

Supplementary Information

A new inovirus from the human blood encodes proteins with nuclear subcellular localization

Nikolay Popgeorgiev^{1, a*}, Mart Krupovic², Julien Hiblot³, Laura Fancello⁴, Sonia Monteil-Bouchard⁵, and Christelle Desnues^{5, †}

TableS1: tBlastN searches against SRA, EST and TSA databases using RIP1 ORFs.

Organism name/ Biosample	Origin	Accession #	Sequencing technique	SRA, EST, TSA Accession #	Sequences related to RIP1 ORFs:	Number of sequences detected/ E-value
NCBI SRA metagenomic database						
<i>Homo sapiens</i> / Serum	Japan	SAMD00008984	Illumina	DRX000952	1,3,7,15	7
<i>Homo sapiens</i> /Nasopharyngeal swabs	Kazakhstan	SAMN36835585	Ion Torrent	SRX21246575, SRX21246574, SRX21246571, SRX21246569, SRX21246568, SRX21246565, SRX21246535, SRX21246557	1,3,5,9, 10,12,14	152
<i>Homo sapiens</i> /Stool	China	SAMN28109099	Illumina	SRX15540575, SRX15540587	3	>200
<i>Homo sapiens</i> /Amniotic fluids	China	SAMN34062693 SAMN34062696	Illumina	SRX19859627, SRX19859630	1,3	38
NCBI EST and TSA databases						
<i>Sus scrofa</i> / 4-Cell-Embryo	USA	SAMN00174820	cDNA	CN025072.1	1	8e ⁻¹⁹
<i>Schistosoma mansoni</i> / mixed pool	Brasil	SAMN00172663	cDNA	CF497648.1	1,2,3	5e ⁻⁴⁸
<i>Tupaia chinensis</i> / liver	China	SAMN00794552	TSA/Illumina	JU128735.1	3	2e ⁻⁴³

