTBVL-PV-25	2a	47225973	47243423	17451
	eRTBVL-PV-26	2a	54061417	54062719	1303
	eRTBVL-PV-27	2a	58931067	58932414	1348
	eRTBVL-PV-28	2a	71165042	71165962	921
	eRTBVL-PV-29	2a	71169435	71169586	152
	eRTBVL-PV-30	2a	78166359	78166480	122
	eRTBVL-PV-31	2b	5021135	5024424	3290
	eRTBVL-PV-32	2b	5037068	5038890	1823
	eRTBVL-PV-33	2b	6904677	6907431	2755
	eRTBVL-PV-34	2b	36489758	36494092	4335
	eRTBVL-PV-35	2b	66654276	66655684	1409
	eRTBVL-PV-36	2b	69686352	69687814	1463
	eRTBVL-PV-37	3a	17048243	17055577	7335
	eRTBVL-PV-38	3a	18975833	18977519	1687
	eRTBVL-PV-39	3a	22202340	22211891	9552
	eRTBVL-PV-40	3a	23635211	23636775	1565
	eRTBVL-PV-41	3a	23649947	23650211	265
	eRTBVL-PV-42	3a	27086736	27091242	4507
	eRTBVL-PV-43	3a	27278692	27280678	1987
	eRTBVL-PV-44	3a	27284340	27291423	7084
	eRTBVL-PV-45	3a	29012157	29019199	7043
	eRTBVL-PV-46	3a	29032267	29039791	7525
	eRTBVL-PV-47	3a	30115297	30117958	2662
	eRTBVL-PV-48	3a	31039341	31041245	1905
	eRTBVL-PV-49	3a	31363124	31363320	197
	eRTBVL-PV-50	3a	34466398	34467041	644

eRTBVL-PV-51	3a	36333378	36342631	9254
eRTBVL-PV-52	3a	36688260	36693536	5277
eRTBVL-PV-53	3a	37629973	37633118	3146
eRTBVL-PV-54	3a	39401056	39404714	3659
eRTBVL-PV-55	3b	12616354	12618618	2265
eRTBVL-PV-56	3b	26853453	26859064	5612
eRTBVL-PV-57	3b	31701704	31703562	1859
eRTBVL-PV-58	3b	32696373	32697653	1281
eRTBVL-PV-59	3b	40572250	40579432	7183
eRTBVL-PV-60	4a	12095696	12098621	2926
eRTBVL-PV-61	4a	24956560	24956862	303
eRTBVL-PV-62	4a	28251292	28251592	301
eRTBVL-PV-63	4a	38897669	38898534	866
eRTBVL-PV-64	4a	39276138	39289592	13455
eRTBVL-PV-65	4a	48391477	48394031	2555
eRTBVL-PV-66	4a	49977558	49977988	431
eRTBVL-PV-67	4b	18665850	18669188	3339
eRTBVL-PV-68	4b	33907710	33908319	610
eRTBVL-PV-69	4b	35532334	35538644	6311
eRTBVL-PV-70	4b	37923002	37925105	2104
eRTBVL-PV-71	4b	38116213	38116340	128
eRTBVL-PV-72	4b	39391045	39395661	4617
eRTBVL-PV-73	4b	44589292	44591341	2050
eRTBVL-PV-74	5a	291818	292676	859
eRTBVL-PV-75	5a	403628	404270	643
eRTBVL-PV-76	5a	537386	539476	2091
eRTBVL-PV-77	5a	1705162	1707157	1996
eRTBVL-PV-78	5a	25124191	25124747	557
eRTBVL-PV-79	5a	26893710	26895791	2082
eRTBVL-PV-80	5a	27159207	27162028	2822
eRTBVL-PV-81	5a	29367274	29367389	116
eRTBVL-PV-82	5a	30364773	30373249	8477
eRTBVL-PV-83	5a	32237478	32240559	3082
eRTBVL-PV-84	5a	33946546	33947353	808
eRTBVL-PV-85	5b	14019566	14019694	129
eRTBVL-PV-86	5b	17285832	17287870	2039
eRTBVL-PV-87	5b	18543592	18543705	114
eRTBVL-PV-88	5b	38752761	38754667	1907
eRTBVL-PV-89	5b	58114293	58117387	3095
eRTBVL-PV-90	5b	61963868	61966905	3038
eRTBVL-PV-91	6a	32118919	32119472	554
eRTBVL-PV-92	6a	32129531	32130729	1199
eRTBVL-PV-93	6a	51058056	51058667	612
eRTBVL-PV-94	6b	24543079	24546892	3814
eRTBVL-PV-95	6b	28836574	28836797	224
eRTBVL-PV-96	6b	40524874	40525114	241
eRTBVL-PV-97	6b	43799880	43803816	3937
eRTBVL-PV-98	7a	10920119	10921303	1185
eRTBVL-PV-99	7a	25391005	25396094	5090
eRTBVL-PV-100	7a	26462463	26464622	2160
eRTBVL-PV-101	7a	27938955	27940390	1436
eRTBVL-PV-102	7a	42521996	42532240	10245
eRTBVL-PV-103	7a	53076722	53079154	2433
eRTBVL-PV-104	7a	53082542	53089261	6720
eRTBVL-PV-105	7b	29821097	29821241	145
eRTBVL-PV-106	7b	43144300	43144990	691
eRTBVL-PV-107	8a	10383269	10384590	1322
eRTBVL-PV-108	8a	21844492	21847395	2904
eRTBVL-PV-109	8a	22096925	22106930	10006
eRTBVL-PV-110	8a	31701941	31705519	3579
eRTBVL-PV-111	8b	4498582	4498797	216
eRTBVL-PV-112	8b	6450782	6451953	1172
eRTBVL-PV-113	8b	7352084	7355464	3381
eRTBVL-PV-114	8b	7469610	7471106	1497
eRTBVL-PV-115	8b	7480433	7482572	2140
eRTBVL-PV-116	8b	13620835	13624963	4129
eRTBVL-PV-117	8b	14448781	14450192	1412
eRTBVL-PV-118	8b	20300082	20301288	1207
eRTBVL-PV-119	8b	27990273	27990479	207
eRTBVL-PV-120	8b	29154119	29156902	2784
eRTBVL-PV-121	8b	38379699	38379950	252
eRTBVL-PV-122	8b	40082699	40082929	231
eRTBVL-PV-123	8b	41195046	41200523	5478
eRTBVL-PV-124	8b	46784628	46784869	242
eRTBVL-PV-125	8b	51564142	51564390	249
eRTBVL-PV-126	9a	14972502	14972623	122
eRTBVL-PV-127	9a	24714989	24716215	1227
eRTBVL-PV-128	9a	38273290	38273518	229
eRTBVL-PV-129	9a	49598040	49598303	264
eRTBVL-PV-130	9a	49700136	49706383	6248
eRTBVL-PV-131	9a	59493974	59494385	412
eRTBVL-PV-132	9a	64307855	64308103	249
eRTBVL-PV-133	9a	81151500	81151622	123

eRTBVL-PV-134	9b	24690102	24690230	129
eRTBVL-PV-135	9b	24909503	24909873	371
eRTBVL-PV-136	contig00232	4119	5003	885
eRTBVL-PV-137	contig00232	7490	9047	1558
eRTBVL-PV-138	contig04516	6698	7223	526
eRTBVL-PV-139	contig07306	2963	3657	695
eRTBVL-PV-140	contig10874	1993	5133	3141
eRTBVL-PV-141	contig13560	1	2265	2265
eRTBVL-PV-142	contig14534	6742	10793	4052
eRTBVL-PV-143	contig14817	4964	9035	4072
eRTBVL-PV-144	contig15669	2199	7033	4835
eRTBVL-PV-145	contig16552	8541	8895	355
eRTBVL-PV-146	contig16803	8624	10091	1468
eRTBVL-PV-147	contig17490	9165	9900	736
eRTBVL-PV-148	contig19340	7612	9275	1664
eRTBVL-PV-149	contig23934	7311	8442	1132
eRTBVL-PV-150	contig26420	1	5153	5153
eRTBVL-PV-151	contig27282	6121	7901	1781
eRTBVL-PV-152	contig28007	5462	7772	2311
eRTBVL-PV-153	contig33093	1	1980	1980
eRTBVL-PV-154	contig38724	825	6032	5208
eRTBVL-PV-155	contig40240	1739	4490	2752
eRTBVL-PV-156	contig42682	89	2542	2454
eRTBVL-PV-157	contig56269	639	5335	4697
eRTBVL-PV-158	contig57181	1	1173	1173
eRTBVL-PV-159	contig65105	617	4883	4267
eRTBVL-PV-160	contig65840	856	4864	4009
eRTBVL-PV-161	contig66971	3	1315	1313
eRTBVL-PV-162	contig67407	11	4786	4776
eRTBVL-PV-163	contig68515	3285	4746	1462
eRTBVL-PV-164	contig69369	1	4715	4715
eRTBVL-PV-165	contig70699	1	4455	4455
eRTBVL-PV-166	contig71529	1226	1492	267
eRTBVL-PV-167	contig78634	1	4286	4286
eRTBVL-PV-168	contig84261	3627	4179	553
eRTBVL-PV-169	contig92762	3401	3649	249
eRTBVL-PV-170	contig93078	35	3928	3894
eRTBVL-PV-171	contig95396	4	190	187
eRTBVL-PV-172	contig95625	146	563	418
eRTBVL-PV-173	contig97108	1	3776	3776
eRTBVL-PV-174	contig100980	4	3719	3716
eRTBVL-PV-175	contig101618	1	3539	3539
eRTBVL-PV-176	contig103268	370	1713	1344
eRTBVL-PV-177	contig104236	1042	1290	249
eRTBVL-PV-178	contig104740	1	3639	3639
eRTBVL-PV-179	contig109135	3270	3441	172
eRTBVL-PV-180	contig109458	2659	3532	874
eRTBVL-PV-181	contig109787	915	3527	2613
eRTBVL-PV-182	contig111675	2844	3484	641
eRTBVL-PV-183	contig113595	1003	1263	261
eRTBVL-PV-184	contig114926	1	3122	3122
eRTBVL-PV-185	contig117138	1	461	461
eRTBVL-PV-186	contig118733	1	3339	3339
eRTBVL-PV-187	contig119510	1	2358	2358
eRTBVL-PV-188	contig120941	1	3302	3302
eRTBVL-PV-189	contig121505	6	3128	3123
eRTBVL-PV-190	contig123457	122	1353	1232
eRTBVL-PV-191	contig125167	3	340	338
eRTBVL-PV-192	contig132626	7	2827	2821
eRTBVL-PV-193	contig133531	25	165	141
eRTBVL-PV-194	contig134477	666	2749	2084
eRTBVL-PV-195	contig134491	1	3069	3069
eRTBVL-PV-196	contig135050	579	843	265
eRTBVL-PV-197	contig140967	1	2947	2947
eRTBVL-PV-198	contig143720	2498	2908	411
eRTBVL-PV-199	contig146055	1	2890	2890
eRTBVL-PV-200	contig146409	145	2547	2403
eRTBVL-PV-201	contig150822	1659	1905	247
eRTBVL-PV-202	contig151393	3	757	755
eRTBVL-PV-203	contig153292	4	680	677
eRTBVL-PV-204	contig153458	3	2785	2783
eRTBVL-PV-205	contig154783	109	2768	2660
eRTBVL-PV-206	contig157396	5	609	605
eRTBVL-PV-207	contig158232	34	1885	1852
eRTBVL-PV-208	contig160475	1	2691	2691
eRTBVL-PV-209	contig162278	6	204	199
eRTBVL-PV-210	contig165276	280	2634	2355
eRTBVL-PV-211	contig165278	1	1924	1924
eRTBVL-PV-212	contig167429	1	201	201
eRTBVL-PV-213	contig170391	1	2490	2490
eRTBVL-PV-214	contig170962	1	2476	2476
eRTBVL-PV-215	contig171497	53	1116	1064
eRTBVL-PV-216	contig174249	73	2525	2453

