Supporting Information

Sharma et al. 10.1073/pnas.0805694105



Fig. 51. SimPlots illustrate the similarity of recombinants R1 to R5 to either parental element CRM1A or CRM1B in different regions of the element. BioEdit was used to generate an 80% consensus sequence for each CRM1 subgroup using separate multiple sequence alignments generated by ClustalX from each of the two parental (CRM1A and CRM1B) and five recombinant (R1–R5) subgroups. The similarity of each recombinant to CRM1A (*green*) and CRM1B (*orange*) was plotted using SimPlot with a sliding window of 200 nt and a step size of 20 nt. Detailed recombination breakpoints for all recombinants are shown in Fig. S2. (a) Recombinant R1 more closely resembles parent CRM1B along most of its length, with the exception of the RNAseH region (*circled*). The R1 consensus was derived from all 89 R1 sequences. Note that the sequence similarity is close to 100% to one or the other parent, except for the region around 1,100 nt (*arrow*) and 1,900 nt (*arrowhead*). The former is a highly variable AT-rich region of the UTR containing a duplication within the R1 consensus that results in a big gap in the alignment of A and B with R1, while the latter is a UTR region with high sequence variation. High sequence variation among individual B sequences in part of the UTR indicates that the consensus B sequence is not representative of the B parent in this region. (*b*) The R1 consensus for this figure was derived from three of the four oldest R1 sequences. One R1 sequence similarity with "B" because they lack the duplication (*arrow*) and the ambiguous region of the UTR (*arrowhead*), resulting in a much cleaner (and slightly shorter) SimPlot (compare with Fig. S1 *a*). (*c*–f) SimPlots of recombinants R2 to R5 demonstrate sequence similarities of each recombinant to either parent CRM1B or CRM1B in different regions of the element, as represented schematically in Fig. 1.









U ⊲



v ▼

Supporting Figure 2a:



Fig. S2. Recombination breakpoints in CRM1 and CRM4 elements. Multiple sequence alignments of the recombinant CRM1 and CRM4 subgroups with their respective parents are shown. The consensus sequence of the recombinant is shown in the middle and those of the parents are at the top and bottom of each alignment. Identities to the top-most sequence are indicated by (.) dots. The recombinant sequence derived from A and B lineages is highlighted in yellow and gray, respectively. The recombination breakpoints identified by "2 parental, 1 derived" Maximum Chi Squared Test is marked by asterisk "*" for each recombinant. Note: The A polyprotein consensus was generated using the A polyprotein associated with the A and R3 LTRs, while that for the B polyprotein was generated from B polyprotein associated with the B LTRs and their deletion derivatives. The B_{Δ} indicates subgroup B_{Δ} 12a3a, while R1, R2, R3, R4 and R5 represent different recombinant groups. (a) Two recombination breakpoints flank the RH_A in CRM1 R1 (middle sequence). The flanking RT and Integrase domains (*black* and *blue* underline, respectively) are included in this alignment to better visualize the recombination breakpoints. Both recombination breakpoints R2, R3 and R5, as well as CRM4 R1.

b. CRM1 R2 polyprotein (integrase)

PNAS PNAS

C	RM1_R3 LTRs
	130 140 150 160 170 180 190 200 210 220 230 240
	TGCCAGAGATCTTAACTTCGTCATGCTACTGAAGAATGAGGGCCCCAGAAGAATAGACGACCAGCCCAACTGCGGCCCATAGTGGACGACC-TAGGGT
	····A·····AT.gA.A
_	
c	RM1 R5 polyprotein (integrase)
c	RM1 R5 polyprotein (integrase)
c	RM1 R5 polyprotein (integrase)
c	RM1 R5 polyprotein (integrase) 4200 4210 4220 4230 4240 4250 4260 4270 4280 4290 4300 42 TOCKALLAL TOROT TOCTOOT AGAILAGAILAGAILAGAILAGAILAGAILAGAILAGAI
с	RM1 R5 polyprotein (integrase) 4200 4210 4220 4230 4240 4250 4260 4270 4280 4290 4300 4 TOCHALAL TOCHAGE TOCTOGE AGE AND ADDREAD ADD

