Supporting Information

Sharma *et al.* 10.1073/pnas.0805694105

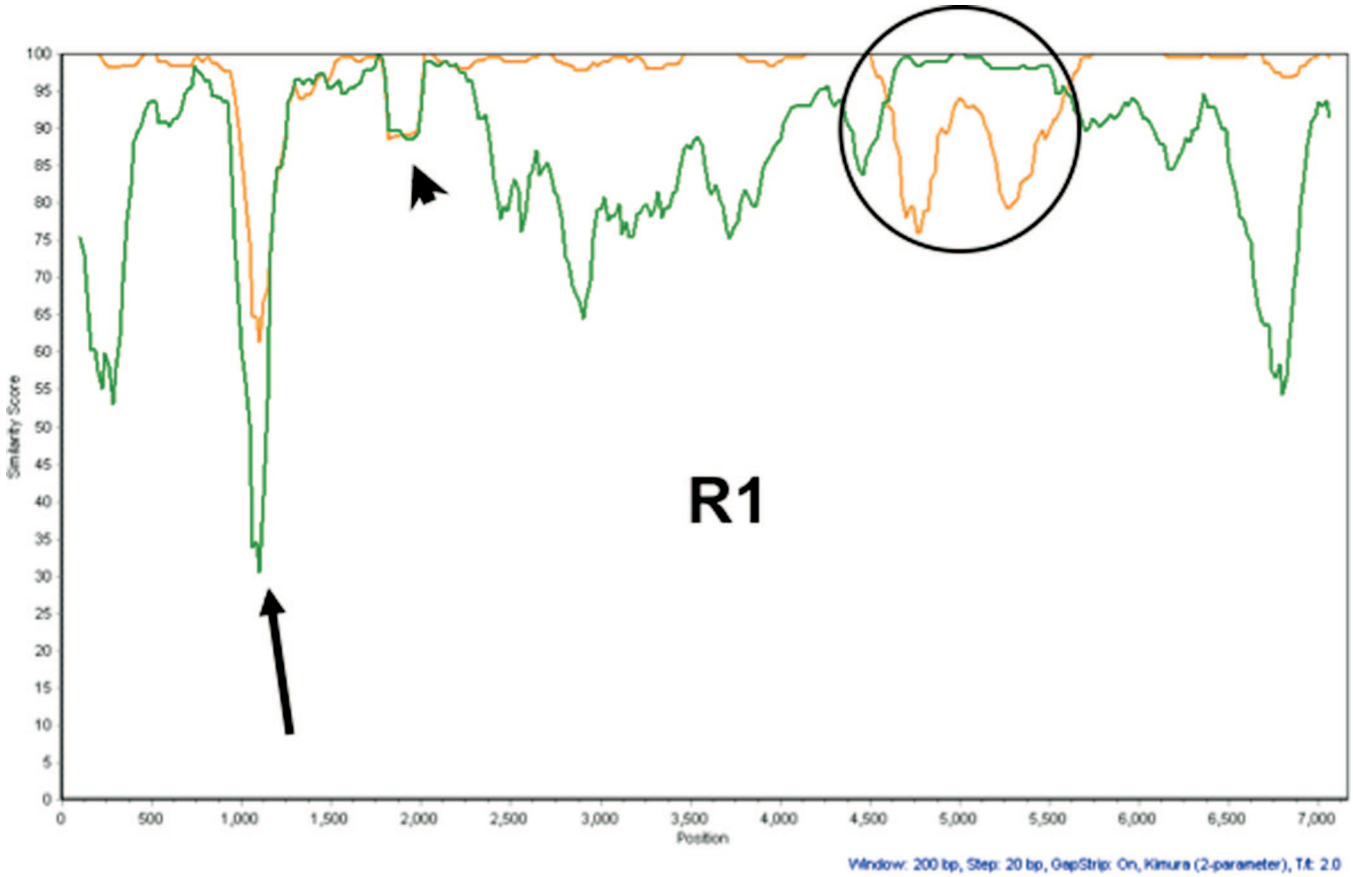


Fig. S1. SimPlots illustrate the similarity of recombinants R1 to R5 to either parental element CRM1A or CRM1B in different regions of the element. BioEdit was used to generate an 80% consensus sequence for each CRM1 subgroup using separate multiple sequence alignments generated by ClustalX from each 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 was not included because of the presence of a duplication in the UTR. Note that the consensus derived from these oldest R1 elements shows high 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 CRM1A or CRM1B in different regions of the element, as represented schematically in Fig. 1.