eRTBVL-PV-217	contig174589	1046	1812	767
eRTBVL-PV-218	contig177937	869	2409	1541
eRTBVL-PV-219	contig179587	138	2471	2334
eRTBVL-PV-220	contig180188	3	2110	2108
eRTBVL-PV-221	contig181060	2150	2398	249
eRTBVL-PV-222	contig183630	1	2421	2421
eRTBVL-PV-223	contig185217	1737	2404	668
eRTBVL-PV-224	contig186933	1496	2343	848
eRTBVL-PV-225	contig190104	108	1864	1757
eRTBVL-PV-226	contig192188	3	2338	2336
eRTBVL-PV-227	contig194732	1	1342	1342
eRTBVL-PV-228	contig195631	1291	2244	954
eRTBVL-PV-229	contig198139	1	2215	2215
eRTBVL-PV-230	contig198445	1510	1935	426
eRTBVL-PV-231	contig199405	1672	1936	265
eRTBVL-PV-232	contig203536	1	1952	1952
eRTBVL-PV-233	contig205781	1	2153	2153
eRTBVL-PV-234	contig207077	1136	2196	1061
eRTBVL-PV-235	contig208928	1501	2179	679
eRTBVL-PV-236	contig209643	1	2125	2125
eRTBVL-PV-237	contig213253	1	1744	1744
eRTBVL-PV-238	contig214761	1	2011	2011
eRTBVL-PV-239	contig219176	17	2083	2067
eRTBVL-PV-240	contig224042	1	1811	1811
eRTBVL-PV-241	contig225014	1	1715	1715
eRTBVL-PV-242	contig225749	168	651	484
eRTBVL-PV-243	contig227007	3	2025	2023
eRTBVL-PV-244	contig228339	4	2015	2012
eRTBVL-PV-245	contig228372	1442	2014	573
eRTBVL-PV-246	contig228859	1	2010	2010
eRTBVL-PV-247	contig229067	54	1028	975
eRTBVL-PV-248	contig229465	141	1658	1518
eRTBVL-PV-249	contig231175	1	1992	1992
eRTBVL-PV-250	contig231442	1	1989	1989
eRTBVL-PV-251	contig231584	5	1985	1981
eRTBVL-PV-252	contig232904	6	1978	1973
eRTBVL-PV-253	contig233274	1	1976	1976
eRTBVL-PV-254	contig233557	730	1829	1100
eRTBVL-PV-255	contig233659	1	1716	1716
eRTBVL-PV-256	contig234168	23	1943	1921
eRTBVL-PV-257	contig234585	1	1966	1966
eRTBVL-PV-258	contig235297	4	1960	1957
eRTBVL-PV-259	contig236274	1	1952	1952
eRTBVL-PV-260	contig237162	646	1005	360
eRTBVL-PV-261	contig237578	1095	1222	128
eRTBVL-PV-262	contig238896	1	1769	1769
eRTBVL-PV-263	contig239290	1	1927	1927
eRTBVL-PV-264	contig239869	1	1925	1925
eRTBVL-PV-265	contig242182	382	1831	1450
eRTBVL-PV-266	contig244029	320	1893	1574
eRTBVL-PV-267	contig244252	1	1472	1472
eRTBVL-PV-268	contig245258	1	267	267
eRTBVL-PV-269	contig245512	8	1884	1877
eRTBVL-PV-270	contig257568	1675	1796	122
eRTBVL-PV-271	contig258047	1	1789	1789
eRTBVL-PV-272	contig261400	1	1773	1773
eRTBVL-PV-273	contig263223	360	1760	1401
eRTBVL-PV-274	contig263512	29	1748	1720
eRTBVL-PV-275	contig265351	2	1725	1724
eRTBVL-PV-276	contig265719	321	1742	1422
eRTBVL-PV-277	contig268416	383	1726	1344
eRTBVL-PV-278	contig268591	299	1725	1427
eRTBVL-PV-279	contig270409	1	1712	1712
eRTBVL-PV-280	contig275184	35	1682	1648
eRTBVL-PV-281	contig277910	4	1664	1661
eRTBVL-PV-282	contig279270	1	1652	1652
eRTBVL-PV-283	contig281585	1	1641	1641
eRTBVL-PV-284	contig283415	1	1630	1630
eRTBVL-PV-285	contig285707	2	771	770
eRTBVL-PV-286	contig288359	633	894	262
eRTBVL-PV-287	contig288979	1	853	853
eRTBVL-PV-288	contig289100	1	1594	1594
eRTBVL-PV-289	contig293261	663	1229	567
eRTBVL-PV-290	contig295141	1006	1558	553
eRTBVL-PV-291	contig297360	7	1542	1536
eRTBVL-PV-292	contig297814	163	1539	1377
eRTBVL-PV-293	contig298531	1	1538	1538
eRTBVL-PV-294	contig300580	2	1523	1522
eRTBVL-PV-295	contig300617	2	1142	1141
eRTBVL-PV-296	contig305203	1281	1494	214
eRTBVL-PV-297	contig305513	1	1497	1497
eRTBVL-PV-298	contig305634	1	1494	1494
eRTBVL-PV-299	contig306280	528	1493	966

eRTBVL-PV-300	contig306622	1	1492	1492
eRTBVL-PV-301	contig306653	1	1003	1003
eRTBVL-PV-302	contig307089	1	1488	1488
eRTBVL-PV-303	contig309113	3	1477	1475
eRTBVL-PV-304	contig310311	1	1469	1469
eRTBVL-PV-305	contig313827	1	1451	1451
eRTBVL-PV-306	contig315721	1	1304	1304
eRTBVL-PV-307	contig321520	60	1408	1349
eRTBVL-PV-308	contig322702	1	1399	1399
eRTBVL-PV-309	contig326065	1	1253	1253
eRTBVL-PV-310	contig326231	3	1383	1381
eRTBVL-PV-311	contig332136	1	1353	1353
eRTBVL-PV-312	contig333208	3	1347	1345
eRTBVL-PV-313	contig334212	60	1342	1283
eRTBVL-PV-314	contig336240	1	1331	1331
eRTBVL-PV-315	contig336457	8	1331	1324
eRTBVL-PV-316	contig337431	639	1042	404
eRTBVL-PV-317	contig338152	1	1321	1321
eRTBVL-PV-318	contig338172	1	1322	1322
eRTBVL-PV-319	contig339500	22	886	865
eRTBVL-PV-320	contig340213	14	857	844
eRTBVL-PV-321	contig344685	252	1221	970
eRTBVL-PV-322	contig344888	1	1289	1289
eRTBVL-PV-323	contig348320	1	1273	1273
eRTBVL-PV-324	contig352320	1	1252	1252
eRTBVL-PV-325	contig352647	1	1028	1028
eRTBVL-PV-326	contig355292	1	1236	1236
eRTBVL-PV-327	contig356003	410	1226	817
eRTBVL-PV-328	contig357263	1	1227	1227
eRTBVL-PV-329	contig357316	1	1226	1226
eRTBVL-PV-330	contig357639	1	1228	1228
eRTBVL-PV-331	contig359726	1	1215	1215
eRTBVL-PV-332	contig360466	1	1210	1210
eRTBVL-PV-333	contig360775	1	994	994
eRTBVL-PV-334	contig360899	1	1212	1212
eRTBVL-PV-335	contig361910	1	1207	1207
eRTBVL-PV-336	contig362040	1	1201	1201
eRTBVL-PV-337	contig363012	2	1201	1200
eRTBVL-PV-338	contig367810	1	1175	1175
eRTBVL-PV-339	contig368425	1	1176	1176
eRTBVL-PV-340	contig368692	1	806	806
eRTBVL-PV-341	contig368694	2	1113	1112
eRTBVL-PV-342	contig369964	4	1170	1167
eRTBVL-PV-343	contig372350	3	1157	1155
eRTBVL-PV-344	contig372969	1	1154	1154
eRTBVL-PV-345	contig373903	3	1153	1151
eRTBVL-PV-346	contig375667	1	1145	1145
eRTBVL-PV-347	contig377500	1	1136	1136
eRTBVL-PV-348	contig380972	288	1117	830
eRTBVL-PV-349	contig381076	281	1113	833
eRTBVL-PV-350	contig381622	1	574	574
eRTBVL-PV-351	contig381861	4	1116	1113
eRTBVL-PV-352	contig383034	1	1109	1109
eRTBVL-PV-353	contig385629	1	1100	1100
eRTBVL-PV-354	contig389089	1	1086	1086
eRTBVL-PV-355	contig389975	18	912	895
eRTBVL-PV-356	contig390383	1	1083	1083
eRTBVL-PV-357	contig393429	357	1068	712
eRTBVL-PV-358	contig394054	89	1050	962
eRTBVL-PV-359	contig394306	1	1065	1065
eRTBVL-PV-360	contig395579	1	1061	1061
eRTBVL-PV-361	contig396072	1	1053	1053
eRTBVL-PV-362	contig396293	1	1053	1053
eRTBVL-PV-363	contig396505	2	859	858
eRTBVL-PV-364	contig396853	1	1056	1056
eRTBVL-PV-365	contig397903	2	1044	1043
eRTBVL-PV-366	contig400171	1	1041	1041
eRTBVL-PV-367	contig402295	16	1031	1016
eRTBVL-PV-368	contig404910	3	120	118
eRTBVL-PV-369	contig405455	2	1020	1019
eRTBVL-PV-370	contig406308	1	708	708
eRTBVL-PV-371	contig407275	152	1011	860
eRTBVL-PV-372	contig407697	1	738	738
eRTBVL-PV-373	contig407872	2	1008	1007
eRTBVL-PV-374	contig408004	1	1008	1008
eRTBVL-PV-375	contig408196	1	1009	1009
eRTBVL-PV-376	contig409388	2	1005	1004

Table S7 Distribution of eGPV1 and eGPV2 segments in Poaceae genomes

Species	Segment name	Family ^a	Position in grass genomes				Position in eGPV ^b		Length (bp)
			Chromosome/scaffold	Strand	Start	End	Start	End	
<i>Oryza sativa</i>	eGPV1-OS1	eGPV1 ^{Or}	1	-	8849149	8847147	3966	5967	2003
	eGPV1-OS2	eGPV1 ^{Or}	2	-	11313137	11312114	3309	4332	1024
	eGPV1-OS3	eGPV1 ^{Or}	2	-	11313900	11313138	31	799	763
	eGPV1-OS4	eGPV1 ^{Or}	2	+	11313901	11314199	356	684	299
	eGPV1-OS5	eGPV1 ^{Or}	5	-	20231116	20224846	63	6425	6271
	eGPV1-OS6	eGPV1 ^{Or}	7	-	11604767	11601272	2748	6526	3496
	eGPV1-OS7	eGPV1 ^{Or}	7	+	11604768	11605869	270	1384	1102
	eGPV1-OS8	eGPV1 ^{Or}	7	-	27575480	27572485	3557	6553	2996
					27572484	27571939	1	560	546
	eGPV1-OS9	eGPV1 ^{Or}	11	-	11865464	11860427	1221	6434	5038
	eGPV1-OS10	eGPV1 ^{Or}	11	+	19845104	19848248	3384	6365	3145
	eGPV1-OS11	eGPV1 ^{Or}	12	+	19505111	19511369	42	6355	6259
	eGPV2-OS1	eGPV2 ^{Or}	1	+	8840504	8847149	4	6589	6646
	eGPV2-OS2	eGPV2 ^{Or}	1	-	23908392	23902732	967	6643	5661
	eGPV2-OS3	eGPV2 ^{Or}	1	-	34125359	34121219	1488	5638	4141
	eGPV2-OS4	eGPV2 ^{Or}	1	+	34125360	34126305	459	1404	946
	eGPV2-OS5	eGPV2 ^{Or}	2	+	11304753	11311418	1	6650	6666
	eGPV2-OS6	eGPV2 ^{Or}	2	-	11312045	11311420	1	624	626
	eGPV2-OS7	eGPV2 ^{Or}	3	+	22472858	22472966	5056	5164	109
	eGPV2-OS8	eGPV2 ^{Or}	3	+	22472973	22474562	5056	6653	1590
					22474563	22474682	1	122	120
	eGPV2-OS9	eGPV2 ^{Or}	3	-	22478535	22474683	582	4442	3853
	eGPV2-OS10	eGPV2 ^{Or}	4	-	23050761	23049961	5296	6549	801
	eGPV2-OS11	eGPV2 ^{Or}	4	+	23051051	23052842	4835	6653	1792
					23052843	23052909	1	66	67
	eGPV2-OS12	eGPV2 ^{Or}	7	+	11608404	11609584	1821	3001	1181
	eGPV2-OS13	eGPV2 ^{Or}	8	+	25041604	25041980	6258	6652	377
	eGPV2-OS14	eGPV2 ^{Or}	8	-	25047553	25041981	271	5829	5573
	eGPV2-OS15	eGPV2 ^{Or}	9	+	20408412	20408865	62	515	454
	eGPV2-OS16	eGPV2 ^{Or}	9	+	20415024	20421219	510	6611	6196
	eGPV2-OS17	eGPV2 ^{Or}	9	-	21426023	21423507	4191	6653	2517
	eGPV2-OS18	eGPV2 ^{Or}	9	-	21431914	21430262	2545	4198	1653
	eGPV2-OS19	eGPV2 ^{Or}	9	-	21432081	21431915	5168	5350	167
	eGPV2-OS20	eGPV2 ^{Or}	10	+	8405909	8410513	2057	6653	4605
	eGPV2-OS21	eGPV2 ^{Or}	11	+	8293415	8299957	69	6646	6543
	eGPV2-OS22	eGPV2 ^{Or}	11	-	28547732	28541505	23	6653	6228
	eGPV2-OS23	eGPV2 ^{Or}	11	-	28591917	28585688	23	6653	6230