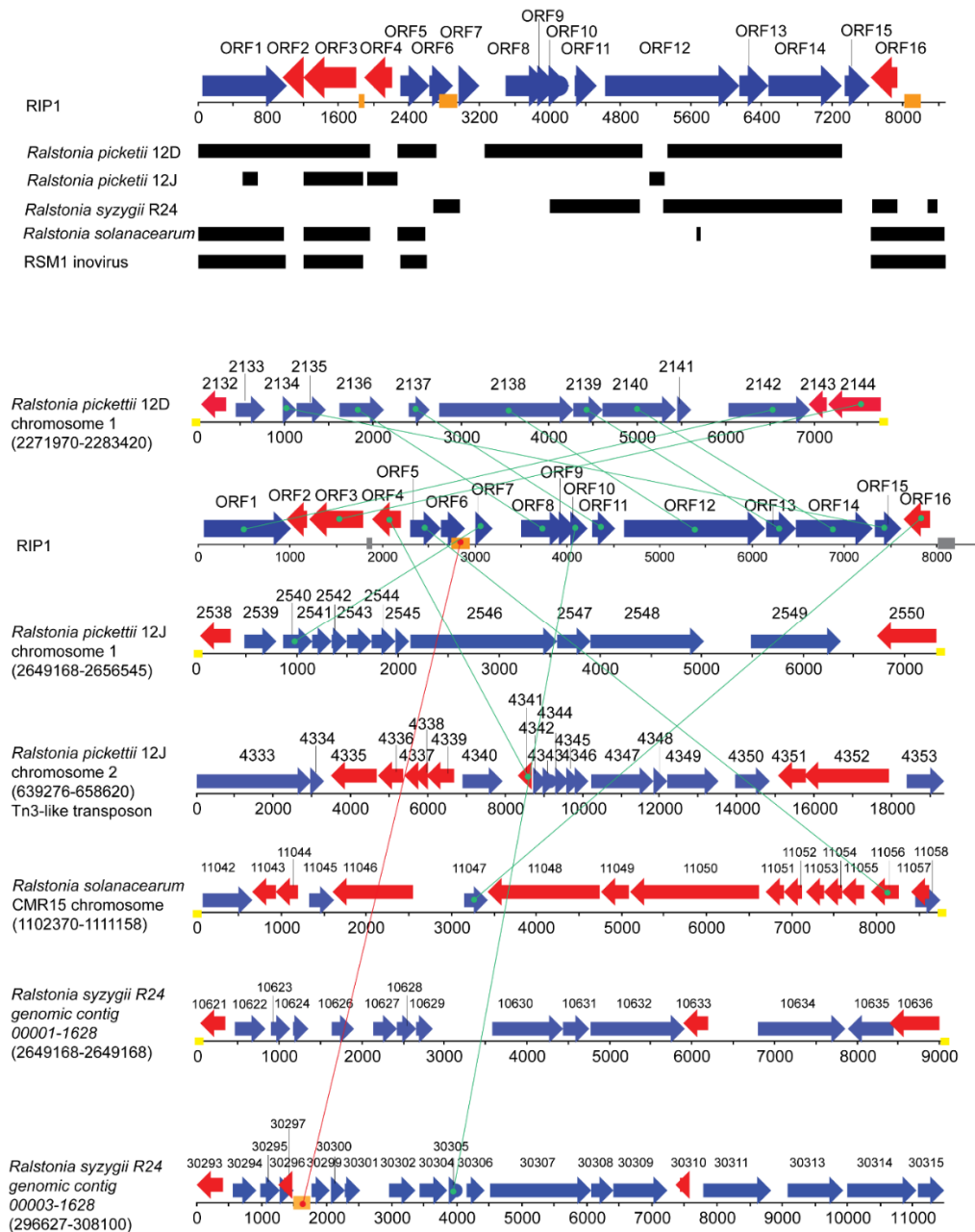


Figure S1: RIP1 has a chimeric genomic organization.

(A) Schematic representation of local alignment results between RIP1 genome and genomes of *Ralstonia pickettii* 12D, *Ralstonia syzygii* R24, *Ralstonia pickettii* 12J, *Ralstonia solanacearum* CRM15 and RSM1 phage. The matching regions as well as their locations on the RIP1 genome are reported using black dashes. (B) Genomic organization comparison between RIP1 genome and homologous bacterial regions. Forward/reverse ORFs are represented with blue/red arrows respectively. The presence of *attP* sites are represented with yellow boxes. Best Blast hits correspondence between RIP1 ORFs and bacterial ORFs are represented with colored lines.

The presence of *cis* elements are outlined with beige boxes. DNA sequence homologous to *Ralstonia syzyzii* R24 non coding region located inside ORF6 is reported with orange boxes.

