

Evolutionary patterns of chimeric retrogenes in *Oryza* species

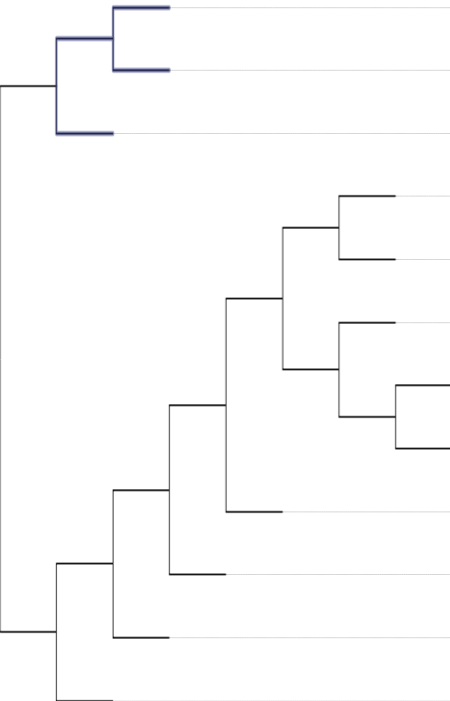
Yanli Zhou¹, Chengjun Zhang^{1,2*}

¹Germplasm Bank of Wild species, Kunming Institute of Botany, Chinese Academy of Sciences. No. 132 Lanhei Road, Kunming 650201, Yunnan, China

²Haiyan Engineering & Technology Center, Kunming Institute of Botany, Chinese Academy of Science. Jiaxing 314300, Zhejiang, China

*Correspondence author: Chengjun Zhang; email: zhangchengjun@mail.kib.ac.cn

Table S2. Species analysed in this study



ID	Species/subspecies	Short Name	Genome	IRGC ACC
1	<i>Oryza grandiglumis</i> (Doell Prod.)	<i>Grandi</i>	CCDD	105664
2	<i>Oryza alta</i> (Swallen)	<i>Alta</i>	CCDD	100967
3	<i>Oryza officinalis</i>	<i>OWR</i>	CC	/
4	<i>Oryza Sativa</i> L. <i>japonica</i>	<i>Japonica</i>	AA	/
5	<i>Oryza rufipogon</i> (Griff.)	<i>Rufi</i>	AA	80643
6	<i>Oryza Sativa</i> L. <i>indica</i>	<i>Indica</i>	AA	/
7	<i>Oryza nivara</i> (Sharma et Shastry)	<i>Nivara a</i>	AA	80622
8	<i>Oryza nivara</i> (Sharma et Shastry)	<i>Nivara b</i>	AA	80582
9	<i>Oryza glaberrima</i> (Steud.)	<i>Glab</i>	AA	103600
10	<i>Oryza longistaminata</i> (A. Chev. et Roehr)	<i>Longi</i>	AA	103886
11	<i>Oryza punctata</i>	<i>YSD8</i>	BB	/
12	<i>Oryza australiensis</i> (Domin)	<i>Austra</i>	EE	86530

Note: ‘/’ indicates the genomes of model species (*Oryza sativa* L. *indica* and *Oryza sativa* L. *japonica*) or wild accessions (*Oryza punctata* and *Oryza officinalis*) from the Wang laboratory were used as reference genomes. The phylogenetic tree on the left presents the phylogenetic relationships among the 12 analysed rice species.

Table S3. Details regarding the primers designed for PCR and sequencing

ID		Primer pair	primer	Product (bp)
<i>RCG1</i>	Chimerical retrogene	1	CGAGATTAACATTCTCATC GCTTCAGGTTACAGGTTAG	1260
		2	CAAAACAGCCGGATAGATAC GACATTGTCTCCCATCCGAG	1005
	Parental gene	3	GTCCTAGAAGAAGATGGTCG CCAAGATTCAACAGACAAG	1085
		4	CCAATTATCCAGCGATAGTG GACATTGTCTCCCATCCAAG	1196
<i>RCG2</i>	Chimerical retrogene	1	CTGGAAGGATGCATGGAATGG CTCAGTTCCTGTAGGGCCTG	1094
	Parental gene	2	GAACTCCAGTTTAAAGGTTCTG CACAGCTTCGAATTATCAACTC	1912
		3	CAGGCAGGTGAGGTTTCCTGG CTGAGTTTCTGTTCTGATGG	1332
<i>RCG3</i>	Chimerical retrogene	1	GCAGCGGTACATATTGATGG ACGCTGTAGACTCCATTGGG	1377
	Parental gene	2	TCATCCAACATGGGAAGGAG ACCTTGACCTCTCCACAAGC	1166
<i>RCG4</i>	Chimerical retrogene	1	TGCCATCCTCACTACGAAGAC TTATGTGGATCGTCTTCCGC	1155
	Parental gene	2	ACATGACGACTGCTTGATCG TCTCTCGTCATCAAGTGAGGG	1181
<i>RCG5</i>	Chimerical retrogene	1	ATTCAAGGAAAGCTGGGTTG GCGAAGAGACGAAGTTACCTG	1170
	Parental gene	2	TGGGAGATCTGGCAAAGATG GCACTGTCATCGGATCATTC	1165
<i>RCG6</i>	Chimerical retrogene	1	GGCACTTCACCGAAGGGTAG GGATCCGTTATCTCCTCTTCC	1087
	Parental gene	2	ATAGCGAGCAGGTCGGTTC ATAACAGCAGGGGCAAAGC	1257
<i>RCG7</i>	Chimerical retrogene	1	GGTTGAAAGGAAGAACCGC TAGCAACTGGTGCAAGGTTTC	966
	Parental gene	2	TTCGTCTCCATGTGTTCTCTC ACAGCATTGATCCACGATTC	881

Table S4 Chimerical retrogene and parental gene blat against IR8 genome

	Gene ID in Plant cell paper	Chimeric gene in 9311	Chimeric retrogene in IR8	Parental gene in 9311	Parental gene in IR8
<i>RCG1</i>	Chr03_4107	Chr03:27608262-27613159 (4897)	Chr03: 26049124.. 26054020 (4896)	Chr02:14775401-14780342 (4941)	Chr02: 14335789..4340666 (4877)
<i>RCG2</i>	Chr04_4524	Chr04:30664128-30669070 (4942)	Chr04: 31889774..31894673 (4899)	Chr04:30673408-30679130 (5722)	Chr04: 31713113..31894438 (543826)
<i>RCG3</i>	Chr12_904	Chr12:6133222-6139034 (5813)	Chr12: 5723488..5727995 <4507>	Chr11:19034040-19040398 (6359)	Chr12: 5801686..5807082 (5396)
<i>RCG4</i>	Chr10_2602	Chr10:17747410-17752061 (4652)	Chr10: 20792601..20812983 <3709>	Chr09:4376856-4381567 (4711)	Chr09: 5374301..5378751 (4450)
<i>RCG5</i>	Chr01_5436	Chr01:36521615-36526443 (4828)	Chr01:34682823..34687645 (4822)	Chr05:18541573-18546535 (4962)	Chr05:18326109..18330814 (4705)
<i>RCG6</i>	Chr02_1920	Chr02:12785385-12789823 (4438)	Chr02:12059249..12063688 (4439)	Chr07:9892475-9896854 (4379)	Chr07:10350983..10355365 (4382)
<i>RCG7</i>	Chr08_3454	Chr08:24470675-24475311 (4636)	Chr08:24715472..24720109 (4637)	Chr10:11013463-11018192 (4729)	Chr10:14428416..14432001 (3585)

The sequences of chimeric retrogenes and the corresponding parental genes were used as queries for a blast search of the *Indica* (IR8) rice genome, which was sequenced via PacBio technology. The output of the blast search is provided in round brackets. When the blast search output was too long, the sequences were shortened with gene-specific primers. The resulting data are provided in angle brackets.