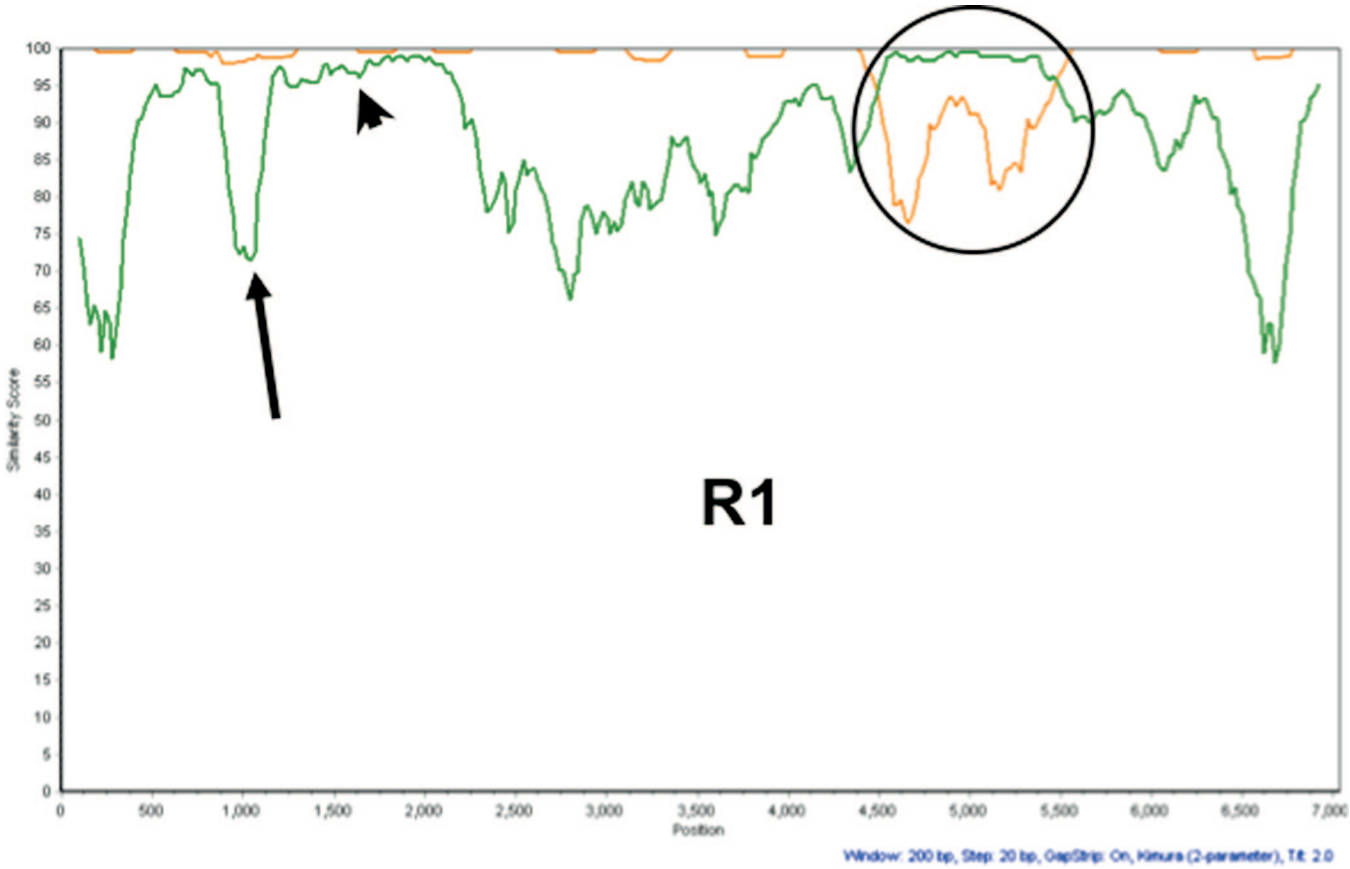


Figure S1. *Continued*

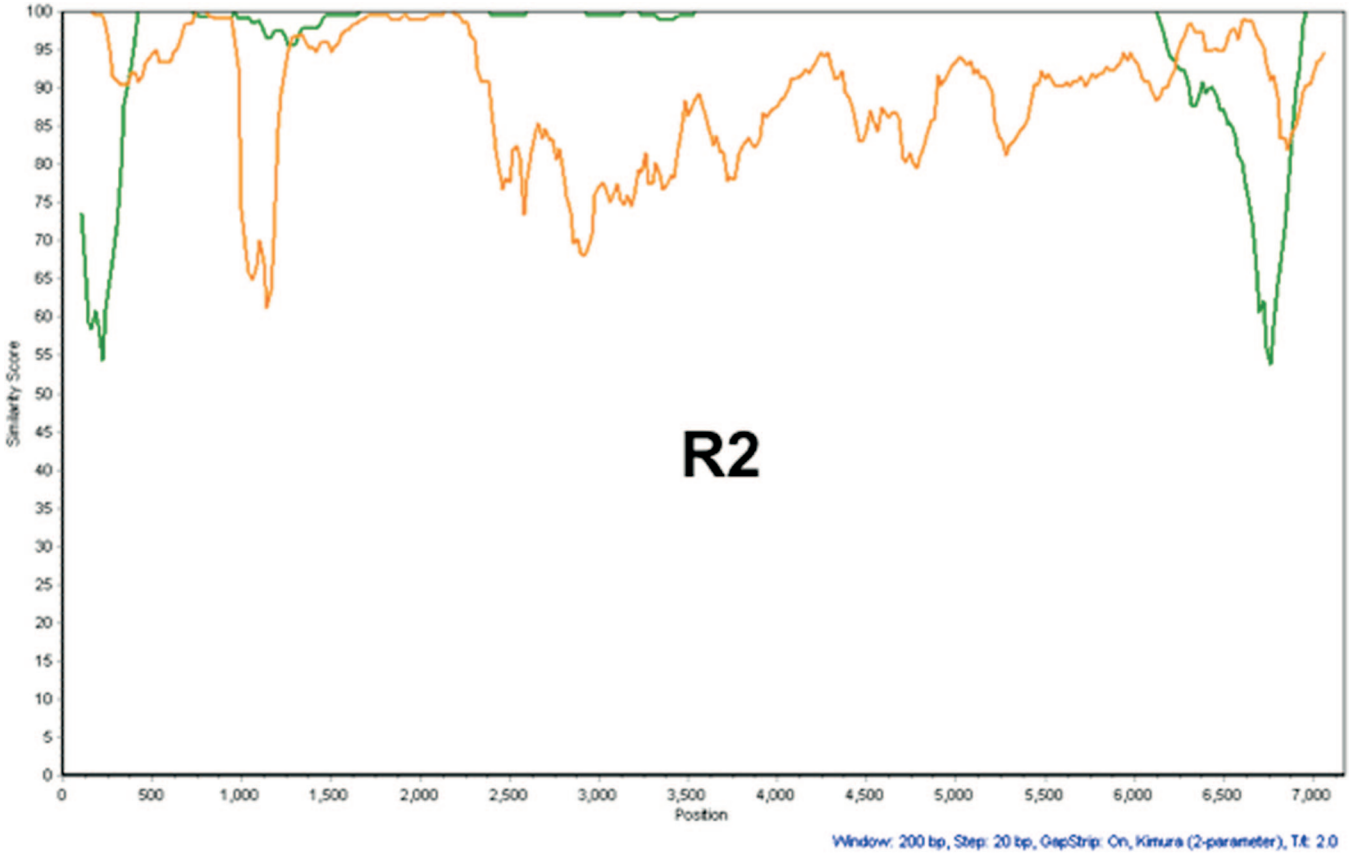