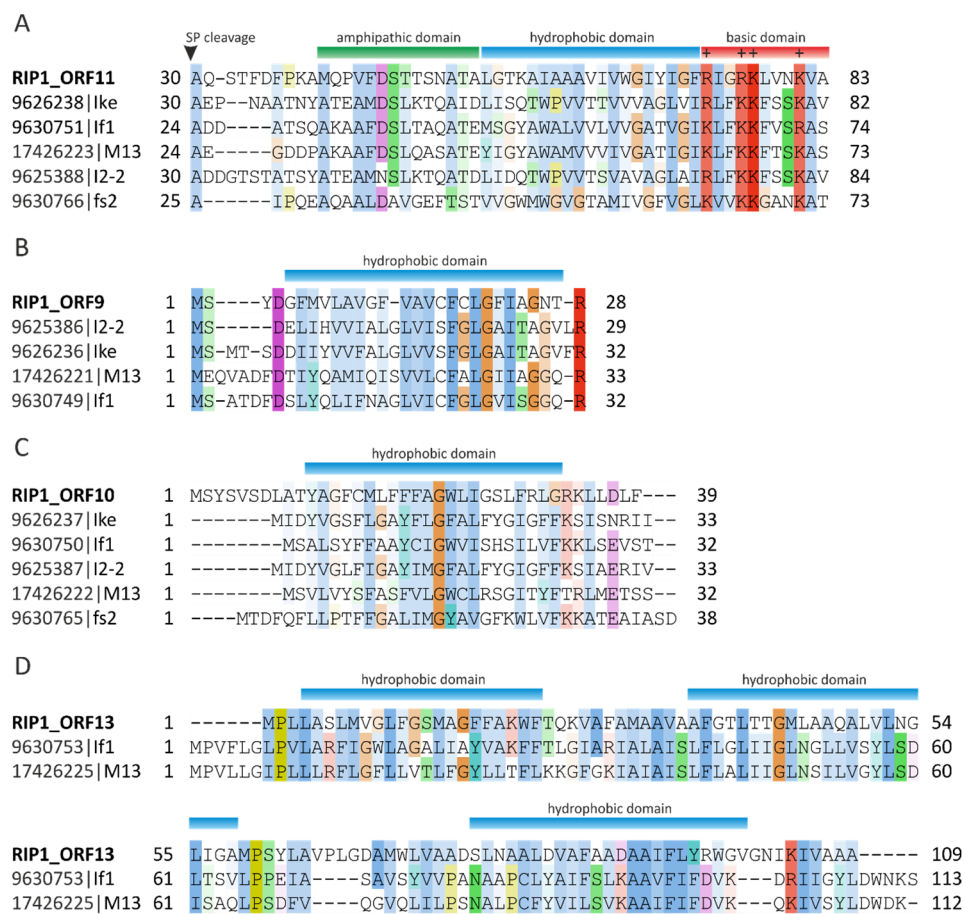


Figure S2: Multiple alignments between RIP1 ORFs and homologous inovirus proteins.

(A) Multiple sequence alignment between ORF9 and pVII homologs of inoviruses. The location of a single hydrophobic domain is represented with a blue bar. (B) Multiple sequence alignment between ORF10 and pIX-like proteins of inoviruses. (C) Multiple sequence alignment between ORF11 and pVIII-like inoviral homologs. The amphipathic, hydrophobic and basic domains are represented with green, blue and red bars, respectively. Positively charged residues are indicated with (+). (D) Multiple sequence alignment between ORF13 and pVI-like inoviral homologs. Abbreviations: SP, signal peptidase; Ike- Enterobacteria phage Ike; If1- Enterobacteria phage If1; I2-2- Enterobacteria phage I2-2; M13- Enterobacteria phage M13; fs2- Vibrio phage fs2. GenBank identifiers for the depicted proteins are provided in the figure.