Table S5 PCR sequencing of retrogenes and parental genes

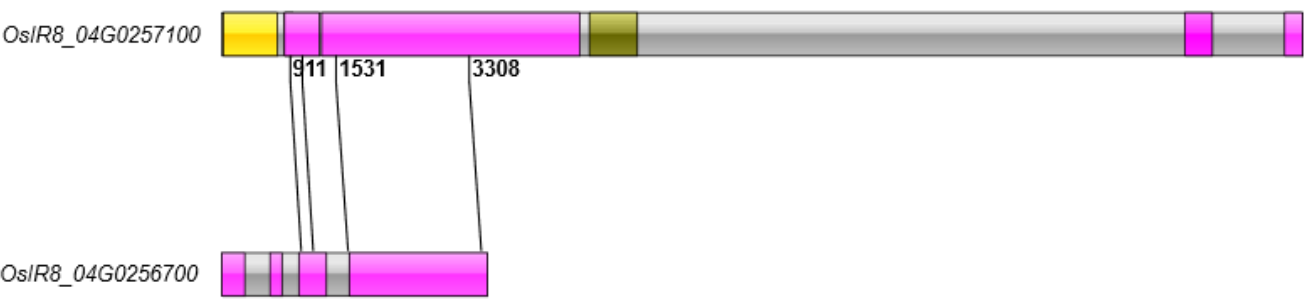
	<i>RCG1</i>		<i>RCG2</i>		<i>RCG3</i>		<i>RCG4</i>		<i>RCG5</i>		<i>RCG6</i>		<i>RCG7</i>	
	C	P	C	P	C	P	C	P	C	P	C	P	C	P
<i>Grandi</i>	x	x	x	7	x	x	x	15	x	8	x	na	x	5
<i>Longi</i>	1	5	5	6	3	x	7	14	5	x	1	x	3	10
<i>Alta</i>	x	x	x	3	x	5	8	16	x	7	x	na	1	6
<i>Austra</i>	7	x	x	x	x	x	x	x	2	x	x	x	x	8
<i>Rufi</i>	1	3	x	x	x	x	2	10	3	6	1	3	2	11
<i>Nivara a</i>	1	4	x	4	1	5	6	13	4	6	1	na	3	9
<i>Nivara b</i>	1	4	x	4	x	6	4	12	5	6	1	x	3	9
<i>Glab</i>	na	2	x	2	4	na	5	11	1	9	2	x	3	na
<i>Indica</i>	1	5	1	1	3	8	1&*	9	5	6	1	4*	4&*	7
<i>Japonica</i>	na	6	1	1	2	9	x	x	x	x	x	x	x	x
<i>YSD8</i>	na	x	x	x	x	x	x	x	x	x	x	x	x	x
<i>OWR</i>	x	na	x	x	x	x	x	x	x	x	x	x	x	x
<i>Total number of sequences</i>	7		7		9		16		9		4		9	

C, chimeric retrogene; P: parental gene; x: no amplicon; na: no high-quality sequence was obtained; *: *Indica* reference sequence was used; &: cloned sequence did not exactly match the 9311 reference sequence.

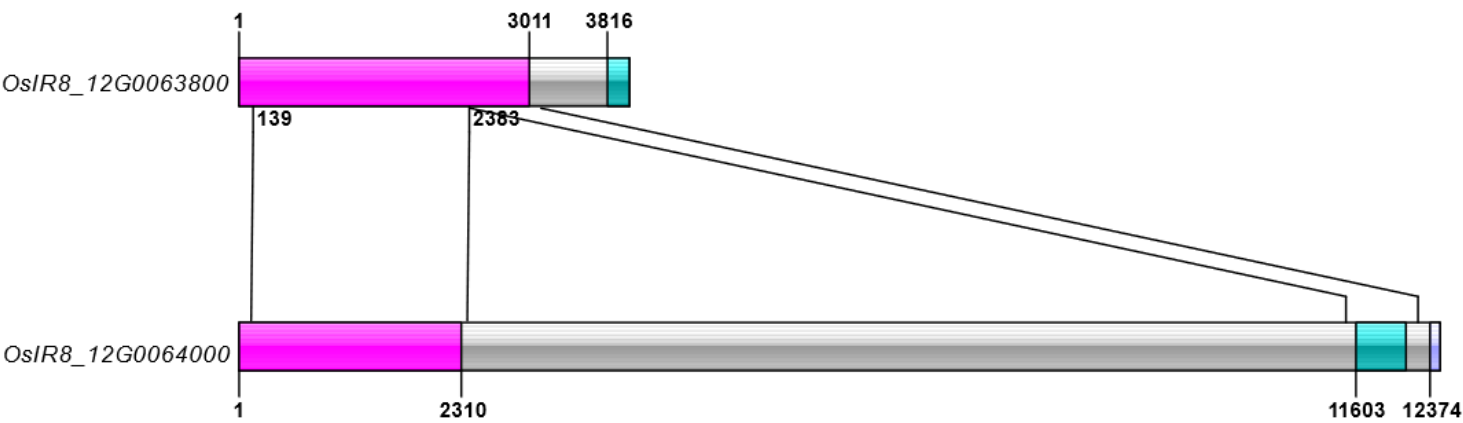
Total number of sequences refers to the number of sequence types used for analysing phylogeny, and corresponds to the maximum value in the C and P columns for each retrogene.

The same numbers in the two columns indicate a sequence type exactly same in one or several species, and correspond to the numbers in the phylogenetic tree presented in Figure 1.