Figure S1. *Continued*

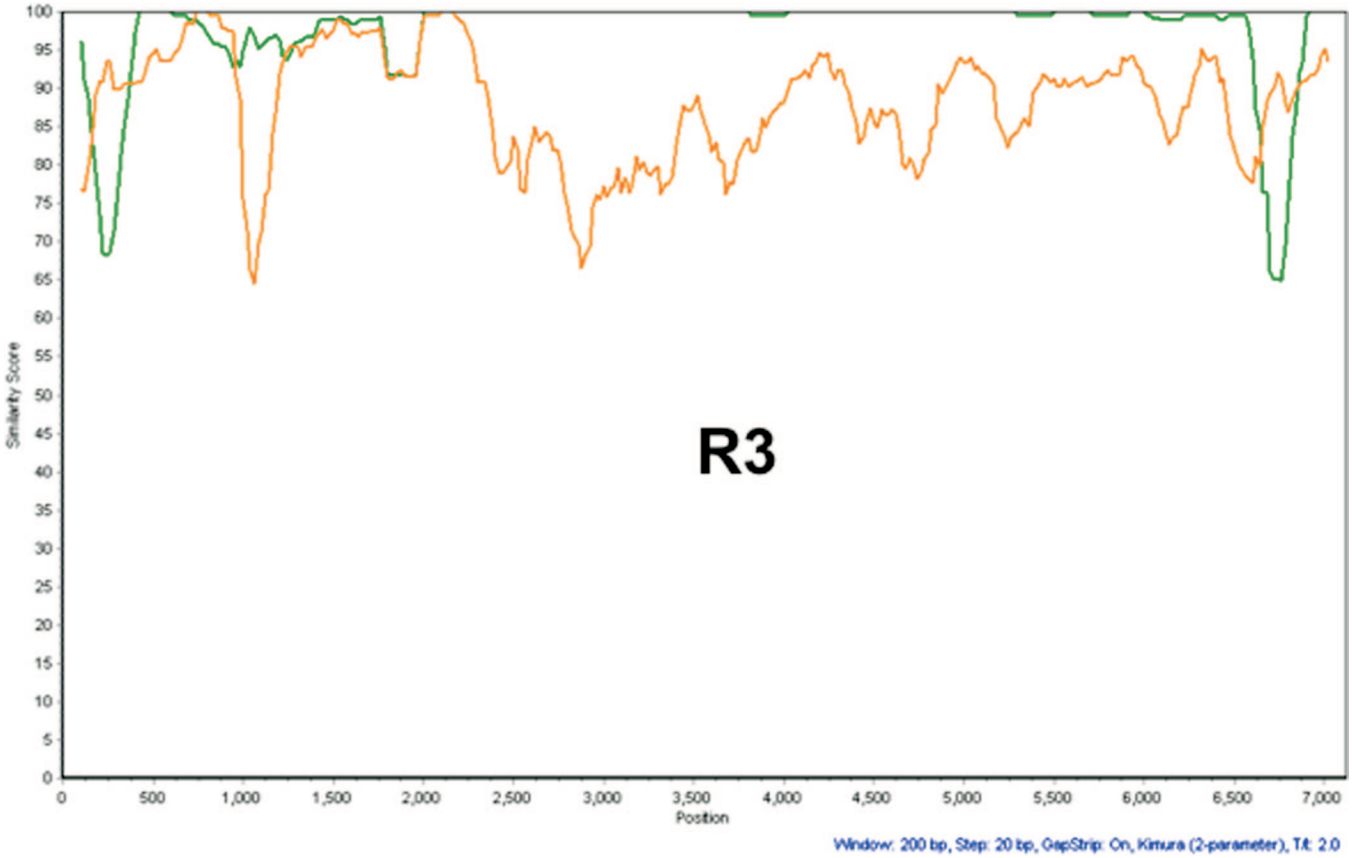


Figure S1. *Continued*

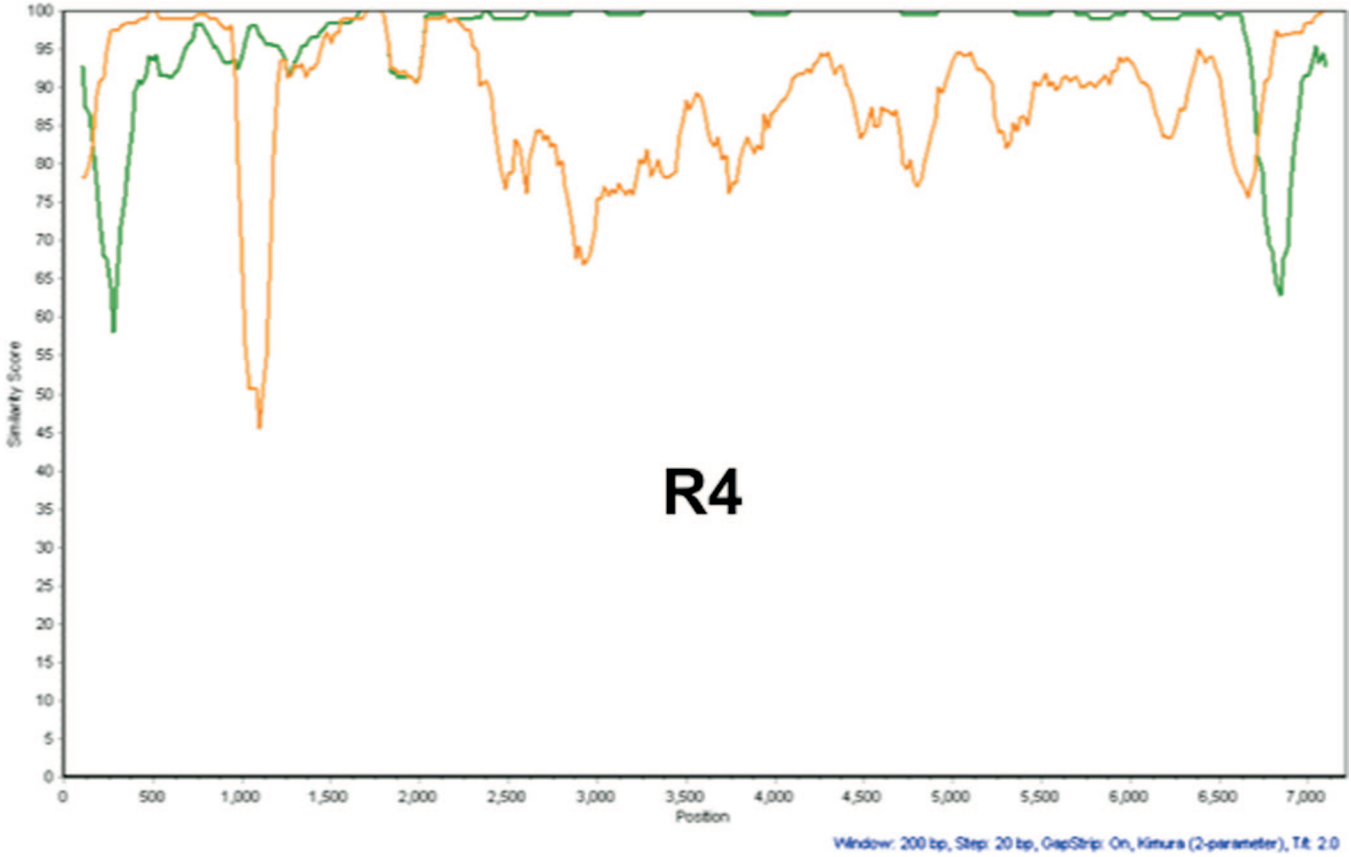