	eGPVX-OS1	Unsigned	6	-	2756800	2756688	81/81	193/193	113
<i>Oryza glaberrima</i>	eGPV1-OG1	eGPV1 ^{Or}	4	-	26260103	26253601	15	6560	6503
	eGPV1-OG2	eGPV1 ^{Or}	5	-	15500612	15494086	63	6419	6527
	eGPV1-OG3	eGPV1 ^{Or}	7	+	7758507	7764924	1	6523	6418
	eGPV1-OG4	eGPV1 ^{Or}	9	+	16969612	16970506	278	1172	895
	eGPV1-OG5	eGPV1 ^{Or}	9	+	16970503	16971982	4748	6228	1480
	eGPV1-OG6	eGPV1 ^{Or}	10	-	8397098	8392335	1567	6153	4764
	eGPV1-OG7	eGPV1 ^{Or}	11	-	6592706	6587470	1285	6560	5237
	eGPV1-OG8	eGPV1 ^{Or}	11	-	9224687	9219391	1221	6434	5297
	eGPV1-OG9	eGPV1 ^{Or}	11	+	15896886	15897190	2705	3009	305
	eGPV1-OG10	eGPV1 ^{Or}	11	-	15899867	15899177	1517	2240	691
	eGPV1-OG11	eGPV1 ^{Or}	Oglab02_unplaced013	-	19352	17967	5145	6520	1386
	eGPV1-OG12	eGPV1 ^{Or}	Oglab02_unplaced013	-	22131	20365	1569	3369	1767
	eGPV2-OG1	eGPV2 ^{Or}	1	-	24848084	24847398	4731	5417	687
	eGPV2-OG2	eGPV2 ^{Or}	1	-	24850826	24849905	2675	3596	922
	eGPV2-OG3	eGPV2 ^{Or}	4	-	15508021	15507736	6265	6549	286
	eGPV2-OG4	eGPV2 ^{Or}	4	-	15508460	15508338	5314	5437	123
	eGPV2-OG5	eGPV2 ^{Or}	4	+	15508778	15510896	4835	6653	2119
					15510897	15510963	1	66	67
	eGPV2-OG6	eGPV2 ^{Or}	4	-	26265352	26260098	1663	6651	5255
	eGPV2-OG7	eGPV2 ^{Or}	4	+	26265353	26265878	1048	1592	526
	eGPV2-OG8	eGPV2 ^{Or}	6	+	16933152	16940095	54	6653	6944
					16940096	16940152	1	57	57
	eGPV2-OG9	eGPV2 ^{Or}	10	+	6511230	6516580	1434	6593	5351
	eGPV2-OG10	eGPV2 ^{Or}	10	+	8397190	8401147	734	4471	3958
	eGPV2-OG11	eGPV2 ^{Or}	10	+	8403814	8405594	4466	6246	1781
	eGPV2-OG12	eGPV2 ^{Or}	11	+	6585334	6587452	2334	4453	2119
	eGPV2-OG13	eGPV2 ^{Or}	11	-	15893636	15893277	4663	5023	360
	eGPV2-OG14	eGPV2 ^{Or}	11	+	15893637	15894763	1170	2312	1127
	eGPV2-OG15	eGPV2 ^{Or}	11	+	15895817	15897227	2708	4118	1411
	eGPV2-OG16	eGPV2 ^{Or}	12	-	13437088	13433323	1488	4597	3766
	eGPV2-OG17	eGPV2 ^{Or}	12	+	13437089	13437510	459	891	422
	eGPV2-OG18	eGPV2 ^{Or}	12	+	13437516	13438725	5418	6611	1210
	eGPV2-OG19	eGPV2 ^{Or}	Oglab12_unplaced047	+	58	4574	357	5062	4517
	eGPVX-OG1	Unsigned	11	-	6045997	6045480	6045/613	6560/665	518
					6045479	6044654	1/1	1284/128	826
<i>Oryza glumaepatula</i>	eGPV1-OU1	eGPV1 ^{Or}	10	-	15996112	15995870	5350	5592	243
	eGPV1-OU2	eGPV1 ^{Or}	11	-	11666533	11665031	1221	2725	1503
	eGPV2-OU1	eGPV2 ^{Or}	3	+	4729510	4730673	5485	6653	1164
	eGPV2-OU2	eGPV2 ^{Or}	4	-	20425951	20425145	5296	6549	807
	eGPV2-OU3	eGPV2 ^{Or}	4	+	20426253	20428023	4843	6653	1771
					20428024	20428093	1	66	70
	eGPV2-OU4	eGPV2 ^{Or}	12	+	21618343	21619520	3736	4914	1178

	eGPVX-OU1	Unsigned	1	-	25261149	25260838	6237/633	6542/663	312
	eGPVX-OU2	Unsigned	6	-	25698260	25697867	249/250	642/643	394
	eGPVX-OU3	Unsigned	12	+	21626409	21626532	64376530	6560/665	124
<i>Oryza meridionalis</i>	eGPV1-OM1	eGPV1 ^{Or}	4	-	23170204	23169068	4526	5662	1137
	eGPV1-OM2	eGPV1 ^{Or}	6	+	11359819	11361254	2627	3551	1436
	eGPV1-OM3	eGPV1 ^{Or}	7	-	18949381	18949008	4535	4908	374
	eGPV1-OM4	eGPV1 ^{Or}	8	-	5997777	5996654	616	1739	1124
	eGPV1-OM5	eGPV1 ^{Or}	8	-	5999649	5999224	5833	6137	426
	eGPV1-OM6	eGPV1 ^{Or}	8	+	9583173	9583690	4195	4578	518
	eGPV1-OM7	eGPV1 ^{Or}	8	-	10030167	10030053	4128	4242	115
	eGPV1-OM8	eGPV1 ^{Or}	9	-	11799894	11798813	3529	4248	1082
	eGPV1-OM9	eGPV1 ^{Or}	11	-	7758285	7758179	1	107	107
	eGPV1-OM10	eGPV1 ^{Or}	11	+	7758637	7759163	2055	2446	527
	eGPV2-OM1	eGPV2 ^{Or}	1	+	843164	844272	5006	5785	1109
	eGPV2-OM2	eGPV2 ^{Or}	1	-	19176727	19176596	3867	3998	132
	eGPV2-OM3	eGPV2 ^{Or}	2	+	19424534	19424933	5150	5410	400
	eGPV2-OM4	eGPV2 ^{Or}	3	+	2366827	2368542	2080	3410	1716
	eGPV2-OM5	eGPV2 ^{Or}	3	+	2369641	2370890	3258	4414	1250
	eGPV2-OM6	eGPV2 ^{Or}	8	+	9592885	9593588	4729	5204	704
	eGPV2-OM7	eGPV2 ^{Or}	8	-	10030062	10029035	4510	5344	1028
	eGPVX-OM1	Unsigned	9	+	3194279	3194422	6043/613	6186/627	144
<i>Oryza punctata</i>	eGPV1-OP1	eGPV1 ^{Or}	1	+	36462406	36468381	496	6528	5976
	eGPV1-OP2	eGPV1 ^{Or}	7	-	14512232	14518685	55	6547	6454
	eGPV1-OP3	eGPV1 ^{Or}	12	+	17242165	17243967	1	1802	1803
	eGPV1-OP4	eGPV1 ^{Or}	12	+	17245498	17250240	1797	6538	4743
	eGPV2-OP1	eGPV2 ^{Or}	2	-	18174237	18173741	1002	1500	497
	eGPV2-OP2	eGPV2 ^{Or}	2	+	18174238	18174792	1520	2085	555
	eGPV2-OP3	eGPV2 ^{Or}	2	+	18174893	18175461	3969	4537	569
	eGPV2-OP4	eGPV2 ^{Or}	2	+	18175564	18175698	1551	1685	135
	eGPV2-OP5	eGPV2 ^{Or}	2	+	18175703	18178269	4053	6628	2567
	eGPV2-OP6	eGPV2 ^{Or}	2	-	18178605	18178270	31	369	336
	eGPV2-OP7	eGPV2 ^{Or}	2	+	18178606	18180215	357	1988	1610
	eGPV2-OP8	eGPV2 ^{Or}	2	+	18180216	18182535	4359	6650	2320
	eGPV2-OP9	eGPV2 ^{Or}	3	-	1877451	1870791	1	6629	6661
	eGPV2-OP10	eGPV2 ^{Or}	3	-	1877899	1877449	6228	6653	451
	eGPV2-OP11	eGPV2 ^{Or}	4	-	11094038	11089082	44	5005	4957
	eGPV2-OP12	eGPV2 ^{Or}	7	-	14019246	14018987	3763	4022	260
	eGPV2-OP13	eGPV2 ^{Or}	7	+	14019291	14021601	4347	6653	2311
					14021602	14021994	1	408	393
	eGPV2-OP14	eGPV2 ^{Or}	7	-	14022570	14022140	3935	4365	431
	eGPV2-OP15	eGPV2 ^{Or}	9	+	2488988	2494897	476	6653	5910
	eGPV2-OP16	eGPV2 ^{Or}	9	+	2508244	2509198	5707	6653	955
					2509199	2509685	1	490	487

	eGPV2-OP17	eGPV2 ^{Or}	9	-	2510214	2509778	6191	6646	437
	eGPV2-OP18	eGPV2 ^{Or}	9	-	2511022	2510216	5703	6471	807
	eGPV2-OP19	eGPV2 ^{Or}	9	+	2511023	2513219	1330	3521	2197
	eGPV2-OP20	eGPV2 ^{Or}	9	+	2513220	2515543	4196	6514	2324
	eGPV2-OP21	eGPV2 ^{Or}	12	-	17241666	17239641	4690	6629	2026
	eGPV2-OP22	eGPV2 ^{Or}	12	-	17241795	17241667	1758	1891	129
	eGPV2-OP23	eGPV2 ^{Or}	12	-	17242162	17241796	2142	2513	367
<i>Phyllostachys heterocycla</i>	eGPV1-PH1	eGPV1 ^{So}	PH01000128	-	1080539	1080132	5190	5599	408
	eGPV1-PH2	eGPV1 ^{So}	PH01000128	-	1086293	1084861	2838	4275	1433
	eGPV1-PH3	eGPV1 ^{So}	PH01003803	-	13301	12391	3433	4347	911
	eGPV2-PH1	eGPV2 ^{So}	PH01002077	+	92577	92764	4355	4540	188
	eGPV2-PH2	eGPV2 ^{So}	PH01002077	-	99502	98026	2678	4159	1477
	eGPV2-PH3	eGPV2 ^{So}	PH01002077	-	99603	99500	1509	1612	104
	eGPV2-PH4	eGPV2 ^{So}	PH01003803	-	50917	50558	2589	2949	360
	eGPVX-Ph1	Unsigned	PH01003803	-	53145	51790	257/258	1630/161	1356
<i>Sorghum bicolor</i>	eGPV1-SB1	eGPV1 ^{So}	1	+	54922322	54928778	1	6548	6457
					54928779	54928805	1	27	27
	eGPV1-SB2	eGPV1 ^{So}	4	+	16773368	16779949	1	6500	6582
	eGPV1-SB3	eGPV1 ^{So}	6	+	37559217	37562676	124	3584	3460
	eGPV1-SB4	eGPV1 ^{So}	6	-	37563756	37562692	5028	6030	1065
	eGPV1-SB5	eGPV1 ^{So}	6	-	37571348	37566378	22	5033	4971
	eGPV1-SB6	eGPV1 ^{So}	10	-	16463710	16457225	3	6548	6486
					16457224	16457110	1	116	115
	eGPV1-SB7	eGPV1 ^{So}	10	+	16463785	16463938	6199	6352	154
	eGPV2-SB1	eGPV2 ^{So}	3	-	293657	286801	1	6879	6857
	eGPV2-SB2	eGPV2 ^{So}	3	-	18263770	18263397	4846	5222	374
	eGPV2-SB3	eGPV2 ^{So}	3	-	18270253	18267703	4321	6880	2551
	eGPV2-SB4	eGPV2 ^{So}	4	+	46230217	46237238	13	6893	7022
	eGPV2-SB5	eGPV2 ^{So}	5	+	13007912	13013640	1152	6893	5729
	eGPV2-SB6	eGPV2 ^{So}	5	+	13013665	13020548	1	6872	6884
	eGPV2-SB7	eGPV2 ^{So}	5	-	13715088	13710895	1955	6179	4194
	eGPV2-SB8	eGPV2 ^{So}	5	+	33532220	33538691	196	6759	6472
	eGPV2-SB9	eGPV2 ^{So}	5	+	47930070	47936859	10	6796	6790
	eGPV2-SB10	eGPV2 ^{So}	5	-	58530701	58530474	5519	5751	228
	eGPV2-SB11	eGPV2 ^{So}	5	+	58530702	58531602	1	926	901
	eGPV2-SB12	eGPV2 ^{So}	5	+	58531603	58536535	1960	6892	4933
	eGPV2-SB13	eGPV2 ^{So}	5	+	58536624	58537273	1	674	650
	eGPV2-SB14	eGPV2 ^{So}	5	-	58541750	58537274	569	5057	4477
	eGPV2-SB15	eGPV2 ^{So}	6	+	31383876	31384128	6628	6891	253
	eGPV2-SB16	eGPV2 ^{So}	6	+	31384129	31388146	2887	6893	4018
					31388147	31391037	1	2965	2891
	eGPV2-SB17	eGPV2 ^{So}	6	+	37580566	37581112	6352	6893	547
	eGPV2-SB18	eGPV2 ^{So}	6	+	37581137	37587995	1	6893	6859

	eGPV2-SB19	eGPV2 ^{So}	6	-	37587996	37588062	1	68	67
	eGPV2-SB20	eGPV2 ^{So}	7	+	44430825	44423970	4	6893	6856
	eGPV2-SB21	eGPV2 ^{So}	7	-	36892423	36897609	1	5688	5187
	eGPV2-SB22	eGPV2 ^{So}	7	-	36898552	36897755	5324	6123	798
	eGPV2-SB23	eGPV2 ^{So}	Super_19	+	36901720	36898553	2513	6140	3168
	eGPV2-SB24	eGPV2 ^{So}	Super_19	+	790000	792947	677	3641	2948
	eGPV2-SB25	eGPV2 ^{So}	Super_19	+	793795	795128	5082	6429	1334
	eGPV2-SB26	eGPV2 ^{So}	Super_19	+	795337	796253	4127	5041	917
	eGPV2-SB27	eGPV2 ^{So}	Super_19	+	796354	796770	6461	6891	417
	eGPVX-SB1	Unsigned	5	+	796771	800729	2887	6867	3959
<i>Panicum virgatum</i>	eGPV1-PV1	eGPV1 ^{So}	7a	-	12982623	12982413	6326/666	6548/689	211
	eGPV1-PV2	eGPV1 ^{So}	7a	+	40139014	40138744	3241	3511	271
	eGPV2-PV1	eGPV2 ^{So}	3b	-	40139022	40141657	3274	5904	2636
	eGPV2-PV2	eGPV2 ^{So}	7a	+	5042450	5042142	1991	2322	309
	eGPV2-PV3	eGPV2 ^{So}	7a	+	35052031	35054066	1	2699	2036
	eGPV2-PV4	eGPV2 ^{So}	7a	-	39593523	39594633	2684	3795	1111
	eGPV2-PV5	eGPV2 ^{So}	7a	-	39595884	39595051	1146	1969	834
	eGPV2-PV6	eGPV2 ^{So}	Contig04516	-	39602918	39601951	1	1079	968
	eGPV2-PV7	eGPV2 ^{So}	Contig171646	-	10517	6657	2054	5989	3861
	eGPV2-PV8	eGPV2 ^{So}	Contig171646	-	646	389	1414	1675	258
	eGPV2-PV9	eGPV2 ^{So}	Contig171646	-	953	716	404	639	238
	eGPV2-PV10	eGPV2 ^{So}	Contig171646	-	1469	1173	333	618	297
	eGPVX-PV1	Unsigned	7a	+	2090	1972	31	151	119
	eGPVX-PV2	Unsigned	8b	+	35051553	35051689	6154/648	6294/663	137
	eGPVX-PV3	Unsigned	Contig04516	-	47417775	47417980	337/336	542/541	206
					6300	6085	6110/644	6310/665	216

^aBecause the eGPV1 and eGPV2 sequences show a local high similarity, the classification of a small amount of short segments only derived from the highly similar region can't be unambiguously determined, and are marked as "Unsigned".