Figure S2. Continued

							1010010001							
	an TON	oalei	CACC	OCANC	ATGAG	TACGACT	ленесносы	seace receaci	Texilococer	crice				TICAL
						•••••					*********			
									· · · · · · T · · · ·		*********			
				1										
a' 🛄				- 11					CT			2		
				_										
			· • · AS	R	.CR						TCAGETOGGE	AACTOCACT	TEADGHREEA	ettee
		180		190	.cx	200	210	220	230	240	250	260	270	280
		180		190	.CR	200	210	220 CATAGTOGA	230	240	250	260	270	200
	LANTO	180		190	.CR	200	210 	220 	230 230	240	250 250	260	276	200
	JAATG	180		190	.CR	200 00.00000	210 AAC 100000	220 220 20	230	240	250	260	270 	200
	33.A TG	180	OCAGA	190	.CR	200	210 	220 	230 	240	250 	260 	270	280
	38.ATG	180		190 	.CR	200	210 	220 	230 	240 199009000	250 250	260	270 	280
	3A.A.TQ	180	000404	190 	.CR	200	210 	220 	230 a.oct.agga	240	25-0 77.00011000	260 	270 1110200000 	280

Fig. S3. Deletion derivatives of B LTR. Each deletion derivative of the CRM1B LTR is defined by a characteristic deletion as indicated on this multiple sequence alignment of the 5' end of the CRM1B LTR and its deletion derivatives. The deletions are numbered (1, 2, or 3) based on the region with the deletion, followed by a letter to indicate different extent of deletion. $B_{\Delta 12a3a'}$ differs from $B_{\Delta 12a3a}$ by two additional trinucleotide deletions and base changes.