Figure S1. *Continued*

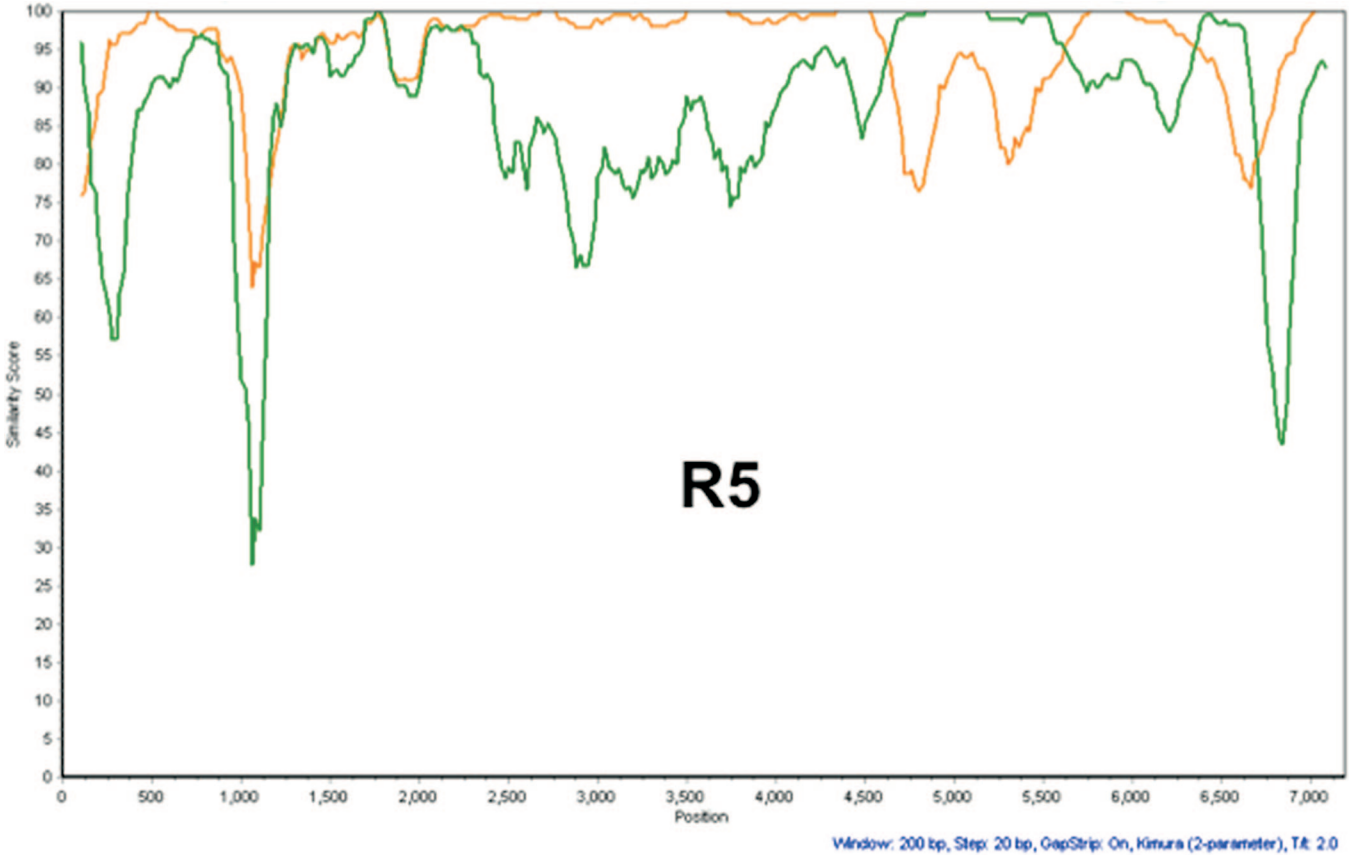


Figure S1. *Continued*

Supporting Figure 2a:

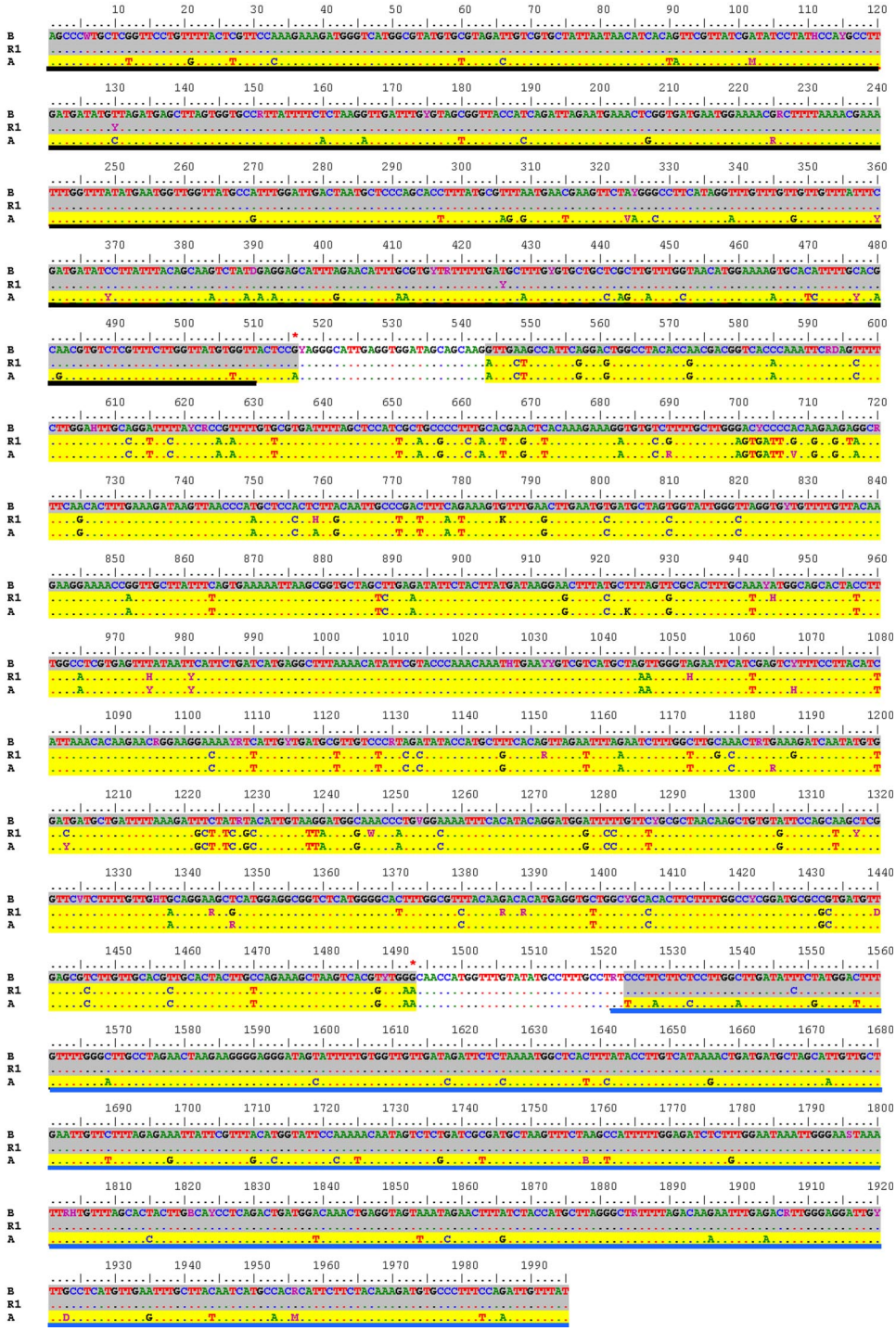


Fig. S2. Recombination breakpoints in CRM1 and CRM4 elements. Multiple sequence alignments of the recombinant CRM1 and CRM4 subgroups with their respective parents are shown. The consensus sequence of the recombinant is shown in the middle and those of the parents are at the top and bottom of each alignment. 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 polyproteins 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 map upstream of a region of high sequence similarity shared by CRM1 A and B. (*b–e*) The recombination breakpoints determined for CRM1 recombinants R2, R3 and R5, as well as CRM4 R1.