AK106715



AK072107



AK105722

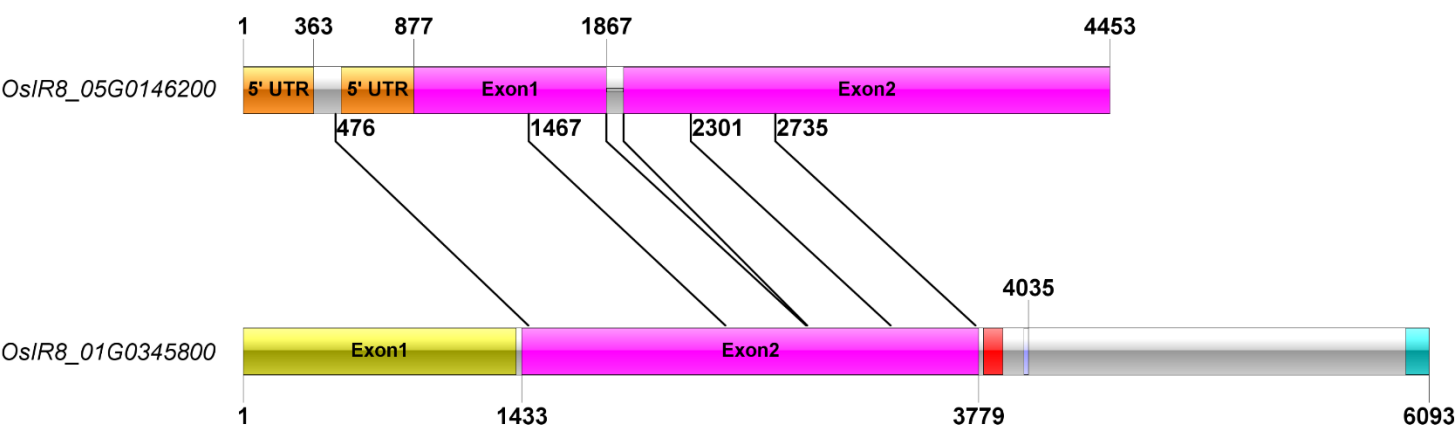
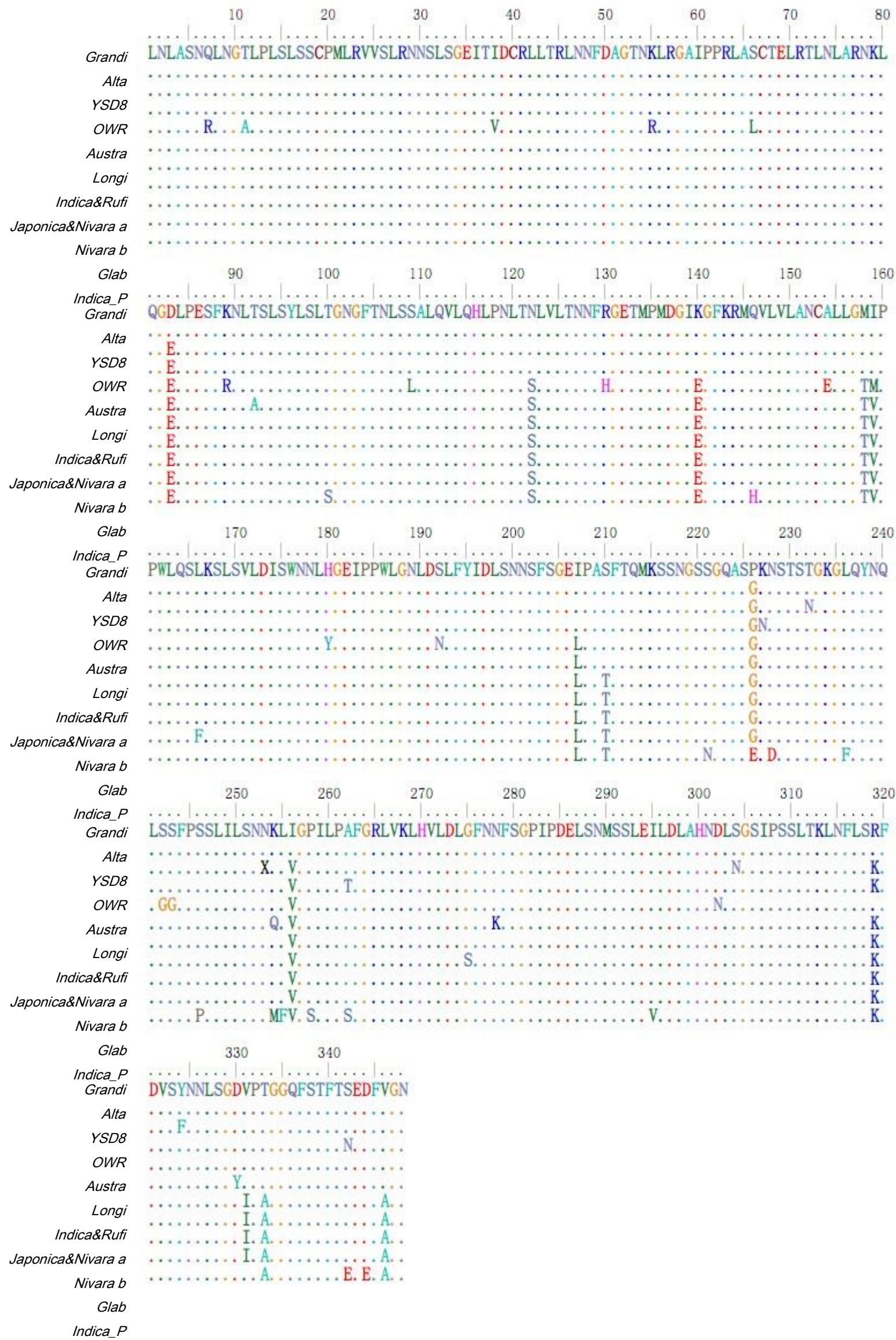
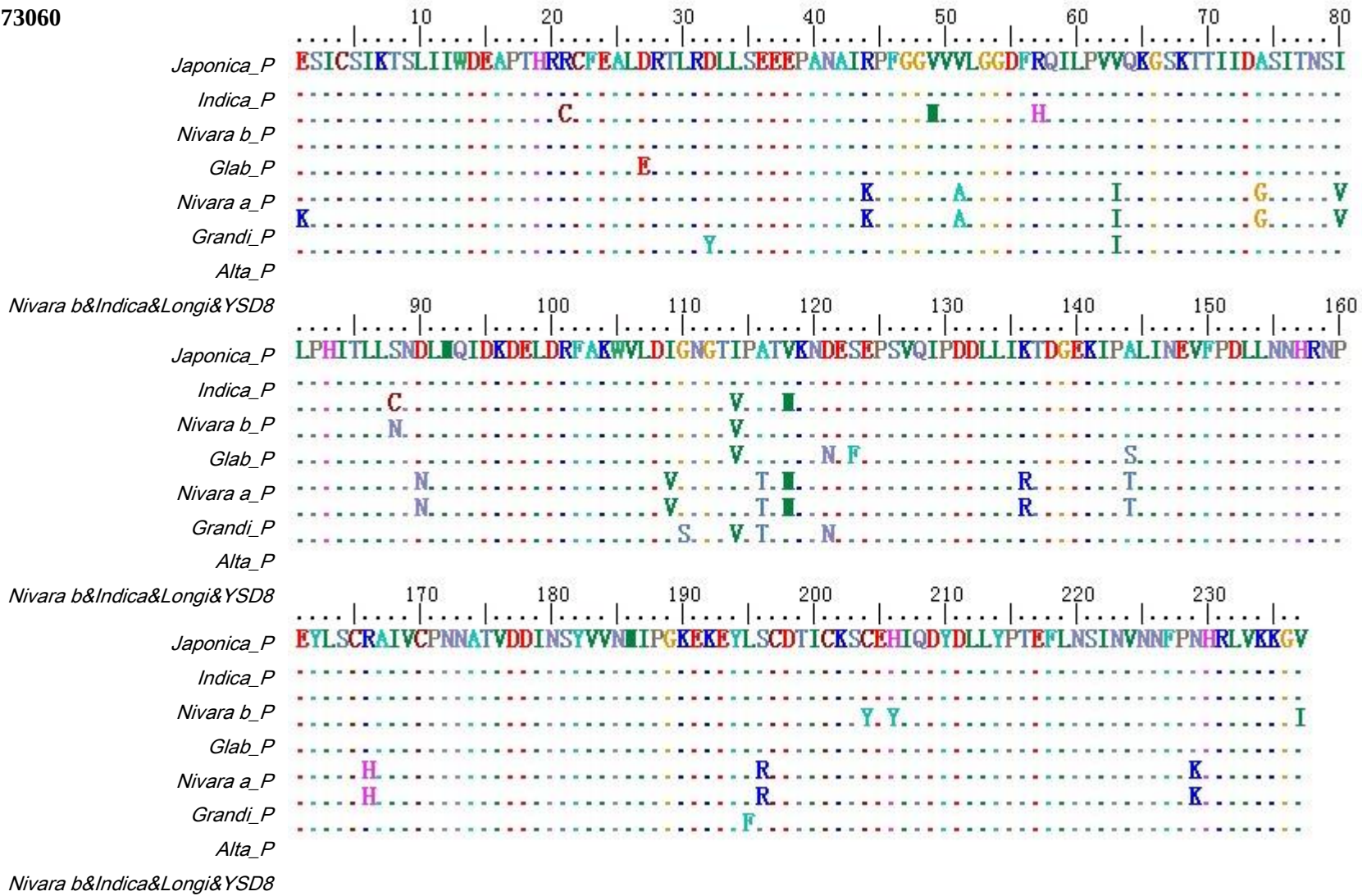