		Motif I		Motif II		Motif III		
RIP1_ORF1	123	VT ¹²³ HM ¹²⁴ VT ¹²⁵ LT ¹²⁶ TRE	(34)	KRGAW ³⁷¹ HH ³⁷² MA ³⁷³ VAV ³⁷⁴ SGR	(37)	LDALAS ²³³ Y ²³⁴ IA ²³⁵ KY ²³⁶ IG ²³⁷ K	233/316	
117676352 RSM1	123	VT ¹²³ HM ¹²⁴ VT ¹²⁵ LT ¹²⁶ TRE	(34)	KRGAW ³⁷¹ HH ³⁷² MA ³⁷³ VAV ³⁷⁴ SGR	(37)	LDALAS ²³³ Y ²³⁴ IA ²³⁵ KY ²³⁶ IG ²³⁷ K	233/311	
209901325 RSM3	73	VT ⁷³ HM ⁷⁴ VT ⁷⁵ LT ⁷⁶ TRE	(34)	KRGAW ³⁷¹ HH ³⁷² MA ³⁷³ VAV ³⁷⁴ SGR	(37)	LDALAS ¹⁸³ Y ¹⁸⁴ IA ¹⁸⁵ QV ¹⁸⁶ HR ¹⁸⁷ Q	183/262	<i>Inoviridae</i>
156535144 p12J	129	LD ¹²⁹ HLL ¹³⁰ TL ¹³¹ TY ¹³² RG	(41)	QSGGI ³⁶¹ HI ³⁶² HV ³⁶³ GVR ³⁶⁴ GW	(36)	VHRVA ²⁴⁵ AY ²⁴⁶ LS ²⁴⁷ KY ²⁴⁸ IT ²⁴⁹ K	245/350	<i>(Inovirus)</i>
9626323 Pf3	57	HALAL ⁵⁷ TLL ⁵⁸ TV ⁵⁹ RD	(34)	RRGV ³⁶¹ PH ³⁶² LH ³⁶³ CA ³⁶⁴ AF ³⁶⁵ FD	(36)	AIGWF ¹⁶⁶ QY ¹⁶⁷ VS ¹⁶⁸ KH ¹⁶⁹ AAR	166/278	
75093009 Lf	155	AW ¹⁵⁵ YFL ¹⁵⁶ TL ¹⁵⁷ TY ¹⁵⁸ RD	(44)	QRFR ²³ PH ²⁴ YH ²⁵ VML ²⁶ WVP	(23)	ARNCV ²⁶¹ GY ²⁶² LAK ²⁶³ YAS ²⁶⁴ K	261/346	
45597172 SVTS2	93	NLS ⁹³ FLL ⁹⁴ TL ⁹⁵ TY ⁹⁶ AV	(45)	KRGAV ²⁴¹ HH ²⁴² FH ²⁴³ IIL ²⁴⁴ NQ ²⁴⁵ K	(24)	NEDV ²⁰¹ VK ²⁰² YL ²⁰³ AK ²⁰⁴ YIV ²⁰⁵ K	201/309	<i>Inoviridae</i>
74095490 SVGI13	96	NL ⁹⁶ GFT ⁹⁷ TL ⁹⁸ TY ⁹⁹ AE	(43)	QRGA ²⁴¹ VHH ²⁴² FH ²⁴³ IIF ²⁴⁴ SEY	(24)	NEFV ²⁰² SK ²⁰³ YVA ²⁰⁴ KY ²⁰⁵ VV ²⁰⁶ K	202/345	<i>(Plectrovirus)</i>
77020121 Chp4	81	QNC ⁸¹ FLL ⁸² TY ⁸³ ED	(40)	KLQR ³⁶¹ PH ³⁶² YH ³⁶³ LLI ³⁶⁴ YNY	(36)	TRQS ¹⁹⁶ AG ¹⁹⁷ TV ¹⁹⁸ ARY ¹⁹⁹ SL ²⁰⁰ K	196/315	
12085145 φMH2K	61	DNIF ⁶¹ LTL ⁶² TY ⁶³ DD	(46)	KTKR ³⁷¹ PH ³⁷² WH ³⁷³ VLI ³⁷⁴ FNF	(37)	TLD ¹⁸³ SAS ¹⁸⁴ YV ¹⁸⁵ ARY ¹⁸⁶ AAR	183/315	<i>Microviridae</i>
19424731 SpV4	72	HNW ⁷² FV ⁷³ TL ⁷⁴ TY ⁷⁵ SD	(47)	KRM ³⁷¹ RPH ³⁷² YH ³⁷³ IC ³⁷⁴ FF ³⁷⁵ NL	(37)	TYHS ¹⁹⁵ AN ¹⁹⁶ Y ¹⁹⁷ TAR ¹⁹⁸ YT ¹⁹⁹ TK	195/320	
226596530 HRPV-1	222	SGVM ²²² VTL ²²³ TT ²²⁴ DP	(49)	HIGL ⁴¹¹ PH ⁴¹² LHV ⁴¹³ CV ⁴¹⁴ FGV	(41)	GKSV ³⁵¹ AG ³⁵² YLG ³⁵³ KY ³⁵⁴ LS ³⁵⁵ K	351/500	<i>Pleolipoviridae</i>
394922815 HRPV-2	273	NAV ²⁷³ FCT ²⁷⁴ LT ²⁷⁵ TD ²⁷⁶ P	(67)	SAGY ¹¹² PH ¹¹³ LHV ¹¹⁴ LF ¹¹⁵ FDV	(112)	SKTAG ⁴⁹¹ SY ⁴⁹² VG ⁴⁹³ KY ⁴⁹⁴ IS ⁴⁹⁵ K	491/655	