^bThe eGPV1/eGPV2 segments in the Ehrhartoideae and Bambusoideae genomes are mapped to the viral genome of eGPV1s/eGPV2s constructed from the *O. sativa* genome, and the eGPV1/eGPV2 segments in the Panicoideae genomes are mapped to the viral genome of eGPV1s/eGPV2s constructed from the *S. bicolor* genome; the unsigned segments are mapped to both eGPV1 and eGPV2 viral genomes.

Table S8 Mapping of eGPV1 and eGPV2 loci in the *O. sativa* genome into other *Oryza* genomes

Information of eGPV loci in <i>O. sativa</i> genome			Mapping into other <i>Oryza</i> genomes				Remark
Locus number (position) ^a	Segments ^b	Family ^c	<i>O. glaberrima</i>	<i>O. glumaepatula</i>	<i>O. meridionalis</i>	<i>O. punctata</i>	
Locus1 (1:8840504-8849149)	eGPV1-OS1, eGPV2-OS1	eGPV1, 2	A (1.6923482-6923778)	A (1.10192607-10194752)	A (1≈8993855)	A (1≈8415586)	specific
Locus2 (1:23902732-23908392)	eGPV2-OS2	eGPV2	A (1.16686414-16686464)	A (1.26043345-26062879)	A (1.21706486-21706494)	A (1≈25718279)	specific
Locus3 (1:34121219-34126305)	eGPV2-OS3, eGPV2-OS4	eGPV2	P (1.24847398-24850826)	A (1.37035535-37039966)	A (1≈32368856)	A (1≈36538050)	shared
Locus4 (2:11304685-11314199)	eGPV1-OS2, eGPV1-OS3, eGPV1-OS4, eGPV2-OS5, eGPV2-OS6	eGPV1, 2	A (2≈9990278)	A (2≈10786295)	A (2≈12965281)	A (2≈11085428)	specific
Locus5 (3:22472858-22478535)	eGPV2-OS7, eGPV2-OS8, eGPV2-OS9	eGPV2	A (3.21127466-21144299)	A (3.24188491-24207759)	A (3.23820900-23831718)	A (3≈24683602)	specific
Locus6 (4:23049961-23052909)	eGPV2-OS10, eGPV2-OS11	eGPV2	P (4.15507736-15510963)	P (4.20425145-20428093)	P (by PCR)	A (4.20701665-20702064)	shared
Locus7 (5:20224846-20231116)	eGPV1-OS5	eGPV1	P (5.15494086-15500612)	A (5.20226666-20226726)	A (5≈17927496)	NA	shared
Locus8 (6:2756579-2756800)	eGPVX-OS1	unsigned	NA	NA	NA	NA	undetermined
Locus9 (7:11601272-11609584)	eGPV1-OS6, eGPV1-OS7, eGPV2-OS12	eGPV1, 2	A (7≈9685762)	A (7≈10281454)	NA	A (7≈12715201)	specific
Locus10 (7:27571939-27575480)	eGPV1-OS8	eGPV1	A (7.19735543-19735565)	A (7.26263270-26263297)	A (7.24129203-24129252)	A (7.29102683-29102710)	specific
Locus11 (8:25041604-25047553)	eGPV2-OS13, eGPV2-OS14	eGPV2	A (8.18879456-18879465)	A (8.23608448-23608461)	A (8.19905457-19905466)	NA	specific
Locus12 (9:20408412-20408865)	eGPV2-OS15	eGPV2	A (9≈15164188)	A (9≈20532533)	A (9≈16147417)	A (9≈23782384)	specific
Locus13 (9:20415024-20421219)	eGPV2-OS16	eGPV2	A (9≈15166306)	A (9≈20549604)	A (9≈16163653)	A (9≈23789563)	specific
Locus14 (9:21423506-21432081)	eGPV2-OS17, eGPV2-OS18, eGPV2-OS19	eGPV2	NA	A (9≈21601883)	A (9≈17419017)	NA	undetermined
Locus15 (10:8405909-8410513)	eGPV2-OS20	eGPV2	A (10≈12586160)	A (10.7693053-7693094)	NA	NA	specific
Locus16 (11:8293415-8299957)	eGPV2-OS21	eGPV2	A (11≈6580050)	A (11≈8208302)	NA	NA	specific
Locus17 (11:11860427-11865464)	eGPV1-OS9	eGPV1	P (11.9219391-9224687)	P (11.11665031-11666533)	A (11≈12525767)	A (by PCR)	shared
Locus18 (11:19845104-19848248)	eGPV1-OS10	eGPV1	A (11≈15223836)	A (11≈19127289)	A (11.16885801-16885858)	NA	specific
Locus19 (11:28541504-28591917)	eGPV2-OS22, eGPV2-OS23	eGPV2	A (6≈11963074)	NA	NA	A (7≈12374530)	specific
Locus20 (12:19505111-19511369)	eGPV1-OS11	eGPV1	A (Oglab12_unplaced156:40610-41300)	A (12.19470470-19470551)	A (12≈13183476)	A (12≈20578764)	specific

^aThe segments fairly close to each other in the chromosome are taken as a locus in the mapping analysis (locus positions are provided in the brackets).

^bNames of the segments within a locus.

^cClassification of the segments within a locus.

P, presence of an orthologous insertion at the corresponding locus (highlighted by orange background; positions of orthologous insertions are shown in the brackets); A, absence of an orthologous insertion at the corresponding locus (predicted position or range of empty sites are shown in the brackets); NA, not available or missing data.

Table S9 Mapping of eGPV1 and eGPV2 loci in the genomes of *O. glaberrima*, *O. glumaepatula*, *O. meridionalis* and *O. punctata* into the genome of *O. sativa*

Genome	Information of eGPV loci in the genome			Mapping into <i>O. sativa</i> genome	Remark
	Locus number (position) ^a	Segments ^b	Family ^c		
<i>O. glaberrima</i>	Locus1 (1:24847398-24850826)	eGPV2-OG1, eGPV2-OG2	eGPV2	P (1:34121219-34126305)	shared
	Locus2 (4:15507736-15510963)	eGPV2-OG3, eGPV2-OG4, eGPV2-OG5	eGPV2	P (4:23049961-23052909)	shared
	Locus3 (4:26253608-26265878)	eGPV1-OG1, eGPV2-OG6, eGPV2-OG7	eGPV1, 2	A (4:35405701-35405717)	specific
	Locus4 (5:15494086-15500612)	eGPV1-OG2	eGPV1	P (5:20224846-20231116)	shared
	Locus5 (6:16933152-16940152)	eGPV2-OG8	eGPV2	A (6:22693382-22693889)	specific
	Locus6 (7:7758503-7764924)	eGPV1-OG3	eGPV1	A (7:9011784-9011913)	specific
	Locus7 (9:16969612-16971982)	eGPV1-OG4, eGPV1-OG5	eGPV1	NA	undetermined
	Locus8 (10:6511230-6516569)	eGPV2-OG9	eGPV2	A(10:≈8419468)	specific
	Locus9 (10:8392335-8405594)	eGPV1-OG6, eGPV2-OG10, eGPV2-OG11	eGPV1, 2	A (10:10945698-10945702)	specific
	Locus10 (11:6044654-6045997)	eGPVX-OG1	unsigned	A (11:7617002-7617023)	specific
	Locus11 (11:6585334-6592706)	eGPV1-OG7, eGPV2-OG12	eGPV1, 2	A (11:≈8377686)	specific
	Locus12 (11:9219391-9224687)	eGPV1-OG8	eGPV1	P (11:11860427-11865464)	shared
	Locus13 (11:15893277-15899867)	eGPV1-OG9, eGPV1-OG10, eGPV2-OG13, eGPV2-OG14, eGPV2-OG15	eGPV1, 2	A (11:20971323-20971750)	specific
	Locus14 (12:13433323-13438696)	eGPV2-OG16, eGPV2-OG17, eGPV2-OG18	eGPV2	NA	undetermined
	Locus15 (Oglab02_unplaced013:17967-22131)	eGPV1-OG11, eGPV1-OG12	eGPV1	NA	undetermined
	Locus16 (Oglab12_unplaced047:1-4574)	eGPV2-OG19	eGPV2	NA	undetermined
<i>O. glumaepatula</i>	Locus1 (1:25260891-25261149)	eGPVX-OU1	unsigned	A (1:23156362-23156379)	specific
	Locus2 (3:4729510-4730602)	eGPV2-OU1	eGPV2	A (3:4680875-4688150)	specific
	Locus3 (4:20425145-20428093)	eGPV2-OU2, eGPV2-OU3	eGPV2	P (4:23049961-23052909)	shared
	Locus4 (6:25697867-25698260)	eGPVX-OU2	unsigned	A (6:25598182-23052912)	specific
	Locus5 (10:15995870-15996112)	eGPV1-OU1	eGPV1	A (10:16866946-16867588)	specific
	Locus6 (11:11665031-11666533)	eGPV1-OU2	eGPV1	P (11:11860427-11865464)	shared
	Locus7 (12:21618343-21619520)	eGPV2-OU4,	eGPV2	NA	undetermined
	Locus8 (12:21626409-21626532)	eGPVX-OU3	unsigned	A (12:≈22229076)	specific
<i>O. meridionalis</i>	Locus1 (1:843164-844272)	eGPV2-OM1	eGPV2	A (1:≈1361354)	specific
	Locus2 (1:19176596-19176727)	eGPV2-OM2	eGPV2	A (1:20650545-20652372)	specific
	Locus3 (2:19424534-19424933)	eGPV2-OM3	eGPV2	A (2:≈18184236)	specific
	Locus4 (3:2366827-2370890)	eGPV2-OM4, eGPV2-OM5	eGPV2	A (3:≈1981605)	specific
	Locus5 (4:23169068-23170204)	eGPV1-OM1	eGPV1	A (4:≈28177610)	specific
	Locus6 (6:11359819-11361254)	eGPV1-OM2	eGPV1	A (6:9374355-9374493)	specific
	Locus7 (7:18949008-18949381)	eGPV1-OM3	eGPV1	A (7:≈22528997)	specific

	Locus8 (8:5996654-5999649)	eGPV1-OM4, eGPV1-OM5	eGPV1	A (8:6195297-6195410)	specific
	Locus9 (8:9583173-9583690)	eGPV1-OM6	eGPV1	NA	undetermined
	Locus10 (8:9592885-9593588)	eGPV2-OM6	eGPV2	NA	undetermined
	Locus11 (8:10029035-10030167)	eGPV1-OM7, eGPV2-OM7	eGPV1, 2	A (8:≈13106697)	specific
	Locus12 (9:3194279-3194422)	eGPVX-OM1	unsigned	A (9:≈4554048)	specific
	Locus13 (9:11798813-11799894)	eGPV1-OM8	eGPV1	A (9:14820313-14820314)	specific
	Locus14 (11:7758179-7759163)	eGPV1-OM9, eGPV1-OM10	eGPV1	A (11:≈13741299)	specific
<i>O. punctata</i>	Locus1 (1:36462406-36468381)	eGPV1-OP1	eGPV1	A (1:≈34032877)	specific
	Locus2 (2:18173741-18182535)	eGPV2-OP1, eGPV2-OP2, eGPV2-OP3, eGPV2-OP4, eGPV2-OP5, eGPV2-OP6, eGPV2-OP7, eGPV2-OP8	eGPV2	A (2:≈16229047)	specific
	Locus3 (3:1870791-1877899)	eGPV2-OP9, eGPV2-OP10	eGPV2	A (3:1950314-1952016)	specific
	Locus4 (4:11089082-11094038)	eGPV2-OP11	eGPV2	NA	undetermined
	Locus5 (7:14018987-14022570)	eGPV2-OP12, eGPV2-OP13, eGPV2-OP14	eGPV2	A (7:≈13073519)	specific
	Locus6 (7:14512232-14518685)	eGPV1-OP2	eGPV1	A (7:≈13488047)	specific
	Locus7 (9:2488988-2494897)	eGPV2-OP15	eGPV2	NA	undetermined
	Locus8 (9:2508244-2515543)	eGPV2-OP16, eGPV2-OP17, eGPV2-OP18, eGPV2-OP19, eGPV2-OP20	eGPV2	NA	undetermined
	Locus9 (12:17239641-17250240)	eGPV1-OP3, eGPV1-OP4, eGPV2-OP21, eGPV2-OP22, eGPV2-OP23	eGPV1, 2	A (12:≈19505111)	specific

^aThe segments close to each other (distance < 5kb) in the chromosome are taken as a locus in the mapping analysis (locus positions are provided in the brackets).