Figure S1. Chimeric retrogenes and the parental genes in *Indica* rice (IR8) genome. The sequences of the chimeric retrogenes and the corresponding parental genes were used as queries for blast searches of the IR8 genome, which was sequenced via PacBio technology.

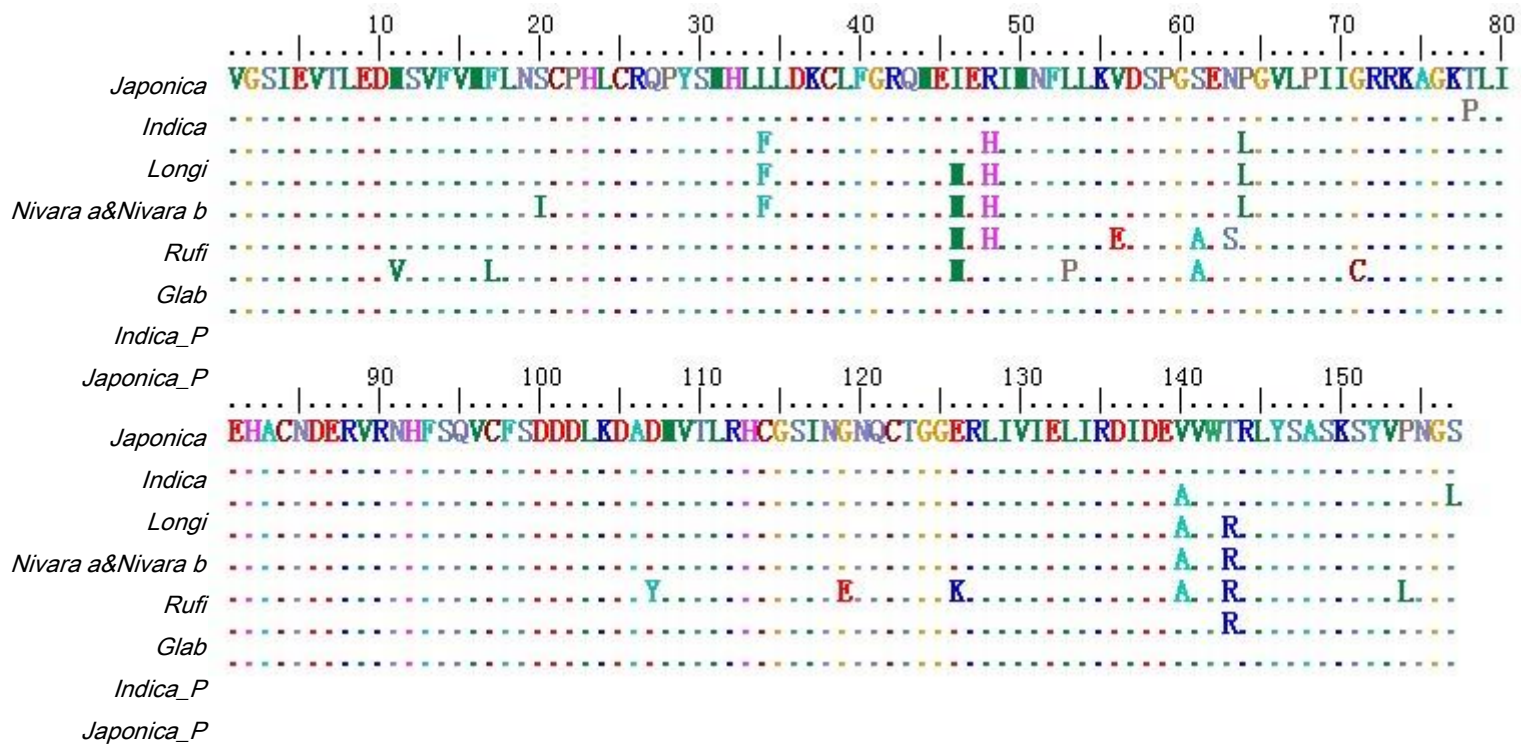
AK070283

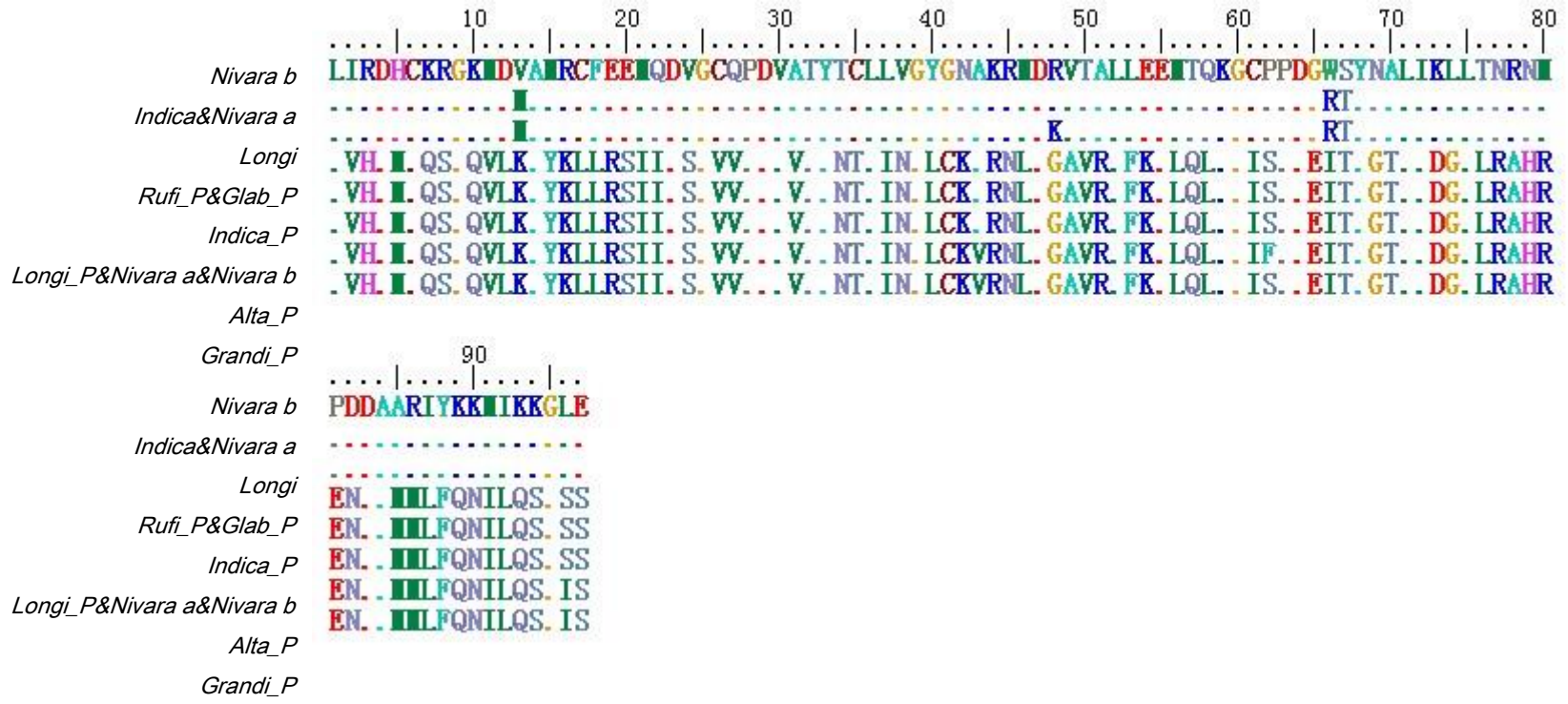


AK073060

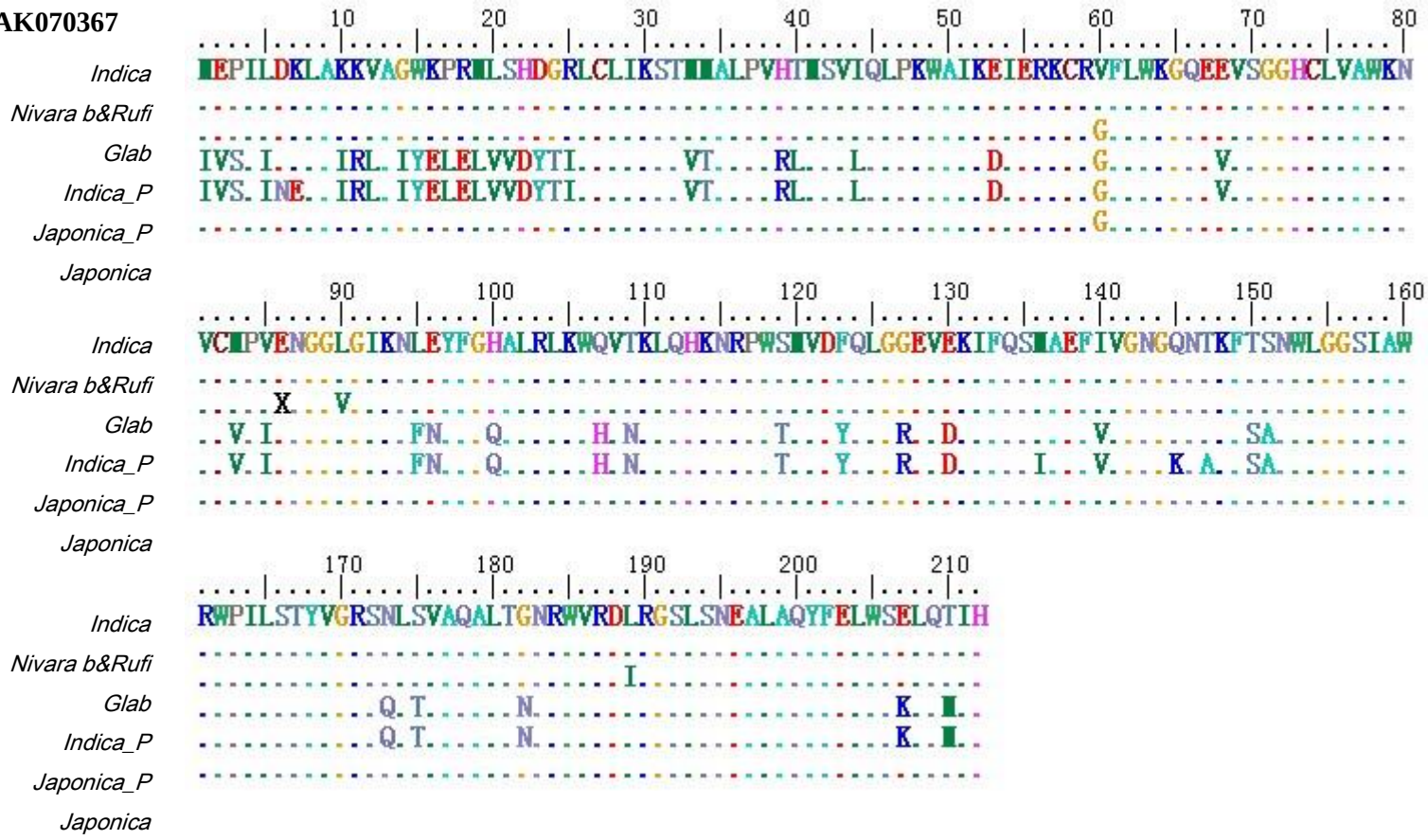


AK069420

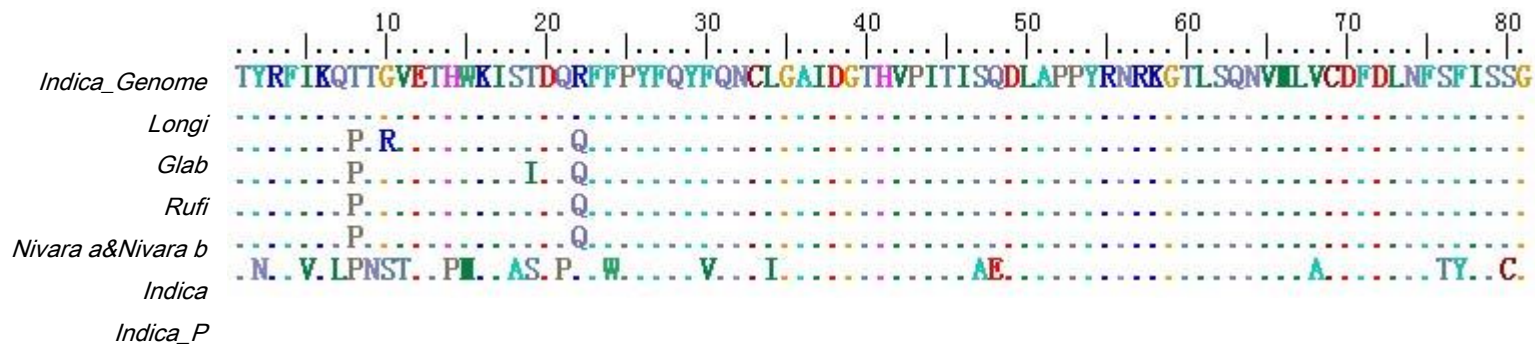


AK064415

AK070367



AK067652



AK068903

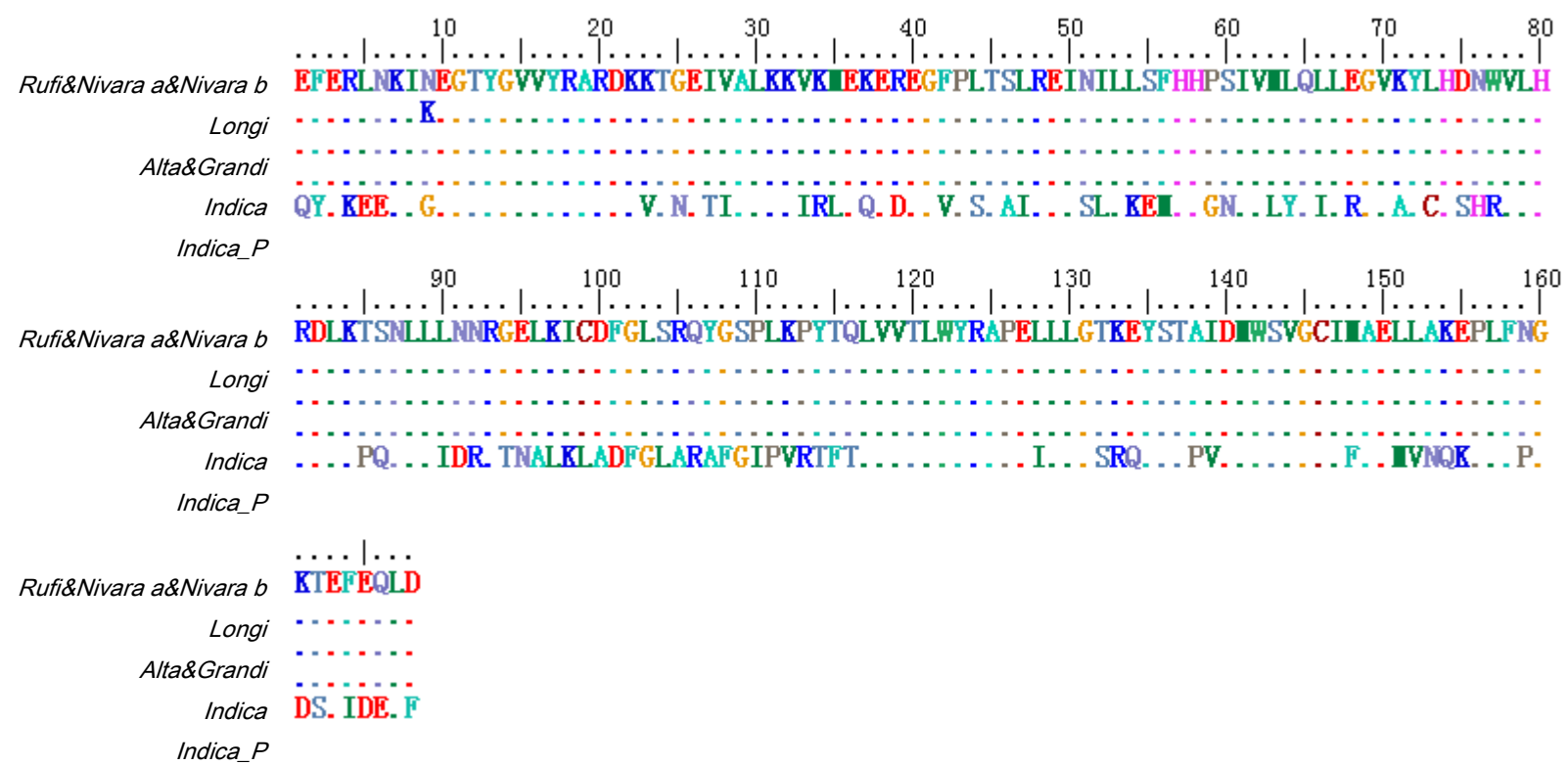
AK068903	10	20	30	40	50	60	70	80																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