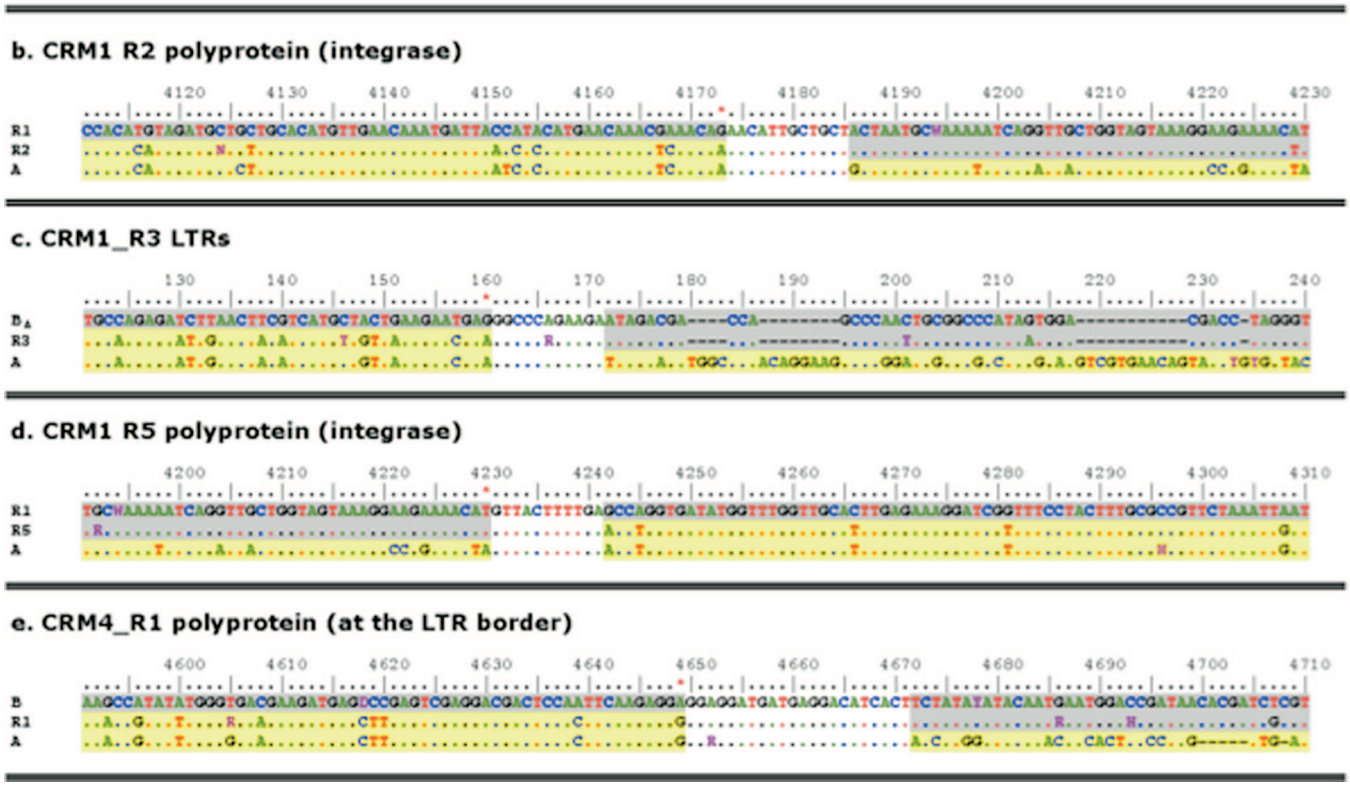


Figure S2. *Continued*

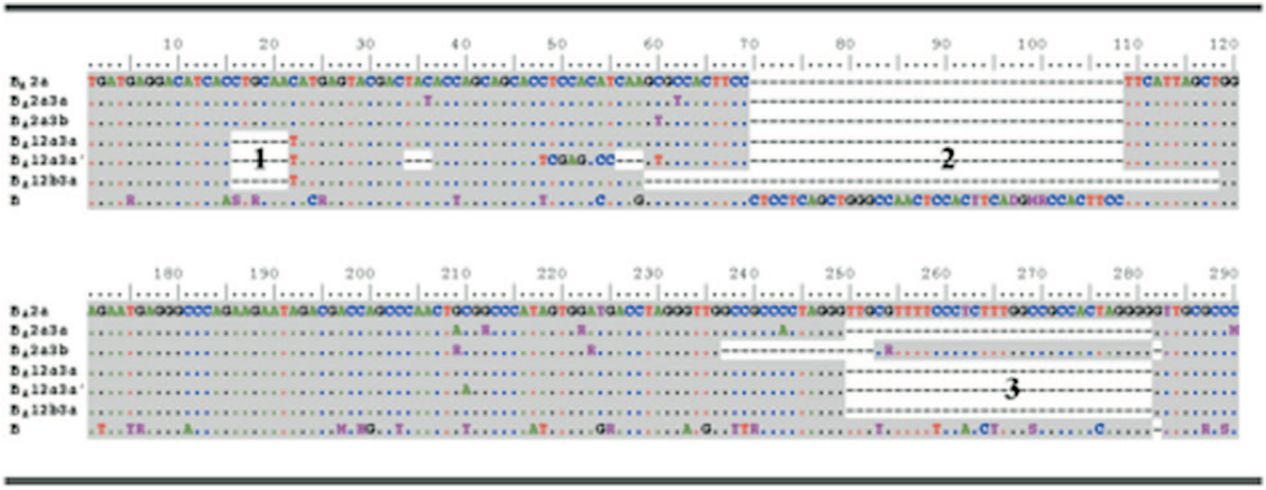


Fig. S3. Deletion derivatives of B LTR. Each deletion derivative of the CRM1B LTR is defined by a characteristic deletion as indicated on this multiple sequence alignment of the 5′ end of the CRM1B LTR and its deletion derivatives. The deletions are numbered (1, 2, or 3) based on the region with the 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.