^bNames of the segments within a locus.

^cClassification of the segments within a locus.

P, presence of an orthologous insertion at the corresponding locus (highlighted by orange background; positions of orthologous insertions are shown in the brackets);

A, absence of an orthologous insertion at the corresponding locus (predicted position or range of empty sites are shown in the brackets); NA, not available due to a mass of uncharacterized bases (missing data), or a large deletion or rearrangement at the corresponding locus.

Table S10 Primers used in this study

1. Primers for orthologous eRTBVLs screening			
Target	Primer name	Primer sequence (from 5' to 3')	Remark
Orthologous JaE1-3	JaE1-3-1	GGACTCTGCATTCACTGTTCAA	P1 (forward)
	JaE1-3-4	TAGGATCAGCCTAACTACACCGT	P4 (reverse)
Orthologous JaE1-4	JaE1-4-1	CGTGTATTTAAAGAGACACATTCAACAATTA	P1 (forward)
	JaE1-4-3	GTGCAAGAGGCTAGAAACAAGA	P3 (forward)
	JaE1-4-4	TATCATACTGTGGCATCATGGGTTC	P4 (reverse)
Orthologous JaE1-7	Ch1-6-1	AGACCCTAAATTTAAATCAAACCTCAC	P1 (forward)
	JaE1-7-2	ATTCTTGGACTTTATCCGGCAC	P2 (reverse)
	Ch1-6-3	GTAAACTAAATAGTGTGTTTTGAACAGTG	P3 (forward)
	Ch1-6-4	GACGGCTGCTACAGAAACCAGA	P4 (reverse)
Orthologous JaE2-2	Ch2-2-1	GCTGCTTCTGATGCAACTACCACA	P1 (forward)
	JaE2-2-2	CTTTAGATGATGTAAAGTCCTACTCGG	P2 (reverse)
	JaE2-2-3	CAAAGGGATACGAGTAGGGGTG	P3 (forward)
	JaE2-2-4	TGATGCCGTCCATCACCTCAT	P4 (reverse)
Orthologous JaE2-3	JaE2-3-1	ATGAGGTGATGGACGGCATCA	P1 (forward)
	JaE2-3-2	CTCGTATCTAGGGGTGGACA	P2 (reverse)
	Ch2-2-3	GTTTGCTACTTCTGCAATATCTTATAATC	P3 (forward)
	JaE2-3-4	GCTGTTCAATTTGATTTCTTTCCCA	P4 (reverse)
Orthologous JaE3-1	Ch3-1-1	GTTAAGAGCGCGAGCAGTAT	P1 (forward)
	Ch3-1-2	GGCTATTCAATATTGATAACATGTCTG	P2 (reverse)
Orthologous JaE3-2	JaE3-2-1	GTCAGCCGCTTCTCTATCTAA	P1 (forward)
	JaE3-2-2	CCGAAGGTAGGTAAGTGTATATC	P2 (reverse)
	JaE3-2-3	GATATGACAGTTACCTACCTTCGG	P3 (forward)
	Ch3-1-4	TTAGATAGGAAGCGGCTGAC	P4 (reverse)
Orthologous JaE4-1	JaE4-1-1	GTACCAAAGCTAACCTACTTTGTC	P1 (forward)
	JaE4-1-2	AGTATAAGTGTGTATATATGGGGCAC	P2 (reverse)
	JaE4-1-3	CAATACTGTTCAATACTGTTAATTCTGATTCTA	P3 (forward)
	JaE4-1-4	TGTTGCTTGCCATTTTATTGAAC	P4 (reverse)
Orthologous JaE4-3	Ch4-3-1	CACTGCCTTGATGCGAAGT	P1 (forward)
	Ch4-3-2	GGCTCTGATACCACCTTTATGATG	P2 (reverse)
	Ch4-3-3	CCCAAAAGATAGATGCTAAATTTGATAG	P3 (forward)
	Ch4-3-4	GCAAGATGACAATTTGGTGACATAG	P4 (reverse)
Orthologous JaE4-4	Ch4-4-3	GAACAGTAATTTGAACAGTAACTGAATAGTAAT	P3 (forward)
	Ch4-4-4	GGGTCACCTATGCGTAACT	P4 (reverse)
Orthologous JaE5-1	Ch5-1-1	GAGAGGAAAACACAGGGTTC	P1 (forward)
	Ch5-1-4	TGGTGCCAGATAGTTAGCAG	P4 (reverse)
Orthologous JaE5-5	Ch5-5-1	GCACGTGATTAGGCTAAGTTC	P1 (forward)
	Ch5-5-2	GCAACCAGATGAATAGAATTCTTTAAGCTA	P2 (reverse)
	Ch5-5-3	TTTATGCTTTATAGGGTGTCTTCTGGGTT	P3 (forward)
	Ch5-5-4	CTGATAGGGCCCATAGATTG	P4 (reverse)
Orthologous JaE6-2	JaE6-2-1	GAGAGCTTGCCATGCATCATGA	P1 (forward)
	JaE6-2-2	GTCTGAAGTATAAGTGGTATTTTCTAC	P2 (reverse)
	JaE6-2-3	ATGCTACAACCTGCTAATCAATATATTCTATC	P3 (forward)
	JaE6-2-4	GCTTATCTGTTTAAATGAAGCAAGCT	P4 (reverse)
Orthologous JaE6-6	JaE6-6-1	CCTATGTTCTTTCAACGCATACCGA	P1 (forward)
	JaE6-6-2	GCTCTGATACCACCTTACTAGGACT	P2 (reverse)
	JaE6-6-3	AGGTAGTAGATGAGAAATTAGAAGTTTC	P3 (forward)
	JaE6-6-4	ACTGCAGTAGAACCTCCAGTG	P4 (reverse)
Orthologous JaE6-7	JaE6-7-1	CACGCATGGTTATGAACTGTATGC	P1 (forward)
	JaE6-7-2	TCCTGTTCTTAAGTCCACTAGGAT	P2 (reverse)
	JaE6-7-3	TGTTCTTGAAATATTGTATATTGTCTGAAGTATAG	P3 (forward)
	JaE6-7-4	CATCCTAAACCAAAGCACGAGGT	P4 (reverse)
Orthologous JaE7-4	Ch7-5-1	AGGAAGTGCCACTACATGGT	P1 (forward)
	Ch7-5-2	CTGAACAGATAAACTTGCATACAAGC	P2 (reverse)

	Ch7-5-3	CATACAAGCTTTCTTCTTACCTTAC	P3 (forward)
	Ch7-5-4	ATTGATGCATATATAACATCGGCTCG	P4 (reverse)
Orthologous JaE7-5	JaE7-5-1	CATCAAATTCCTCATCATAAGCC	P1 (forward)
	JaE7-5-4	CACCACAGTGAAGCAATATGTGGAT	P4 (reverse)
Orthologous JaE7-6	Ch7-6-1	GTAAAAGGGACCGTGGA	P1 (forward)
	Ch7-6-2	GGTCAGGACAGATTGTGAATGTA	P2 (reverse)
	Ch7-6-3	TAACTCTAAAATGAGCGGATTGAAAC	P3 (forward)
	Ch7-6-4	CTCTATCTCCCAACCCAAGA	P4 (reverse)
Orthologous JaE8-19	Ch8-15-1	GCGTATTTCCGGTATGGTAGT	P1 (forward)
	Ch8-15-2	CAGTAACTAAATAGTGTGTTTGAACAGTG	P2 (reverse)
	Ch8-15-3	GAGCGCGCTTGAAGCTTTATATTT	P3 (forward)
	Ch8-15-4	CTCATGAATCGACCCCTCCATTG	P4 (reverse)
Orthologous JaE10-4	JaE10-4-1	GATTTGAAATCTGAGAGCGATTGG	P1 (forward)
	JaE10-4-2	CTATAGACGCTATGGAACCTCG	P2 (reverse)
	JaE10-4-3	CCACCCTGCAAAAGATCCTAAAG	P3 (forward)
	JaE10-4-4	GTGCTACAATGCTTCGAGTTAC	P4 (reverse)
Orthologous JaE10-7	JaE10-7-1	GGATGCAAAGTGTGATCGTCA	P1 (forward)
	JaE10-7-2	GTTGGCCTTACCATTGTGCTT	P2 (reverse)
	JaE10-7-3	TAATTAATTCACAATAAAACCACGAAGAC	P3 (forward)
	JaE10-7-4	CCTGGAGAGGAGGAATCTTTC	P4 (reverse)
Orthologous JaE11-2	JaE11-2-1	TATCACACGCTGTCTCCCTA	P1 (forward)
	JaE11-2-2	GGAATTTGAAATGTAATATCCGTC	P2 (reverse)
	JaE11-2-3	CAGCAGAAGCCTACTCCC	P3 (forward)
	JaE11-2-4-2	GGATATTTACTTTAGTGTCCGGTG	P4 (reverse)
Orthologous JaE11-5	Ch11-4-1	GGTTGGCCCTTCTTCCTTTAT	P1 (forward)
	Ch11-4-2	ATGCCAGTTGCAGAAAATATTGAGAC	P2 (reverse)
	Ch11-4-3	CTCATCTATTAATATCTAGTCTCATTGATG	P3 (forward)
	Ch11-4-4	GCGAAGGCTACGCTACTAAC	P4 (reverse)
Orthologous JaE11-6	Ch11-5-1	CTCACCACATAGGTGCATTG	P1 (forward)
	Ch11-5-2	GGTGTGTTGTATGTCCTCTTATACTT	P2 (reverse)
	Ch11-5-3	AGTATTGAACAGTAATCTCGTTTACCA	P3 (forward)
	Ch11-5-4	CGGCTTTGATAGCACTGTAG	P4 (reverse)
Orthologous JaE12-6	JaE12-6-1	CCATTCTCGTTAACTAGTTTGCCAT	P1 (forward)
	JaE12-6-2	GGTAAGGCATACGGTGAATACC	P2 (reverse)
	JaE12-6-3	ACGATCAGTCGATTATGGCCTG	P3 (forward)
	JaE12-6-4	CTAAGCGTTGAGGCGATAAC	P4 (reverse)
Orthologous InE2-2	InE2-2-3	TTTCCTCAGCTGGGCAAGC	P3 (forward)
	InE2-2-4	CGGGACCATCCAAGAAGATGA	P4 (reverse)
Orthologous InE7-1	InE7-1-1	GCCGGAACCAAATAACAGTC	P1 (forward)
	InE7-1-2	AGATGCAATAAGAATATGAACTTCC	P2 (reverse)
	InE7-1-3	GTGGTTGTGTCTCCGCA	P3 (forward)
	InE7-1-4	TGATGACTGCCCTATCCCTGA	P4 (reverse)
Orthologous InE10-1	InE10-1-1	CCACCCCAATAGGATGAAGA	P1 (forward)
	InE10-1-2B	ATCGGTATCCGTAGAGAAGTAT	P2 (reverse)
	InE10-1-4	TGGAGTAGGCCATTTCTTGG	P4 (reverse)
Orthologous InE11-2	InE11-2-3	CCAACTCAACTCGACAAC	P3 (forward)
	InE11-2-4B	ACAGTATATCGCTCACTCTCATC	P4 (reverse)
Orthologous InE11-4	InE11-4-1	GCCTGATTGCTAGGATGGAGTTA	P1 (forward)
	InE11-4-2	GAGTGCCAACTATGTTCTTCAGATA	P2 (reverse)
Orthologous RuE42	RuE-46-1	ATCGATGCACTCTGATCAGC	P1 (forward)
	RuE-46-4	CGCAGCTAGTCTTCCCAACT	P4 (reverse)
Orthologous RuE44	RuE-48-1	CTGGAGAAGACCATGATTG	P1 (forward)
	RuE-48-2	CTATAGGAGAAACCTCTTTGG	P2 (reverse)
	RuE-48-3	CTATTTCTTTAATCTTTCATTCTAC	P3 (forward)
	RuE-48-4	TGGATGTGAAGGAACACAAC	P4 (reverse)
Orthologous RuE48	RuE-51-1	GATGGTAAACACAGTCTGACC	P1 (forward)
	RuE-51-2	CAACTCCGGCGTAATTCATTG	P2 (reverse)
	RuE-51-3	GTATTGATTCCAAGCAAAGCTTCC	P3 (forward)
	RuE-51-4	ACAACCATGACCACGAGGAG	P4 (reverse)