Rufi	L	I	K	S	T	L	I	A	L	P	V	H	L	S	V	L	Q	L	P	K	W	A	I	K	D	I	E	R	E	C	G	G	F	L	W	K	G	Q	E	D	V	S	G	G	H	C	L	V	A	W	K	N	V	C	I	P	V	Q	N	G	G	L	G	I	K	N	L	D	Y	F	G	Q	A	L	R	L	K	W	H																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																												
Nivara a&Nivara b																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																											

AK070946

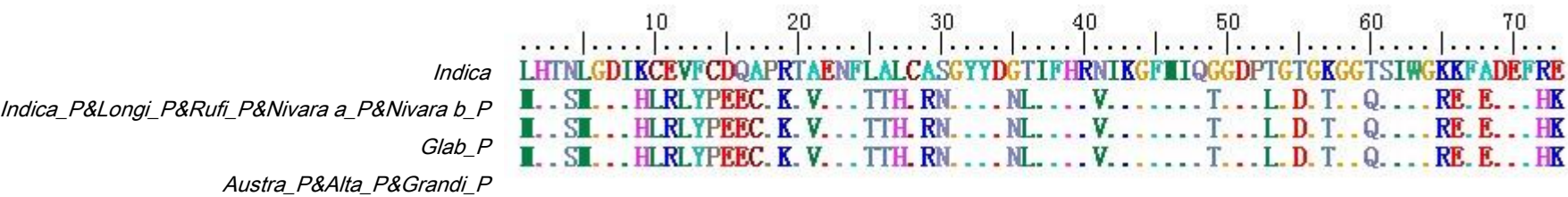
AK070340

	10	20	30	40	50	60	70	80
Indica&Glab&Rufi&Nivara a&Nivara b	YVQVLYNTGARKVVMIGVGQVGCSPNELARYSADGATCVARIDSAIQIFNRRLGFTETTAGCCGCVGRNNGQVTCCLPYEAP							
Grandi&Alta				E				
Austra				E				