PNAS PNAS

571	ATG-gag	382-280	gag-RT	RT-RT	RT-ini	ind-ind	int-rec	TEC-10C	rec-lbr	LTR	
A	A	A	A	A	A	A	A	A	A	A	1 AAAAAAAAAAAA
A	A	λ.	A	A	A	dh	8	8	B	8	1 AAAAAAIhBB88
AB	A	Α.	A	A	A	A	A	A	A	ABa	1 ABAAAAAAAAAAAAA
AB	A	A	A	A	A	A	A	A	A	AB	3 ABAAAAAAAAAAB
AB	A	A	A	A	A	A	A	A	A	AB	ABAAAAAAAAAA
AB	A	Α	A	A	A	A	A	A	A	AB	ABAAAAAAAAAA
AB	A	A	A	A	A	A	A	A	A	ADA	ABAAAAAAAAABA
AB	A	A	A	A	A	A	A	B	8	B.A.	ABAAAAAABBBA
AB	4	A	A	A	A	B	8	8	B	8	1 ABAAAAAS88888
AB	B	8	B	B	4	B	8	8	4	481	1 ARREARARBRAARA
ABA	A	Ă	A	A	A	A	Å	A	A	AB	1 ABAAAAAAAAAAAB
ABA	4	X.	Ä	A	A.	A	× ×	8	B	BA	2 48 4444444468884
ABI	6	11 C	i i i	à	6	<u>A</u>	1 A A	Ř.	R	RA.	4844444446884
ABA	5	8	B	8	Â	8		8	A	ADA	12 ABABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
45.4		-			2				2	600.5	AB (BBBBABBABBAAB)
ARX		-	ä		2				2	68.5	AB (BBBBABBBAAB)
ADA	8	2		6	0	6		6	2	60.1	AD + DDDDADDDAAD +
ABA	2	2	2	-	Q	2	2		<u> </u>	100	ABABBBBBBBBBBBBBBB
ADA			0	8	<u>2</u>				0	40.4	AD IDDDDDDDDDDDDDDD
ABA		2	8		<u>0</u>	0	2	0	0	A003	AB-BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
ABA	8	2	8	8	<u>0</u>	8	2	8	<u>0</u>	ABO	ABABBBBBBBBBBBBBBB
ABI		-			<u>0</u>				<u>^</u>	ADA	ABABBBBBBBBBBBBBB
ABA				8	A			8	A	ABA	ABABBBBABBBAABA
ABA	в	8	8	8	Α.	в	8	8	Α.	ABA	ABABBBBABBBAABA
ABA	8	8	8	8	A.	8	8	8	A.	ABA	ABABBBBABBBAABA
ABA	5	8	В	в	A	5	8	в	A	ABA	ABABBBBBBBBBBBBBB
ABA	В	8	8	8	A	B	8	8	A.	ABA	ABABBBBABBBAAB3
ABA	В	8	8	8	A.	B	8	8	A	ABa	ABABBBBABBBAABA
ABA	8	8	В	B	A	8	8	B	A	ABA	ABABBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
ABA	5	8	B	8	A	5	8	8	A	ABA	ABADBDDADDDAABA
ABA	R	8	8	8	ă.	8	8	8	6	484	AB (BRBBARBAAB)
ARA	8	ě.	ñ	n i	<u>2</u>	e e	ě	ñ	2	684	AB (BBBBABBBABB)
48.4		÷	ň.	ii ii	2		÷.	ii ii	2	454	ABABBBBBBBBBBBBBBB
ADA			0		0		-	0	8	P	E AD (DDDDADDDDD
ADA	2	2	8	0	0	0	2	0		8	AD-0000A00000
ABA	8	8	8	8	<u>A</u> .	8	2	8	8	8	ABABBBBBBBBBBBBB
ABL		•	B		<u>0</u>		•	8			ABABBBBABBBBB
ABL	8	8	в	в	A.	8	8	8		8	ABABBBBABBBBB
ABA	в	8	8	8	A.	B	8	8	B	8	ABABBBBABBBBB
ABA	B	8	8	8	8	8	8	8	B	8	1 AB./88868888888
ABA	A	A	A	A	A	A	- A	A	A	ADA	2 ABAAAAAAAAAAAAA
ABA	A	Α	A	A	A	A	A	A	A	ABA	ABAAAAAAAAAAAAAA
B.A	A	A	Α	A	A	A	A	8	B	BA	8 BAAAAAAAABBBA
ELA.	A	A	A	A	A	A	A	8	8	E.A.	BAAAAAABBBBA
ELA.	A		A	A	A	A	A	8	8	BA.	BAAAAAABBBA
B	A		ch	8	A	B	8	8	B	B	1 BAAchBABBS88
RA.	B	8	8	8	6	B	8	8	4	481	1 BARBARARRAARA
1	5	8	B	8	A	5	8	8	A	ABA	11 DDDDDDADDDDAAD)
R	8	ē.	8	8	ñ.	8	R.	8	6	484	BBBBBBBBBBBBBBBBBBB
ň		2	ĕ	n n	2	5	÷.	ä	2	4.0.1	DDDDDADDDAAD S
		2	2		0	2	2		Q	1000	BBBBBABBBAABS
		-			2		-		2	Amo	REFERENCES OF CARE
0	0	2	0	0	2	0	2	0	2	404	0000040004401
0	0	2	0	0	2	0	2	0	0	ABL	BBBBBABBBAAB3
8	8	0	6	8	A	8	8	8	A.	ABA	BRBSBABSBAAB3
		2	0		2		2	0	A	ABA	BBBBBBABBBBAABA
8	8	8	8	8	A	в	8	8	A.	A81	BBBBBBABBBBAAB3
8	8	8	8	8	A	B	8	8	A.	ABA	88888A888A83
8	8	8	в	8	A	8	8	в	A	ABA	BBBBBBABBBBAABA
8	8	8	в	8	A.	B	8	8	8	8	12 BB558A55685
B	B	8	8	8	A	B	8	8	B	8	88888A88888
B	в	8	8	B	Δ	B	8	8	B	B	BBBBBBABBBBB
0	8	8	B	B	A	8	8	B	8	8	DDDDDADDDDD
1			B		A			8			DDDDDA5DDD5
8	8	8	8	8	6	8	8	8	8	8	BBBBBABBBBB
			B.		2						REFERANCE
		-			2		-				B B B B B A B B B B B B B B B B B B B B
		-		0	2		2	0			00000400000
8	8	0	8	8	0	8	8	8	8	8	RREERVERSER
8	8	8	8	8	A	8	8	8	8	8	88968A96888
8	8	6	8	8	A	8	8	8	8	8	BBBBBABBBBB
8	5	8	в	в	A.	5	8	B	8	5	BBBBBABBBBBB
8	8	8	8	8	8	B	8	8	A	AB	1 888888888AA8
8	8	8	в	8	8	8	8	в	8	5	6 88868886688
8	8	8	B	B	8	8	8	B	8	8	000000000000
В	8	8	8	B	В	B	8	8	В	8	BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
B	8	8	B	8	B	B	8	8	B	8	BBBBBBBBBBBBBBB
0			B				5	B			DDDDDDDDDDD
		-					-				REPRESENTE
80		-	100	10	2.7	100	100	1.0			