Orthologous RuE49	RuE-52-1	CCTGGAAATCTTCATAGCACC	P1 (forward)
	RuE-52-2	TTTGAATAGTAAATAGTAATCAGCCG	P2 (reverse)
	RuE-52-4	AGACTACAGGCTGCTACAGC	P4 (reverse)
2. Primers for eRTBVL-X cluster detection			
Target	Primer name	Primer sequence (from 5' to 3')	Remark
JaE8-5	JaE8-5-1-2	TCACTCAGACAATATACCACACTTG	Forward
	JaE8-5-4	GAAGACTACTCAAGAAGGAATACAG	Reverse
JaE8-8	JaE8-8-1	GAGTGAAGCATAATAGTAATCAGCCG	Forward
	JaE8-8-2	CGCTCTGATATACGAGCACTTATC	Reverse
	JaE8-8-3	GTGTTTTGAACAAAAGTATAGAATAAGAGA	Forward
	JaE8-8-4	TCAGAACAAAACTGATACCACTC	Reverse
JaE8-9	JaE8-9-1	GTCCGAATAGTAAGTGAGAATATTG	Forward
	JaE8-9-2	ACGCTCTGATATACGAACCTTATCAGA	Reverse
JaE8-10	JaE8-10-1	GTGGTATCAGAATATTGTCCGAATAGTA	Forward
	JaE8-10-2	CCACTCTTTTAAACAGTATTTAAACC	Reverse
	JaE8-10-3	CTGAACAGTACTGAACAGTATTCAGC	Forward
	JaE8-10-4	AGTGCATGCCAGCCAATAGCATAA	Reverse
JaE8-11	JaE8-11-3	GTAATCAGCCGAATAAATAGTGTTTTTG	Forward
	JaE8-11-4	TGTCGGGTTTTGGGACACTTATCA	Reverse
JaE8-13	JaE8-13-1	CTTAGAAGTACCCAATCTTGAATATATAAG	Forward
	JaE8-13-2	CTCTGATACCACTCTTCTCAGA	Reverse
	JaE8-13-3	CTTGAATATATAAGATAAGAGTGGTACC	Forward
	JaE8-13-4	CACCTATCAGAAAATATACGCTCTGA	Reverse
JaE8-15	JaE8-15-3	CTCTGATACCACTCTTCTCAGA	Forward
	JaE8-15-4	TGTCGTCAATATACAGTAAGCGT	Reverse
Terminals	JaE8-5-L-1	GATGCCACCAAGTCGTATTAG	Forward
	JaE8-16-4	TGCCGGATGGATTTCAGTAAGAG	Reverse
3. Primers for sequencing the eRTBVL-X cluster gap			
Target	Primer name	Primer sequence (from 5' to 3')	Remark
Terminals	Gap-F	GCCTCCCCACAGTTAAAGAA	Forward
	Gap-R	TGACTGGCTTTCTATTATGACTGC	Reverse
Middle	Gap2-F	CGTGAATCCACAGTAGCGAA	Forward
	Gap2-R	GATACCGATTCTCCAGACGC	Reverse
4. Primers for orthologous eGPV1-OS/eGPV2-OS			
Target	Primer name	Primer sequence (from 5' to 3')	Remark
Orthologous eGPV1-OS9	eRPV1-st6-1b	TGGTATGTGGCAATTGGATGG	P1 (forward)
	eRPV1-st6-2b	ATTGCAAGGCAGCAAAAGAT	P2 (reverse)
	eRPV1-st6-3	CCCTTTTGCTGTTTCGTTGTTGGC	P3 (forward)
	eRPV1-st6-4	TGCTCACCTCCCCACCTCGC	P4 (reverse)
Orthologous eGPV2-OS10 & eGPV2-	eRPV2-st6-1a	CCACTGGGTATCAGCGTTCAATC	P1 (forward)
	eRPV2-st6-2	CCTATATAAGGGGCTCAGGAACAC	P2 (reverse)
	eRPV2-st6-3	GAAGACTTATGAAAGTCAAGTGTGAG	P3 (forward)
	eRPV2-st6-4	GCATATCTTGCAAGAAAATCGTAAGTC	P4 (reverse)
Orthologous eGPV1-OS5	eRPV1-ga2-1	TTGAGGGGCGATTCAATTAG	P1 (forward)
	eRPV1-ga2-2	GGCTCACCATCACACTTCGG	P2 (reverse)
Orthologous eGPV2-OS3 & eGPV2-OS4	eRPV2-ga1-3	CACTGGTTGTCTCCCATTCTATAG	P3 (forward)
	eRPV2-ga1-4	CGTTCGTTTCGCTGTAAATCAG	P4 (reverse)
Orthologous eGPV1-OS8	eRPV1-st5-1	GCAATTAGGCTCCACGAAGGATTTA	P1 (forward)
	eRPV1-st5-4	GTTGGCTCCTGGCCCTTCACAC	P4 (reverse)

References

- Almeyda CV, Eid SG, Saar D, Samuitiene M, Pappu HR (2014). Comparative analysis of endogenous plant pararetroviruses in cultivated and wild *Dahlia* spp. *Virus Genes* 48(1): 140-152.
- Ashby M, Warry A, Bejarano E, Khashoggi A, Burrell M, Lichtenstein C (1997). Analysis of multiple copies of geminiviral DNA in the genome of four closely related *Nicotiana* species suggest a unique integration event. *Plant Molecular Biology* 35: 313 - 321.
- Aswad A, Katzourakis A (2012). Paleovirology and virally derived immunity. *Trends in Ecology & Evolution* 27(11): 627-636.
- Becher H, Ma L, Kelly LJ, Kovarik A, Leitch IJ, Leitch AR (2014). Endogenous pararetrovirus sequences associated with 24 nt small RNAs at the centromeres of *Fritillaria imperialis* L. (Liliaceae), a species with a giant genome. *The Plant Journal* 80: 823-833.
- Bejarano E, Khashoggi A, Witty M, Lichtenstein C (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proceedings of the National Academy of Sciences of the United States of America* 93: 759-764.
- Belyi VA, Levine AJ, Skalka AM (2010). Unexpected inheritance: Multiple integrations of ancient Bornavirus and Ebolavirus/Marburgvirus sequences in vertebrate genomes. *PLoS Pathogens* 6(7): e1001030.
- Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, Jenkins J, Barry K, Lindquist E, Hellsten U, Deshpande S, Wang X, Wu X, Mitros T, Triplett J, Yang X, Ye C-Y, Mauro-Herrera M, Wang L, Li P, Sharma M, Sharma R, Ronald PC, Panaud O, Kellogg EA, Brutnell TP, Doust AN, Tuskan GA, Rokhsar D, Devos KM (2012). Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology* 30: 555-561.
- Bertsch C, Beuve M, Dolja V, Wirth M, Pelsy F, Herrbach E, Lemaire O (2009). Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biology Direct* 4(1): 21.
- Bill CA, Summers J (2004). Genomic DNA double-strand breaks are targets for hepadnaviral DNA integration. *Proceedings of the National Academy of Sciences of the United States of America* 101(30): 11135-11140.
- Boeke J, Corces VG (1989). Transcription and reverse transcription of retrotransposons. *Annual Reviews in Microbiology* 43(1): 403-434.
- Botstein D (1980). A theory of modular evolution for bacteriophages. *Annals of the New York Academy of Sciences* 354(1): 484-491.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10(1): 421.

Catalán P, Kellogg EA, Olmstead RG (1997). Phylogeny of Poaceae subfamily Pooideae based on chloroplastndhf gene sequences. *Molecular Phylogenetics and Evolution* 8(2): 150-166.

Chabannes M, Iskra-Caruana M-L (2013). Endogenous pararetroviruses - a reservoir of virus infection in plants. *Current Opinion in Virology* 3(6): 615-620.

Chen J, Huang Q, Gao D, Wang J, Lang Y, Liu T, Li B, Bai Z, Luis Goicoechea J, Liang C, Chen C, Zhang W, Sun S, Liao Y, Zhang X, Yang L, Song C, Wang M, Shi J, Liu G, Liu J, Zhou H, Zhou W, Yu Q, An N, Chen Y, Cai Q, Wang B, Liu B, Min J, Huang Y, Wu H, Li Z, Zhang Y, Yin Y, Song W, Jiang J, Jackson SA, Wing RA, Wang J, Chen M (2013). Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nature Communications* 4: 1595.

Chessa B, Pereira F, Arnaud F, Amorim A, Goyache F, Mainland I, Kao RR, Pemberton JM, Beraldi D, Stear MJ, Alberti A, Pittau M, Iannuzzi L, Banabazi MH, Kazwala RR, Zhang Y-p, Arranz JJ, Ali BA, Wang Z, Uzun M, Dione MM, Olsaker I, Holm L-E, Saarma U, Ahmad S, Marzanov N, Eythorsdottir E, Holland MJ, Ajmone-Marsan P, Bruford MW, Kantanen J, Spencer TE, Palmarini M (2009). Revealing the history of sheep domestication using retrovirus integrations. *Science* 324(5926): 532-536.

Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, Kanematsu S, Suzuki N (2011). Widespread endogenization of genome sequences of non-retroviral rna viruses into plant genomes. *PLoS Pathogens* 7(7): e1002146.

Crooks GE, Hon G, Chandonia J-M, Brenner SE (2004). WebLogo: A Sequence Logo Generator. *Genome Research* 14(6): 1188-1190.

Darriba D, Taboada GL, Doallo R, Posada D (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8): 1164-1165.

Darriba D, Taboada GL, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* 9(8): 772-772.

dela Paz JS, Stronghill PE, Douglas SJ, Saravia S, Hasenkampf CA, Riggs CD (2012). Chromosome fragile sites in *Arabidopsis* harbor matrix attachment regions that may be associated with ancestral chromosome rearrangement events. *PLoS Genetics* 8(12): e1003136.

Du J, Tian Z, Bowen NJ, Schmutz J, Shoemaker RC, Ma J (2010). Bifurcation and enhancement of autonomous-nonautonomous retrotransposon partnership through LTR swapping in soybean. *The Plant Cell* 22(1): 48-61.

- Durkin SG, Glover TW (2007). Chromosome fragile sites. *Annual Review of Genetics* 41(1): 169-192.
- Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5): 1792-1797.
- Emerman M, Malik HS (2010). Paleovirology—modern consequences of ancient viruses. *PLoS biology* 8(2): e1000301.
- Fütterer J, Potrykus I, Bao Y, Li L, Burns TM, Hull R, Hohn T (1996). Position-dependent ATT initiation during plant pararetrovirus rice tungro bacilliform virus translation. *Journal of Virology* 70(5): 2999-3010.
- Feschotte C, Gilbert C (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews Genetics* 13(4): 283-296.
- Feschotte C, Jiang N, Wessler SR (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics* 3(5): 329-341.
- Froissart R, Roze D, Uzest M, Galibert L, Blanc S, Michalakakis Y (2005). Recombination every day: abundant recombination in a virus during a single multi-cellular host infection. *PLoS biology* 3(3): e89.
- Fujino K, Horie M, Honda T, Merriman DK, Tomonaga K (2014). Inhibition of Borna disease virus replication by an endogenous bornavirus-like element in the ground squirrel genome. *Proceedings of the National Academy of Sciences of the United States of America* 111(36): 13175–13180.
- Fuller DQ, Qin L, Zheng Y, Zhao Z, Chen X, Hosoya LA, Sun G-P (2009). The domestication process and domestication rate in rice: Spikelet bases from the Lower Yangtze. *Science* 323(5921): 1607-1610.
- Gambley C, Geering A, Steele V, Thomas J (2008). Identification of viral and non-viral reverse transcribing elements in pineapple (*Ananas comosus*), including members of two new badnavirus species. *Archives of Virology* 153(8): 1599-1604.
- Gayral P, Blondin L, Guidolin O, Carreel F, Hippolyte I, Perrier X, Iskra-Caruana M-L (2010). Evolution of endogenous sequences of banana streak virus: What can we learn from banana (*Musa* sp.) evolution? *Journal of Virology* 84(14): 7346-7359.
- Gayral P, Iskra-Caruana M-L (2009). Phylogeny of Banana Streak Virus Reveals Recent and Repetitive Endogenization in the Genome of Its Banana Host (*Musa* sp.). *Journal of Molecular Evolution* 69(1): 65-80.
- Geering A, Scharaschkin T, Teycheney P-Y (2010). The classification and nomenclature of endogenous viruses of the family Caulimoviridae. *Archives of Virology* 155(1): 123-131.
- Geering ADW, Olszewski NE, Dahal G, Thomas JE, Lockhart BEL (2001). Analysis of the distribution and structure of integrated Banana streak virus DNA in a range of *Musa* cultivars. *Molecular Plant Pathology* 2(4): 311-321.

207-213.

Geering ADW, Olszewski NE, Harper G, Lockhart BEL, Hull R, Thomas JE (2005). Banana contains a diverse array of endogenous badnaviruses. *Journal of General Virology* 86(2): 511-520.

Geuking MB, Weber J, Dewannieux M, Gorelik E, Heidmann T, Hengartner H, Zinkernagel RM, Hangartner L (2009). Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science* 323(5912): 393-396.

Gifford R, Tristem M (2003). The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* 26(3): 291-315.

Gilbert C, Feschotte C (2010). Genomic fossils calibrate the long-term evolution of hepadnaviruses. *PLoS biology* 8(9): e1000495.