Indica_P&Nivra a_P&Nivara b_P	FL.N..INV...I.LPP...A.HF.WE.GSQDGE.IDY.NNVVIQ..YA..LT..DA...L.KYG.LFI.VLPQIA							
Rufi_P	FL.N..INV...I.LPP...A.HF.WE.GSQDGE.IDY.NNVVIQ..YA..LT..DA...L.KYG.LFI.VLPQIA							
Longi_P	FL.N..INV...I.LPP...A.HF.WE.GSQDGE.IDY.NNVVIQ..YA..LT..DA...L.KYG.LFI.VLPQIA							
Glab_P								
	90	100	110	120				
Indica&Glab&Rufi&Nivara a&Nivara b	CSNRDQHIFWDAFHPSEAAANIIVGRRSYRAESPNDAYPIDIATLASV							
Grandi&Alta				V				
Austra								
Indica_P&Nivra a_P&Nivara b_P	..DASS.VW..E...TD.V.R.LADNVWSG.HTKIC..V.LQQIVKL							
Rufi_P	..DASS.VW..E...TD.V.R.LADNVWSG.HTKIC..V.LQQIVKL							
Longi_P	..DASS.VW..E...TD.V.R.LADNVWSG.HTKIC..V.LQQIVKL							
Glab_P								