394922851 HRPV-6	273	NAV ²⁷³ FCT ²⁷⁴ LT ²⁷⁵ TD ²⁷⁶ P	(67)	SAGY ¹¹² PH ¹¹³ LHV ¹¹⁴ LF ¹¹⁵ FDV	(112)	SKTAG ⁴⁹¹ SY ⁴⁹² VG ⁴⁹³ KY ⁴⁹⁴ IS ⁴⁹⁵ K	491/655	
16128214 E.coli	11	GTW ¹¹ FFT ¹² VN ¹³ LRN	(31)	VVLPE ⁴¹¹ HH ⁴¹² CI ⁴¹³ WT ⁴¹⁴ LP	(41)	YRHH ¹²² VD ¹²³ YI ¹²⁴ YN ¹²⁵ PV ¹²⁶ K	122/165	<i>IS200/IS605 (TnpA)</i>
145423933 PCV2	14	KRW ¹⁴ VFT ¹⁵ LNN ¹⁶ PS	(26)	EGRT ²⁴¹ PH ²⁴² LQG ²⁴³ FAN ²⁴⁴ FV	(24)	DQQN ¹⁰³ KE ¹⁰⁴ YCS ¹⁰⁵ KE ¹⁰⁶ GNL	103/314	
498907918 BCV	13	KRW ¹³ VFT ¹⁴ LNN ¹⁵ PS	(26)	EGKT ²⁴¹ PH ²⁴² LQG ²⁴³ FAN ²⁴⁴ FS	(24)	DKDN ¹⁰² QQ ¹⁰³ YCT ¹⁰⁴ KE ¹⁰⁵ GDV	102/313	<i>Circoviridae</i>
342365722 BFDV	9	RRW ⁹ CF ¹⁰ TL ¹¹ NN ¹² PT	(25)	EQGT ²³ PH ²⁴ LQ ²⁵ GY ²⁶ FH ²⁷ FK	(23)	DSDN ⁹⁶ EK ⁹⁷ YCS ⁹⁸ KE ⁹⁹ GNV	96/288	
81984152 FBNYV	5	ICW ⁵ CF ⁶ TL ⁷ NN ⁸ NPL	(19)	EAGNI ²³ HF ²⁴ QGY ²⁵ IEM ²⁶ K	(23)	QGEAR ⁸⁶ AY ⁸⁷ SM ⁸⁸ KED ⁸⁹ TR	86/286	<i>Nanoviridae</i>
166203210 BBTV	5	VCM ⁵ WFT ⁶ LNN ⁷ PA	(19)	EEGR ²³ TH ²⁴ VQGY ²⁵ VEM ²⁶ K	(23)	QGEAR ⁸⁶ AY ⁸⁷ CM ⁸⁸ KED ⁸⁹ TR	86/286	
20177462 MVDV	6	TCW ⁶ VFT ⁷ LNF ⁸ KG	(19)	RVGD ²⁴ HL ²⁵ LQ ²⁶ GFI ²⁷ QMK	(24)	SDQAK ⁸⁸ AY ⁸⁹ AM ⁹⁰ KED ⁹¹ TR	88/284	
20564209 TYLCV	11	KNY ¹¹ FLT ¹² YFK ¹³ CD	(29)	ENGE ³¹¹ PH ³¹² LH ³¹³ LI ³¹⁴ QFE	(31)	SSDV ¹¹⁰ KSY ¹¹¹ ID ¹¹² KD ¹¹³ GDV	110/359	
41057729 PYVMV	11	KNY ¹¹ FLT ¹² YFK ¹³ CS	(29)	TNGE ³¹¹ PH ³¹² LH ³¹³ VL ³¹⁴ MQFE	(31)	SSDV ¹¹⁰ KSY ¹¹¹ ID ¹¹² KD ¹¹³ GDT	110/354	<i>Geminiviridae</i>
306478726 TCTV	12	KNIF ¹² LT ¹³ YP ¹⁴ HCH	(29)	DNGE ³¹¹ PH ³¹² LH ³¹³ VLI ³¹⁴ QFE	(31)	SSDV ¹¹¹ KSY ¹¹² IE ¹¹³ KD ¹¹⁴ GDT	111/371	
9626995 MVM	25	FSF ²⁵ VFK ²⁶ NENV ²⁷ Q	(85)	KDQG ⁶⁷¹ WH ⁶⁷² CH ⁶⁷³ VLI ⁶⁷⁴ GGK	(67)	GNMIA ²¹⁶ Y ²¹⁷ Y ²¹⁸ FL ²¹⁹ TKK ²²⁰ KI	216/672	
109390400 MPV3	25	FSF ²⁵ VFK ²⁶ NEDV ²⁷ Q	(85)	KDQG ⁶⁷¹ WH ⁶⁷² CH ⁶⁷³ VLI ⁶⁷⁴ GGK	(67)	GNIV ²¹⁶ AY ²¹⁷ Y ²¹⁸ FL ²¹⁹ TKK ²²⁰ KI	216/672	<i>Parvoviridae</i>
29823071 Lull	25	FSF ²⁵ VFK ²⁶ NEDV ²⁷ Q	(85)	KDQG ⁶⁷¹ WH ⁶⁷² CH ⁶⁷³ VLI ⁶⁷⁴ GGK	(67)	GNMVA ²¹⁶ Y ²¹⁷ Y ²¹⁸ FL ²¹⁹ TKK ²²⁰ KI	216/668	