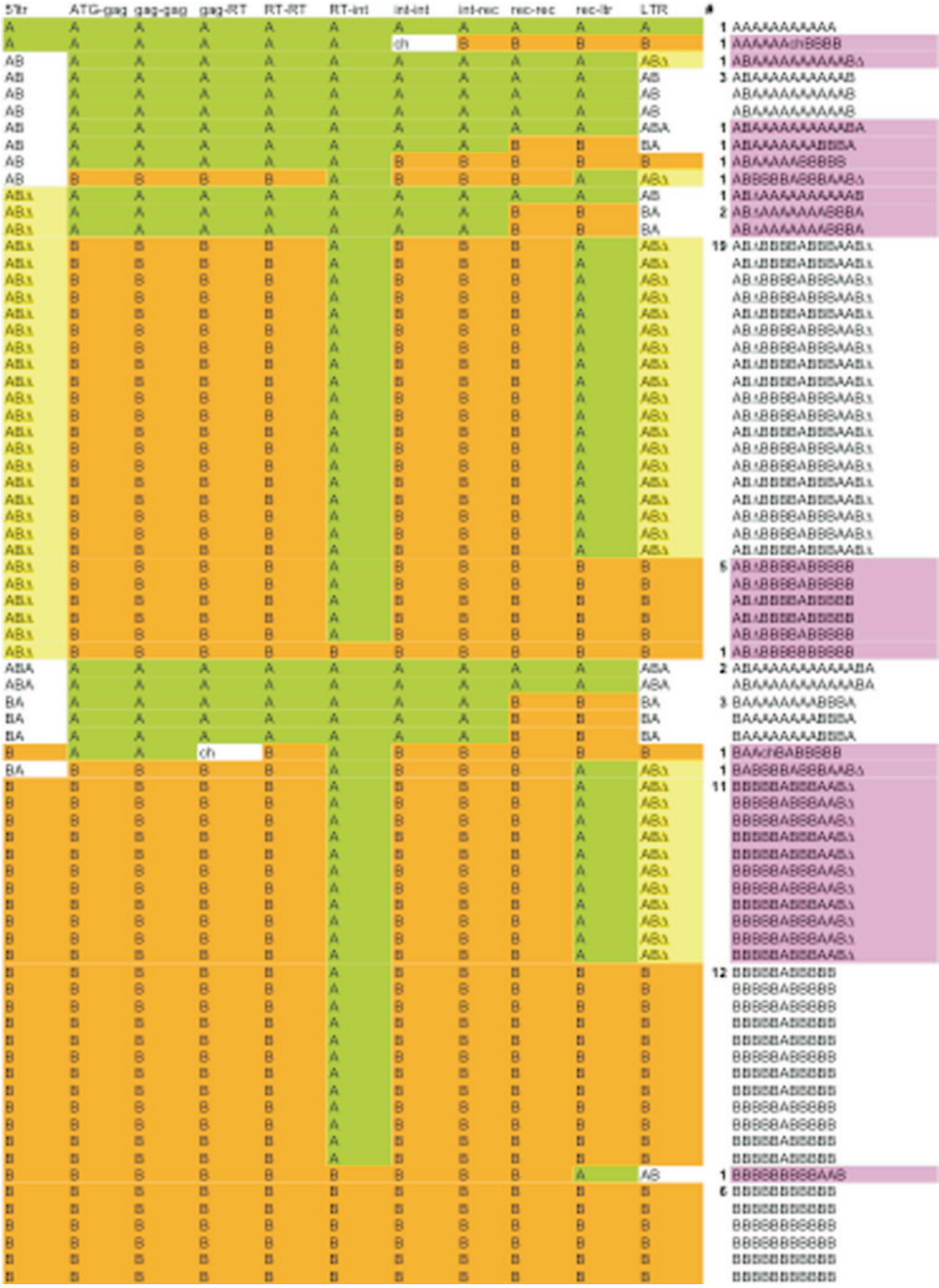
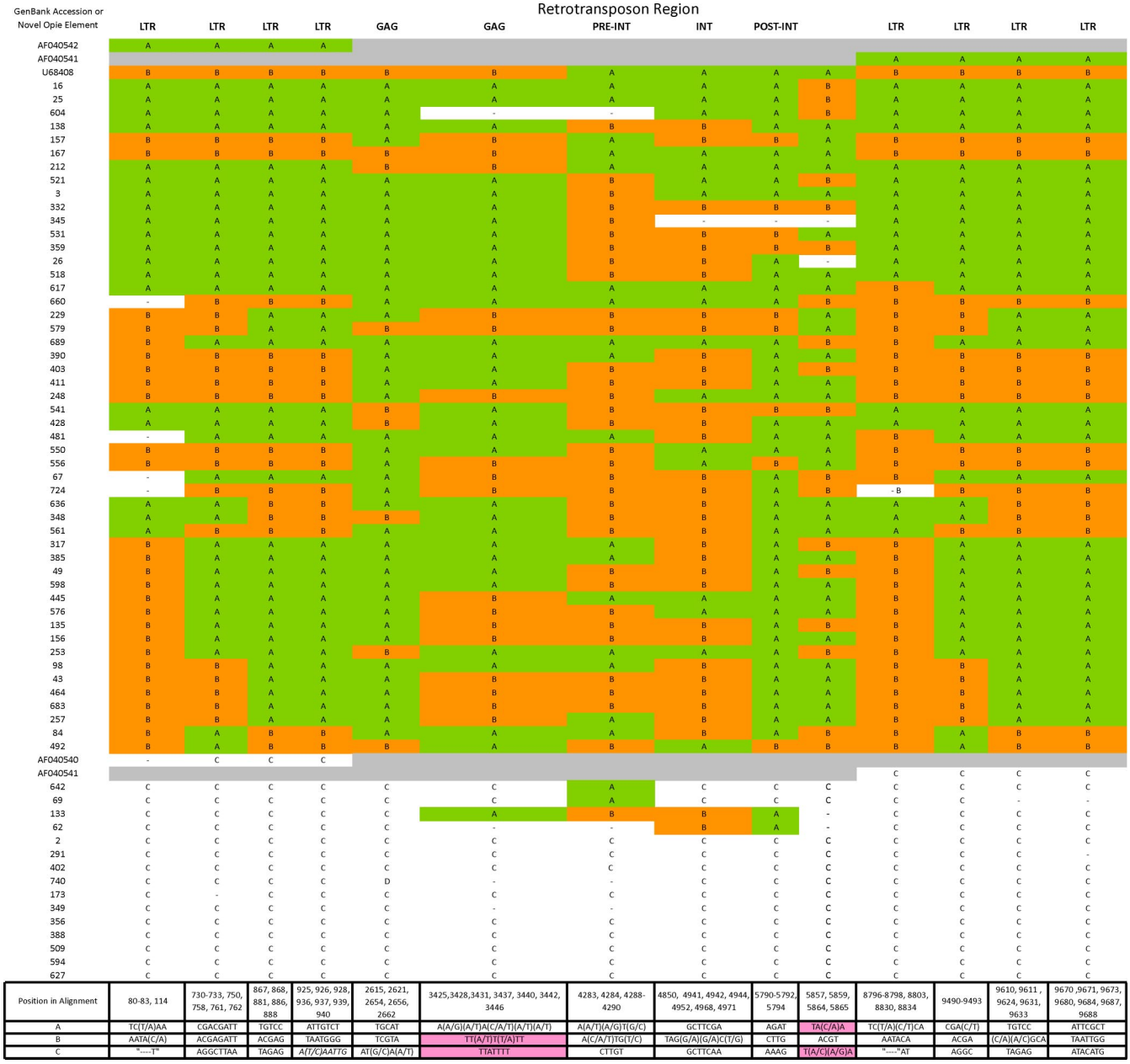


Fig. S4. Allele composition of 75 CRM1 elements lacking TSD and presumed to have resulted from DNA recombination. Each column represents a region of the element, including: 5′LTR, translation start-gag domain, gag domain, gag domain-RT domain, RT domain, RT domain-integrase domain (contains RH), integrase 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.