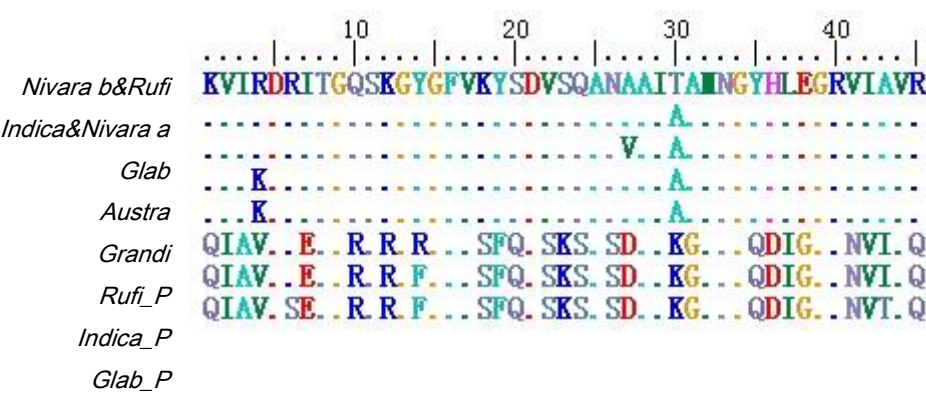
AK073961



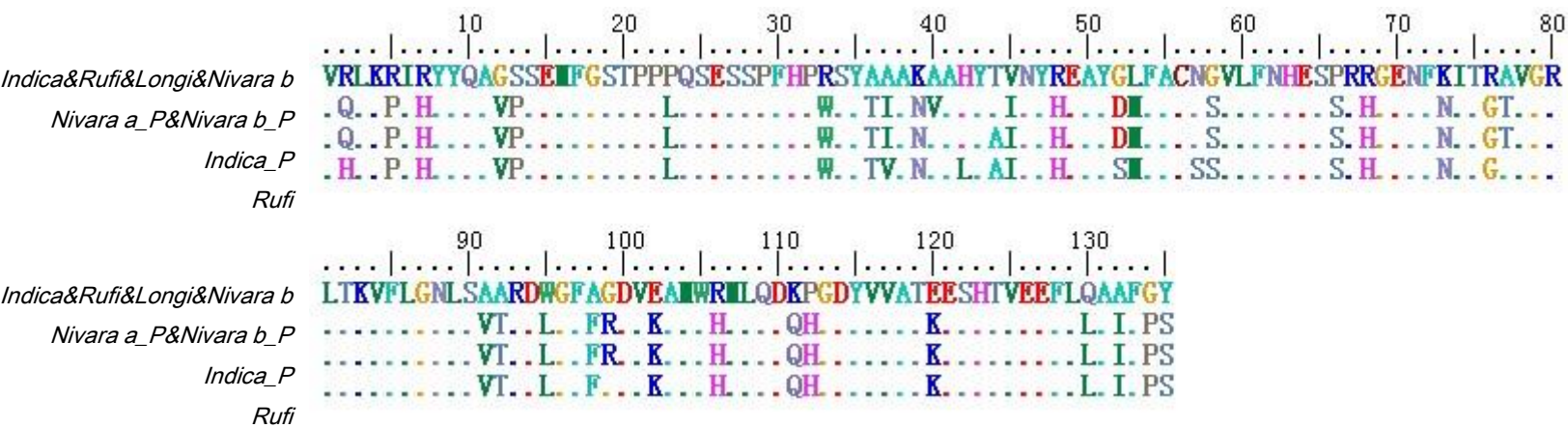
AK105360



AK106308



AK106445



AK108373

	10	20	30	40	50	60	70
Glab_P	TEEEIVAAAKAANAHEFISS	PEGYSTSVGER	TQLSGGQKQRIATARAIVKDP	PRILLLDEATSALDAESE			
Indica_P&Rufi_P&Nivara a_P&Nivara b_P							
Austra_P		T. Q. V.	Q				
Nivara a	S. V. ED. RS. D. NLKD. D. LC. V.				L. N. A. X.		SQ.
Rufi	S. V. ED. RS. D. NLKD. D. LC. V.				L. N. A.		SQ.
Indica_Genome	S. V. ED. RS. D. NLKD. D. WC. V.				L. N. A.	X.	SQ.
Glab	S. V. ED. RS. D. NLKD. D. WC. V.				L. N. A.		SQ.
Longi	S. A. ED. RS. D. NLKD. D. WC. V.				L. N. A.		SQ.
Austra	S. T. ED. RS. D. NLKD. D. WC. V.				L. N. A.		SQ.
Alta	S. V. ED. RS. D. NLKD. D. WC. V.				L. N. A.		SQ.

AK109583

	10	20	30	40	50	60	70	80
Indica_P&Rufi_P&Nivara a_P&Nivara b_P	YAGNKGPPLD	WQQRWKIAVGSARGLAYLHDDCSKQGDRE	FRAEVEIITRVHHRNLVSLVGF	CISGNERLLVYEFV	PNKTL			
Longi_P		R						
Austra_P	NS.					Y. F.		
Indica_cs	. YRYNND. CT. CR. I	NDI. K. C. EE. A. KK. L. T. GSIE. I. R. I. AEKSN. YH. RGS.						
Indica_Genome	. YRHNA. ST. CR. I	LDI. K. C. EE. AR. KK. L. T. GSIE. I. RII. AEKSN. YH. GGS.						
	90	100						
Indica_P&Rufi_P&Nivara a_P&Nivara b_P	DTHLHIGGGFGCVYRG	TLQD						
Longi_P								
Austra_P	. RWIYL. E. S. FE. EIGE							
Indica_cs	. KWIY. E. S. FE. K. SE							
Indica_Genome								

AK064216

	10	20	30	40	50	60	70	80
Longi	EVQSIGRIHHRNLVRIV	GYCKEREQRILVFEF	PGGSLRSILFQTIDTKVDVYSFGVVLLE	ICCRRCQDPVSGPWSWRT				
Japonica_Genome						D.		
Indica&Nivara b		Q.				D.		
Indica_P	. IFL. QFR. PH. KLL. C. D. E. L. Y. R. ENH. KSLNI. S. Y. LLTG. A	EH. R. GT. L						
	90	100	110	120	130			
Longi	EAALGIAKGIEY	LHEGCTSPIGCTSPII	KCDIKPDNILLDTNVRG	IRGYITPE				
Japonica_Genome								
Indica&Nivara b	KI. I. A. LAF. GAS. PV. AST. V. YR. F. AS. R. H. AA.							
Indica_P								

AK069257

	10	20	30
Indica_P	GLPPGFRFHPTDEELVNYYLKRKI	HGLNIEL	
Indica	LK. E. K. D. ARF. LAR. Q. KPLP.		
Nivara a&Japonica	L. E. K. D. ARF. LAR. Q. KPLP.		
Japonica_P		K.	