Figure S3: RIP1 ORF1 encodes a rolling-circle replication initiation protein.

Alignment of the RC-REP protein encoded by RIP1 with the corresponding proteins encoded by ssDNA viruses infecting bacteria (*Inoviridae* and *Microviridae*), archaea (*Pleolipoviridae*), and eukaryotes (*Circoviridae*, *Nanoviridae*, *Geminiviridae* and *Parvoviridae*) as well as TnpA transposase of the IS200/IS605 family from *E. coli*. The protein sequences are denoted by their GenBank identifiers, followed by the corresponding virus names. Only the three conserved motifs (I–III) typical of the RC-REPs are shown. The limits of the depicted motifs are indicated by the residue positions on each side of the alignment, while the number following the slash sign corresponds to the total length of the protein. The numbers in the parenthesis within the alignment indicate the distance between the motifs.

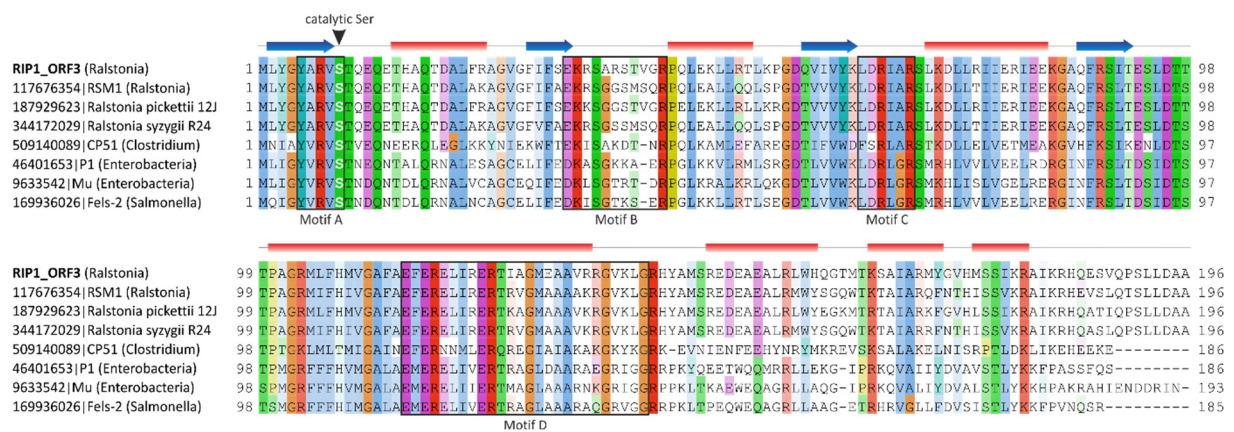


Figure S4: RIP1 ORF3 encodes a serine recombinase.

Alignment of the RIP1 ORF3 product sequence with the corresponding sequences of serine recombinases encoded by various prophages and bacteriophages. The four conserved motifs (A–D) typical of serine recombinases are boxed, while the position of the catalytic serine residue is indicated with an arrowhead. The predicted secondary structure elements (α -helices, red rectangles; β -strands, blue arrows) are shown above the alignment. Proviruses are indicated by the names of the organisms in which they reside, while the host names of the bacteriophages are provided in parenthesis. GenBank identifiers for all depicted proteins are provided in the figure.

TableS2: Summary table of primers used in this study.

Primer	Sequence (5'-3')	Used in
FlagRepF	ATATGGATCCATGGACTACAAGGACGACGATGACAAG ATGCGTGAAACTACGATAGCAG	This study
FlagRepR	ATATCTCGAGTCAACAAACACTGGCTATGGC	This study
FlagResF	ATATGGATCCATGGACTACAAGGACGACGATGACAAG ATGTTGTACGGTTACGCCCCG	This study
FlagResR	ATATCTCGAGTTACGCGGCGTCTAGCAAGC	This study
FlagZotF	ATATGGATCCATGGACTACAAGGACGACGATGACAAG ATGCCGATCTACATCATCACGG	This study
FlagZotR	ATATCTCGAGTCACCATGAGGGGCATCGAG	This study
Rp-F1	ATGATCTAGCTTGCTAGATTGAT	[21]
Rp-R1	ACTGATCGTCGCCTTGGTG	[21]
ZotIntF1	TGTGGATACGTTACGCTCG	This study
ZotIntR2	CGTTCGTCCAGTTTCGCTTC	This study
RP1F	CCCGCAAGCGCGCCAACGCT	This study
RP1R	ACGATCGGGATTGGTATCGC	This study
RP3F	AAACAGTGCGTCGGTTTGTG	This study
RP3R	ATCACGCGAGACGAACAAC	This study