Gilbert C, Meik JM, Dashevsky D, Card DC, Castoe TA, Schaack S (2014). Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. *Proceedings of the Royal Society B: Biological Sciences* 281(1791).

Goldschmidt V, Ciuffi A, Ortiz M, Brawand D, Muñoz M, Kaessmann H, Telenti A (2008). Antiretroviral activity of ancestral TRIM5α. *Journal of Virology* 82(5): 2089-2096.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* 40(D1): D1178-D1186.

Gregor W, Mette MF, Staginnus C, Matzke MA, Matzke AJM (2004). A distinct endogenous pararetrovirus family in *Nicotiana tomentosiformis*, a diploid progenitor of polyploid tobacco. *Plant Physiology* 134(3): 1191-1199.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* 59(3): 307-321.

Hall T (2011). BioEdit: An important software for molecular biology. *GERF Bulletin of Biosciences* 2(1): 60-61.

Hansen CN, Harper G, Heslop-Harrison JS (2005). Characterisation of pararetrovirus-like sequences in the genome of potato (*Solanum tuberosum*). *Cytogenetic and Genome Research* 110(1-4): 559-565.

Harper G, Hart D, Moulton S, Hull R, Geering A, Thomas J (2005). The diversity of banana streak virus isolates in Uganda. *Archives of Virology* 150(12): 2407-2420.

Harper G, Hull R, Lockhart B, Olszewski N (2002). Viral sequences integrated into plant genomes. *Annual*

Review of Phytopathology 40: 119 - 136.

Harper G, Osuji JO, Heslop-Harrison JS, Hull R (1999). Integration of banana streak badnavirus into the *Musa* genome: Molecular and cytogenetic evidence. *Virology* 255(2): 207-213.

Hay J, Grieco F, Druka A, Pinner M, Lee S-C, Hull R (1994). Detection of rice tungro bacilliform virus gene products *in vivo*. *Virology* 205(2): 430-437.

Hay JM, Jones MC, Blakebrough ML, Dasgupta I, Davies JW, Hull R (1991). An analysis of the sequence of an infectious clone of rice tungro bacilliform virus, a plant pararetrovirus. *Nucleic Acids Research* 19(10): 2615-2621.

Hayward A, Grabherr M, Jern P (2013). Broad-scale phylogenomics provides insights into retrovirus–host evolution. *Proceedings of the National Academy of Sciences of the United States of America* 110(50): 20146-20151.

He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu C-I, Shi S (2011). Two evolutionary histories in the genome of rice: The roles of domestication genes. *PLoS Genetics* 7(6): e1002100.

Herzog E, Guerra-Peraza O, Hohn T (2000). The rice tungro bacilliform virus gene II product interacts with the coat protein domain of the viral gene III polyprotein. *Journal of Virology* 74(5): 2073-2083.

Hohn T (2013). Plant pararetroviruses: interactions of cauliflower mosaic virus with plants and insects. *Current Opinion in Virology* 3(6): 629-638.

Hohn T, Richert-Pöggeler K, Staginnus C, Harper G, Schwarzacher T, Teo C, Teycheney P-Y, Iskra-Caruana M-L, Hull R (2008). Evolution of integrated plant viruses. In: Roossinck M (ed) *Plant Virus Evolution*. Springer Berlin Heidelberg, pp 53-81.

Hohn T, Rothnie H (2013). Plant pararetroviruses: replication and expression. *Current Opinion in Virology* 3(6): 621-628.

Holmes EC (2011). The evolution of endogenous viral elements. *Cell Host & Microbe* 10(4): 368-377.

Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Jern P, Gojobori T, Coffin JM, Tomonaga K (2010). Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature* 463(7277): 84-87.

Hu W-S, Rhodes T, Dang Q, Pathak V (2003). Retroviral recombination: review of genetic analyses. *Frontiers in Bioscience* 8: 143-155.

Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu

Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490(7421): 497-501.

Hull R (1996). Molecular biology of rice tungro viruses. *Annual Review of Phytopathology* 34: 275 - 297.

Hull R (2002a). Chapter 8: Virus replication. In: Hull R (ed) *Matthews' Plant Virology (Fourth Edition)*. Academic Press: London, pp 293-372.

Hull R (2002b). Chapter 17: Variation, evolution and origins of plant viruses. In: Hull R (ed) *Matthews' Plant Virology (Fourth Edition)*. Academic Press: London, pp 743-812.

Hull R (2009). *Comparative plant virology*. Academic press.

Iskra-Caruana M-L, Baurens F-C, Gayral P, Chabannes M (2010). A four-partner plant-virus interaction: Enemies can also come from within. *Molecular Plant-Microbe Interactions* 23(11): 1394-1402.

Jackson JA, Trevino AV, Herzig MC, Herman TS, Woynarowski JM (2003). Matrix attachment region (MAR) properties and abnormal expansion of AT island minisatellites in FRA16B fragile sites in leukemic CEM cells. *Nucleic Acids Research* 31(21): 6354-6364.

Jakowitsch J, Mette MF, van der Winden J, Matzke MA, Matzke AJM (1999). Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proceedings of the National Academy of Sciences of the United States of America* 96(23): 13241-13246.

Jiang L, Liu L (2006). New evidence for the origins of sedentism and rice domestication in the Lower Yangzi River, China. *Antiquity* 80(308): 355-361.

Jiang N, Jordan IK, Wessler SR (2002). Dasheng and RIRE2. A nonautonomous long terminal repeat element and its putative autonomous partner in the rice genome. *Plant Physiology* 130(4): 1697-1705.

Johnson WE (2010). Endless forms most viral. *PLoS Genetics* 6(11): e1001210.

Katzourakis A (2013). Paleovirology: inferring viral evolution from host genome sequence data. *Philosophical Transactions of the Royal Society B: Biological Sciences* 368(1626): 20120493.

Katzourakis A, Gifford RJ (2010). Endogenous viral elements in animal genomes. *PLoS Genetics* 6(11): e1001191.

Khan M, Hibino H, Aguiro V, Daquioag R, Opina O (1991). Rice and weed hosts of rice tungro-associated viruses and leafhopper vectors. *Plant Disease* 75(9): 926-930.

Khush G (1997). Origin, dispersal, cultivation and variation of rice. In: Sasaki T and Moore G (eds) *Oryza: From Molecule to Plant*. Springer Netherlands, pp 25-34.

Kojima Y, Ebana K, Fukuoka S, Nagamine T, Kawase M (2005). Development of an RFLP-based rice diversity research set of germplasm. *Breeding Science* 55(4): 431-440.

Kovach MJ, Sweeney MT, McCouch SR (2007). New insights into the history of rice domestication. *Trends in Genetics* 23(11): 578-587.

Kunii M, Kanda M, Nagano H, Uyeda I, Kishima Y, 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(1): 80.

Laney AG, Hassan M, Tzanetakis IE (2012). An integrated badnavirus is prevalent in fig germplasm. *Phytopathology* 102(12): 1182-1189.

Lisch D (2009). Epigenetic regulation of transposable elements in plants. *Annual Review of Plant Biology* 60(1): 43-66.

Liu H, Fu Y, Jiang D, Li G, Xie J, Cheng J, Peng Y, Ghabrial SA, Yi X (2010). Widespread horizontal gene transfer from double-stranded rna viruses to eukaryotic nuclear genomes. *Journal of Virology* 84(22): 11876-11887.

Liu H, Fu Y, Li B, Yu X, Xie J, Cheng J, Ghabrial S, Li G, Yi X, Jiang D (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evolutionary Biology* 11(1): 276.

Liu R, Kishima Y (2014). Chapter 12: Establishment of endogenous pararetroviruses in the rice genome. In: Gaur RK, Hohn T and Sharma P (eds) *Plant Virus–Host Interaction*. Academic Press: Boston, pp 229-240.

Liu R, Koyanagi KO, Chen S, Kishima Y (2012). Evolutionary force of AT-rich repeats to trap genomic and episomal DNAs into the rice genome: lessons from endogenous pararetrovirus. *The Plant Journal* 72(5): 817-828.

Londo JP, Chiang Y-C, Hung K-H, Chiang T-Y, Schaal BA (2006). Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proceedings of the National Academy of Sciences of the United States of America* 103(25): 9578-9583.

Lu T, Yu S, Fan D, Mu J, Shangguan Y, Wang Z, Minobe Y, Lin Z, Han B (2008). Collection and comparative analysis of 1888 full-length cDNAs from wild rice *Oryza rufipogon* Griff. W1943. *DNA Research* 15(5): 285-295.

Magiorkinis EN, Magiorkinis GN, Paraskevis DN, Hatzakis AE (2005). Re-analysis of a human hepatitis B virus (HBV) isolate from an East African wild born Pan troglodytes schweinfurthii: Evidence for interspecies recombination between HBV infecting chimpanzee and human. *Gene* 349(0): 165-171.

Maori E, Lavi S, Mozes-Koch R, Gantman Y, Peretz Y, Edelbaum O, Tanne E, Sela I (2007). Isolation and characterization of Israeli acute paralysis virus, a dicistrovirus affecting honeybees in Israel: evidence for diversity due to intra- and inter-species recombination. *Journal of General Virology* 88(12): 3428-3438.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuve P (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics* 26(19): 2462-2463.

Matsuyama A, Shiraishi T, Trapasso F, Kuroki T, Alder H, Mori M, Huebner K, Croce CM (2003). Fragile site orthologs FHIT/FRA3B and Fhit/Fra14A2: Evolutionarily conserved but highly recombinogenic. *Proceedings of the National Academy of Sciences of the United States of America* 100(25): 14988-14993.

Mekuria TA, Gutha LR, Martin RR, Naidu RA (2009). Genome Diversity and Intra- and Interspecies Recombination Events in Grapevine fanleaf virus. *Phytopathology* 99(12): 1394-1402.

Mette M, Kanno T, Aufsatz W, Jakowitsch J, van der Winden J, Matzke M, Matzke A (2002). Endogenous viral sequences and their potential contribution to heritable virus resistance in plants. *The EMBO Journal* 21: 461 - 469.

Mishmar D, Rahat A, Scherer SW, Nyakatura G, Hinzmann B, Kohwi Y, Mandel-Gutfroind Y, Lee JR, Drescher B, Sas DE, Margalit H, Platzer M, Weiss A, Tsui L-C, Rosenthal A, Kerem B (1998). Molecular characterization of a common fragile site (FRA7H) on human chromosome 7 by the cloning of a simian virus 40 integration site. *Proceedings of the National Academy of Sciences of the United States of America* 95(14): 8141-8146.

Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, Amarasinghe V, Youens-Clark K, Thomason J, Preece J, Pasternak S, Olson A, Jiao Y, Lu Z, Bolser D, Kerhornou A, Staines D, Walts B, Wu G, D'Eustachio P, Haw R, Croft D, Kersey PJ, Stein L, Jaiswal P, Ware D (2014). Gramene 2013: comparative plant genomics resources. *Nucleic Acids Research* 42(D1): D1193-D1199.

Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC (2006). Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucleic Acids Research* 34(suppl 1): D731-D735.

Ndowora T, Dahal G, LaFleur D, Harper G, Hull R, Olszewski N, 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, Richert-Pöggeler KR (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. *The Plant Journal* 50(2): 219-229.

Odell JT, Nagy F, Chua N-H (1985). Identification of DNA sequences required for activity of the cauliflower mosaic virus 35S promoter. *Nature* 313(6005): 810-812.

- Oka H-I (1988). *Origin of cultivated rice*. Japan Scientific Societies Press.
- Ozeri-Galai E, Lebofsky R, Rahat A, Bester Assaf C, Bensimon A, Kerem B (2011). Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites. *Molecular Cell* 43(1): 122-131.
- Pahalawatta V, Druffel K, Pappu H (2008). A new and distinct species in the genus *Caulimovirus* exists as an endogenous plant pararetroviral sequence in its host, *Dahlia variabilis*. *Virology* 376(2): 253-257.
- Patel MR, Emerman M, Malik HS (2011). Paleovirology —ghosts and gifts of viruses past. *Current Opinion in Virology* 1(4): 304-309.
- Peng Z, Lu Y, Li L, Zhao Q, Feng Q, Gao Z, Lu H, Hu T, Yao N, Liu K, Li Y, Fan D, Guo Y, Li W, Lu Y, Weng Q, Zhou C, Zhang L, Huang T, Zhao Y, Zhu C, Liu X, Yang X, Wang T, Miao K, Zhuang C, Cao X, Tang W, Liu G, Liu Y, Chen J, Liu Z, Yuan L, Liu Z, Huang X, Lu T, Fei B, Ning Z, Han B, Jiang Z (2013). The draft genome of the fast-growing non-timber forest species moso bamboo (*Phyllostachys heterocycla*). *Nature Genetics* 45: 456-461.
- Qu R, Bhattacharyya M, Laco GS, De Kochko A, Subba Rao BL, Kaniewska MB, Scott Elmer J, Rochester DE, Smith CE, Beachy RN (1991). Characterization of the genome of rice tungro bacilliform virus: Comparison with commelina yellow mottle virus and caulimoviruses. *Virology* 185(1): 354-364.
- Rajeswaran R, Golyaev V, Seguin J, Zvereva A, Farinelli L, Pooggin M (2014). Interactions of rice tungro bacilliform pararetrovirus and its protein P4 with plant RNA silencing machinery. *Molecular Plant-Microbe Interactions* 27(12): 1370-1378.
- Rebollo R, Romanish MT, Mager DL (2012). Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annual Review of Genetics* 46(1): 21-42.
- Richert-Poggeler KR, Noreen F, Schwarzacher T, Harper G, Hohn T (2003). Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *The EMBO Journal* 22(18): 4836-4845.
- Rohlf F (2000). NTSYS-pc version 2.1: numerical taxonomic and multivariate analysis system. *New York, Exeter Software*.
- Roossinck MJ (1997). Mechanisms of plant virus evolution. *Annual Review of Phytopathology* 35: 191-209.
- Roy A, Shao J, Schneider WL, Hartung JS, Brilansky RH (2014). Population of endogenous pararetrovirus genomes in Carrizo citrange. *Genome Announcements* 2(1): e01063-01013.
- Rudd S, Frisch M, Grote K, Meyers BC, Mayer K, Werner T (2004). Genome-wide *in silico* mapping of scaffold/matrix attachment regions in *Arabidopsis* suggests correlation of intragenic scaffold/matrix

attachment regions with gene expression. *Plant Physiology* 135(2): 715-722.

Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM (2014). Long non-coding RNAs as a source of new peptides. *eLife* 3: e03523.

Söding J, Biegert A, Lupas AN (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* 33(suppl 2): W244-W248.

Sawyer SL, Wu LI, Emerman M, Malik HS (2005). Positive selection of primate TRIM5 α identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the United States of America* 102(8): 2832-2837.

Schibler M, Gerlach D, Martinez Y, Van Belle S, Turin L, Kaiser L, Tapparel C (2012). Experimental human rhinovirus and enterovirus interspecies recombination. *Journal of General Virology* 93(1): 93-101.

Scholthof HB (2005). Plant virus transport: motions of functional equivalence. *Trends in Plant Science* 10(8): 376-382.

Schwartz M, Zlotorynski E, Goldberg M, Ozeri E, Rahat A, Sage CI, Chen BPC, Chen DJ, Agami R, Kerem B (2005). Homologous recombination and nonhomologous end-joining repair pathways regulate fragile site stability. *Genes & Development* 19(22): 2715-2726.

Seal S, Turaki A, Muller E, Kumar PL, Kenyon L, Filloux D, Galzi S, Lopez-Montes A, Iskra-Caruana M-L (2014). The prevalence of badnaviruses in West African yams (*Dioscorea cayenensis-rotundata*) and evidence of endogenous pararetrovirus sequences in their genomes. *Virus Research* 186: 144-154.

Sether DM, Melzer MJ, Borth WB, Hu JS (2012). Pineapple bacilliform CO virus: Diversity, detection, distribution, and transmission. *Plant Disease* 96(12): 1798-1804.

Sharma S, Rabindran R, Robin S, Dasgupta I (2011). Analysis of the complete DNA sequence of rice tungro bacilliform virus from southern India indicates it to be a product of recombination. *Archives of Virology* 156(12): 2257-2262.

Smith GR, Borg Z, Lockhart BEL, Braithwaite KS, Gibbs MJ (2000). Sugarcane yellow leaf virus: a novel member of the Luteoviridae that probably arose by inter-species recombination. *Journal of General Virology* 81(7): 1865-1869.

Staginnus C, Gregor W, Mette MF, Teo C, Borroto-Fernandez E, Machado ML, Matzke M, Schwarzacher T (2007). Endogenous pararetroviral sequences in tomato (*Solanum lycopersicum*) and related species. *BMC Plant Biol* 7(1): 24.

Staginnus C, Iskra-Caruana M, Lockhart B, Hohn T, Richert-Pöggeler K (2009). Suggestions for a nomenclature of endogenous pararetroviral sequences in plants. *Archives of Virology* 154(7): 1189-1193.

Staginnus C, Richert-Pöggeler KR (2006). Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends in Plant Science* 11(10): 485-491.

Su L, Gao S, Huang Y, Ji C, Wang D, Ma Y, Fang R, Chen X (2007). Complete genomic sequence of Dracaena mottle virus, a distinct badnavirus. *Virus Genes* 35(2): 423-429.

Swigoňová Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J (2004). Close split of sorghum and maize genome progenitors. *Genome Research* 14(10a): 1916-1923.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28(10): 2731-2739.

Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, Aono R, Fujii Y, Habara T, Harada E, Kanno M, Kawahara Y, Kawashima H, Kubooka H, Matsuya A, Nakaoka H, Saichi N, Sanbonmatsu R, Sato Y, Shinso Y, Suzuki M, Takeda J, Tanino M, Todokoro F, Yamaguchi K, Yamamoto N, Yamasaki C, Imanishi T, Okido T, Tada M, Ikeo K, Tateno Y, Gojobori T, Lin YC, Wei FJ, Hsing YI, Zhao Q, Han B, Kramer MR, McCombie RW, Lonsdale D, O'Donovan CC, Whitfield EJ, Apweiler R, Koyanagi KO, Khurana JP, Raghuvanshi S, Singh NK, Tyagi AK, Haberer G, Fujisawa M, Hosokawa S, Ito Y, Ikawa H, Shibata M, Yamamoto M, Bruskiewich RM, Hoen DR, Bureau TE, Namiki N, Ohyanagi H, Sakai Y, Nobushima S, Sakata K, Barrero RA, Sato Y, Souvorov A, Smith-White B, Tatusova T, An S, An G, S OO, Fuks G, Fuks G, Messing J, Christie KR, Lieberherr D, Kim H, Zuccolo A, Wing RA, Nobuta K, Green PJ, Lu C, Meyers BC, Chaparro C, Piegu B, Panaud O, Echeverria M (2008). The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* 36(Database issue): D1028-1033.

Tanne E, Sela I (2005). Occurrence of a DNA sequence of a non-retro RNA virus in a host plant genome and its expression: evidence for recombination between viral and host RNAs. *Virology* 332(2): 614-622.

Taylor D, Bruenn J (2009). The evolution of novel fungal genes from non-retroviral RNA viruses. *BMC Biology* 7(1): 88.

Temin HM (1985). Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. *Molecular Biology and Evolution* 2(6): 455-468.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22(22): 4673-4680.

Thorland EC, Myers SL, Gostout BS, Smith DI (2002). Common fragile sites are preferential targets for HPV16 integrations in cervical tumors. *Oncogene* 22(8): 1225-1237.

Torruella M, Gordon K, Hohn T (1989). Cauliflower mosaic virus produces an aspartic proteinase to cleave

its polyproteins. *The EMBO Journal* 8(10): 2819.

Umber M, Filloux D, Muller E, Laboureau N, Galzi S, Roumagnac P, Iskra-Caruana M-L, Pavis C, Teycheney P-Y, Seal SE (2014). The genome of African yam (*Dioscorea cayenensis-rotundata* complex) hosts endogenous sequences from four distinct badnavirus species. *Molecular Plant Pathology* 15(8): 790-801.

Vigne E, Marmonier A, Fuchs M (2008). Multiple interspecies recombination events within RNA2 of Grapevine fanleaf virus and Arabis mosaic virus. *Archives of Virology* 153(9): 1771-1776.

Wei X, Qiao W-H, Chen Y-T, Wang R-S, Cao L-R, Zhang W-X, Yuan N-N, Li Z-C, Zeng H-L, Yang Q-W (2012). Domestication and geographic origin of *Oryza sativa* in China: insights from multilocus analysis of nucleotide variation of *O. sativa* and *O. rufipogon*. *Molecular Ecology* 21(20): 5073-5087.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* 8(12): 973-982.

Xiong Y, Eickbush TH (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *The EMBO Journal* 9(10): 3353.

Yang C-c, Kawahara Y, Mizuno H, Wu J, Matsumoto T, Itoh T (2012). Independent domestication of Asian rice followed by gene flow from *japonica* to *indica*. *Molecular Biology and Evolution* 29(5): 1471-1479.

Yang ZN, Ye XR, Molina J, Roose ML, Mirkov TE (2003). Sequence analysis of a 282-kilobase region surrounding the citrus tristeza virus resistance gene (*Ctv*) locus in *Poncirus trifoliata* L. Raf. *Plant Physiology* 131(2): 482-492.

Yasaka R, Nguyen HD, Ho SY, Duchene S, Korkmaz S, Katis N, Takahashi H, Gibbs AJ, Ohshima K (2014). The temporal evolution and global spread of cauliflower mosaic virus, a plant pararetrovirus. *PLoS One* 9(1): e85641.

Yozwiak NL, Skewes-Cox P, Gordon A, Saborio S, Kuan G, Balmaseda A, Ganem D, Harris E, DeRisi JL (2010). Human Enterovirus 109: a Novel Interspecies Recombinant Enterovirus Isolated from a Case of Acute Pediatric Respiratory Illness in Nicaragua. *Journal of Virology* 84(18): 9047-9058.

Yu S, Mangelsdorf M, Hewett D, Hobson L, Baker E, Eyre HJ, Lapsys N, Le Paslier D, Doggett NA, Sutherland GR, Richards RI (1997). Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell* 88(3): 367-374.

Zhang G, Liu X, Quan Z, Cheng S, Xu X, Pan S, Xie M, Zeng P, Yue Z, Wang W, Tao Y, Bian C, Han C, Xia Q, Peng X, Cao R, Yang X, Zhan D, Hu J, Zhang Y, Li H, Li H, Li N, Wang J, Wang C, Wang R, Guo T, Cai Y, Liu C, Xiang H, Shi Q, Huang P, Chen Q, Li Y, Wang J, Zhao Z, Wang J (2012). Genome sequence of foxtail

millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. *Nature Biotechnology* 30(6): 549-554.

Zhang H, Freudenreich CH (2007). An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in *S. cerevisiae*. *Molecular Cell* 27(3): 367-379.

Zhang H, Zhu J-K (2011). RNA-directed DNA methylation. *Current Opinion in Plant Biology* 14(2): 142-147.

Zhao W, Wang J, He X, Huang X, Jiao Y, Dai M, Wei S, Fu J, Chen Y, Ren X, Zhang Y, Ni P, Zhang J, Li S, Wang J, Wong GKS, Zhao H, Yu J, Yang H, Wang J (2004). BGI - RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Research* 32(suppl 1): D377-D382.

Zheng X-M, Ge S (2010). Ecological divergence in the presence of gene flow in two closely related *Oryza* species (*Oryza rufipogon* and *O. nivara*). *Molecular Ecology* 19(12): 2439-2454.

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007). Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: Severe bottleneck during domestication of rice. *Molecular Biology and Evolution* 24(3): 875-888.

Zhu T, Xu P-Z, Liu J-P, Peng S, Mo X-C, Gao L-Z (2014). Phylogenetic relationships and genome divergence among the AA- genome species of the genus *Oryza* as revealed by 53 nuclear genes and 16 intergenic regions. *Molecular Phylogenetics and Evolution* 70: 348-361.

Zinn E, Vandenberghe LH (2014). Adeno-associated virus: fit to serve. *Current Opinion in Virology* 8: 90-97.

Zong Y, Chen Z, Innes JB, Chen C, Wang Z, Wang H (2007). Fire and flood management of coastal swamp enabled first rice paddy cultivation in east China. *Nature* 449(7161): 459-462.

Zou X-H, Yang Z, Doyle JJ, Ge S (2013). Multilocus estimation of divergence times and ancestral effective population sizes of *Oryza* species and implications for the rapid diversification of the genus. *New Phytologist* 198(4): 1155-1164.

Zou X-H, Zhang F-M, Zhang J-G, Zang L-L, Tang L, Wang J, Sang T, Ge S (2008). Analysis of 142 genes resolves the rapid diversification of the rice genus. *Genome Biology* 9(3): R49.

Acknowledgements

First and foremost, I would like to express my sincere and deep gratitude to my advisor Professor Yuji Kishima, Plant Breeding Laboratory, Hokkaido University, for the continuous guide and overwhelming support of my doctoral study, for his enthusiasm, patience, prompting, friendliness and profound knowledge. Professor Kishima is one of the best scientific teachers I have known. Besides my advisor, I would like to profusely thank Dr. Itsuro Takamure (Hokkaido University) for his kind advices, concern and encouragement, Professor Chikara Masuta (Hokkaido University) and Professor Kanako O. Koyanagi (Hokkaido University) for their insightful advices and comments. I also thank Professor Il-Ryong Choi (International Rice Research Institute), Professor Hideshi Yasui (Kyushu University) and Professor Shoshi Kikuchi (National Institute of Agrobiological Sciences) for their massive supports.

My sincere acknowledgements also go to Jaymee R. Encabo (International Rice Research Institute) and all the members of Plant Breeding Laboratory, especially Ms Ikue Nakano, Ms Naoko Tokuhashi and Ms Mayumi Oseko for administrative help; Ms Takako Takeuchi, Dr. Ruifang Liu, Seiya Ishiguro, Nozomi Saitou, Ryosuke Aiba and Hua Zhou for their great kindness, stimulating discussions and technical supports during my doctoral study. I am thankful to all the staffs of Hokkaido University I have met, and to my friends at Hokkaido University in the past three and half years.

I express my profound appreciation to China Scholarship Council for providing the opportunity and fellowship to my doctoral course. I express my gratitude to Clark Memorial Foundation at Hokkaido University for a doctoral student research grant. I also appreciate National Institute of Genetics, National Institute of Agrobiological Sciences and International Rice Research Institute for materials of *Oryza* species.

Last but not the least, I am heartfully grateful to my beloved wife Huizhen Zheng, mother Judi Liao, father Zhonggai Chen and sister Aiyang Chen for supporting me spiritually throughout my life.

Sunlu Chen
December, 2014