Fig. S4. Allele composition of 75 CRM1 elements lacking TSD and presumed to have resulted from DNA recombination. Each column represents a region of the element, including: 5'LTR, translation start-gag domain, gag domain, gag domain-RT domain, RT domain, RT domain-integrase domain (contains RH), integrase domain, integrase domain-recombination point R2, recombination point R2-recombination point R5, recombination point R5-LTR, LTR. Pink highlight marks the 29 elements that are chimeric based on containing different LTRs or a chimeric polyprotein gene. Note that the LTRs are grouped based on type, including A, B, AB, AB_Δ, ABA and BA. ch, chimeric region.

PNAS PNAS



Fig. S5. Evidence for recombination breakpoints in Opie elements. Full-length and partial Opie elements were identified in a BLAST search of 15,101 Zea mays inbred B73 BACs representing 2,503,489,719 nt using GenBank accession U68408 as a query. Regions (3,500–6,000 nt) flanked by two Opie LTRs (from U68408, AF050450, AF050451, AF050452 or AF050453) in the same orientation represent potential complete Opie elements, which were excised from the genomic BAC sequences including 1,500 flanking nt. Two identical duplicate elements were removed and the remaining 65 full-length elements were aligned using ClustalW and trimmed to the length of the full-length element (alignment available from the authors upon request). Presence of TSDs was determined visually using the alignments, and insertion times for elements flanked by TSDs were estimated using the method of SanMiguel *et al.* [SanMiguel P, Gaut B, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* 20:43–45]. All elements retrieved in this manner had been inserted recently. The alignment was scored for polymorphic sites consisting of at least four transversions. Dashes (-) and gray boxes indicate deletions and missing data, respectively. Lack of polymorphism is indicated in pink highlights.





Fig. S6. Recombination breakpoints in Opie polyprotein. Multiple sequence alignments of two recombinant Opie subgroups with two "parents" are shown. The consensus sequence of the recombinants is shown in the middle and those of the parents are at the top and bottom of each alignment. Note that in these highly recombined elements, the parental and recombinant designations are arbitrary. Identities to the top-most sequence are indicated by (.) dots. The sequence derived from A and B lineages is highlighted yellow and gray, respectively. The recombination breakpoints identified by "2 parental, 1 derived" Maximum Chi Squared Test is marked by asterisk for each recombinant (after 136 for AB and after 162 for BA).