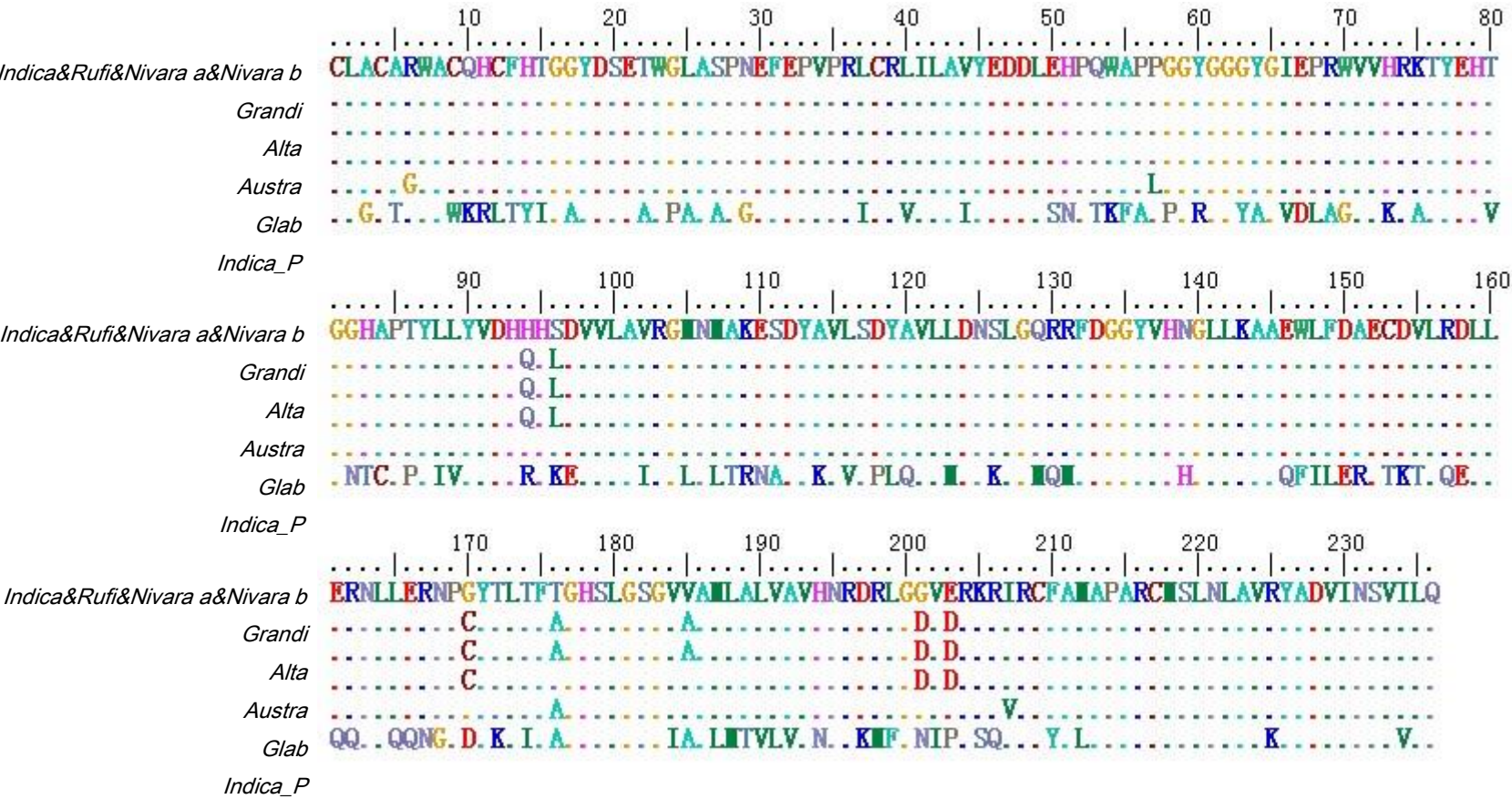


Figure S2. Alignment of the amino acid sequences of another 17 chimeric retrogene pairs

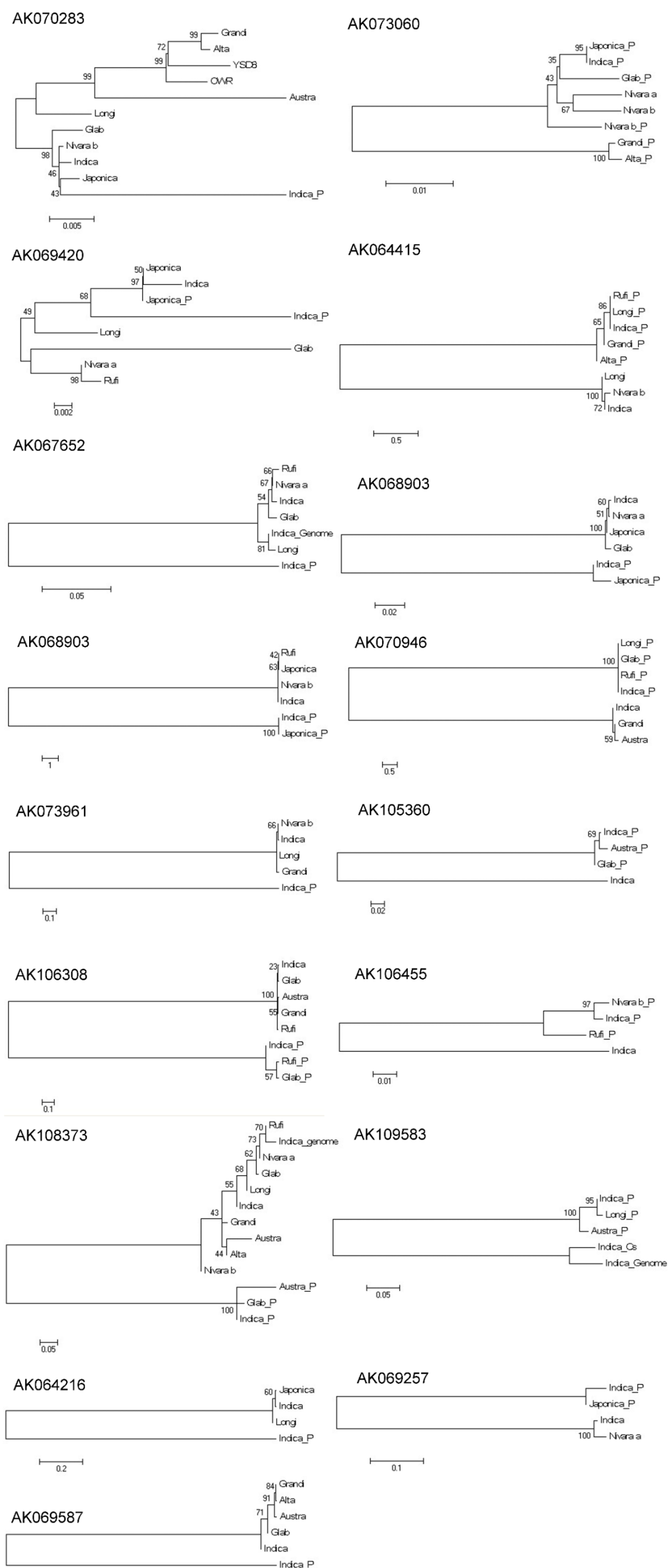


Figure S3. Phylogenetic relationships among 17 chimeric retrogene pairs. Refer to the Figure 1 legend for more details.