Position in Alignment	80-83, 114	730-733, 750, 758, 761, 762	867, 868, 881, 886, 888	925, 926, 928, 936, 937, 939, 940	2615, 2621, 2654, 2656, 2662	3425,3428,3431, 3437, 3440, 3442, 3446	4283, 4284, 4288-4290	4850, 4941, 4942, 4944, 4952, 4968, 4971	5790-5792, 5794	5857, 5859, 5864, 5865	8796-8798, 8803, 8830, 8834	9490-9493	9610, 9611 , 9624, 9631, 9633	9670 ,9671, 9673, 9680, 9684, 9687, 9688
A	TC(T/A)AA	CGACGATT	TGTCC	ATTGTCT	TGCAT	A(A/G)(A/T)A(C/A/T)(A/T)(A/T)	A(A/T)(A/G)T(G/C)	GCTTCGA	AGAT	TA(C/A)A	TC(T/A)(C/T)CA	CGA(C/T)	TGTCC	ATTCGCT
B	AATA(C/A)	ACGAGATT	ACGAG	TAATGGG	TCGTA	TT(A/T)T(T/A)TT	A(C/A/T)TG(T/C)	TAG(G/A)(G/A)C(T/G)	CTTG	ACGT	AATACA	ACGA	(C/A)(A/C)GCA	TAATTGG
C	"----T"	AGGCTTAA	TAGAG	*A(T/C)AATTG*	AT(G/C)A(A/T)	TTATTTT	CTTGT	GCTTCAA	AAAG	T(A/C)(A/G)A	"----"AT	AGGC	TAGAG	ATACATG

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.

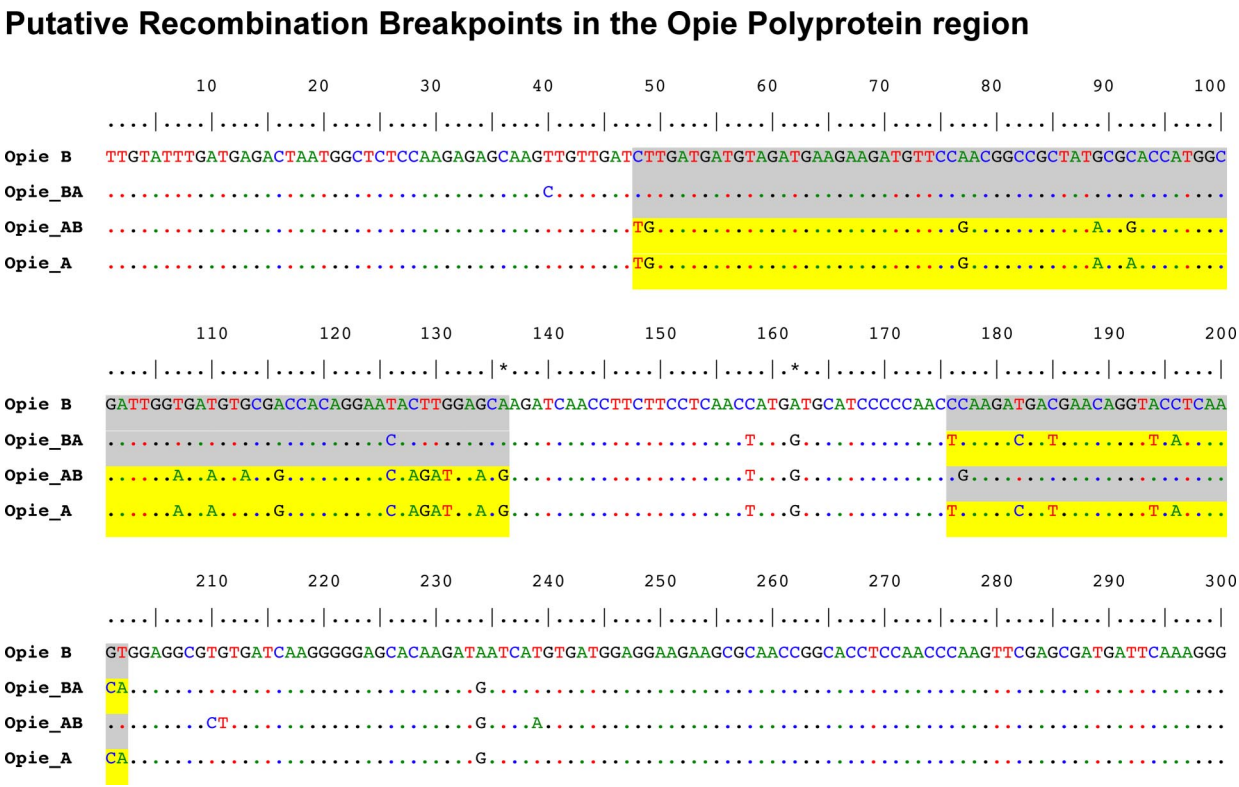


Fig. S6. Recombination breakpoints in Opie polyprotein. Multiple sequence alignments of two recombinant Opie subgroups with two ''parents'' are shown. The consensus sequence of the recombinants is shown in the middle and those of the parents are at the top and bottom of each alignment. Note that in these 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).