Evolution of Chicken Repeat 1 (CR1) Elements: Evidence for Ancient Subfamilies and Multiple Progenitors

Thomas L. Vandergon and Marc Reitman†*

*Laboratory of Molecular Biology and †Diabetes Branch, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health

Chicken repeat 1 (CR1) is an interspersed repetitive element that is a member of the non–long terminal repeat class of retrotransposons. A data set of chicken 95 CR1 elements was compiled and the phylogeny of the 52 elements with the most complete 3′ ends was examined. We interpret the branching pattern as clustering into at least six subfamilies, designated A–F. The presence of highly similar elements within the B, C, D, and F subfamilies is evidence that a distinct progenitor has spawned each of these subfamilies. The nucleotide divergence between members of subfamily C was 5%–8%, suggesting that this subfamily has undergone a relatively recent burst of retrotransposition. The A and E subfamilies may have been spawned from ancestors of these four progenitors or from other, distinct progenitors. The consensus sequences for the six subfamilies showed considerable divergence, implying that the CR1 subfamilies are ancient. The CR1 elements in each subfamily have truncated 5′ ends and a 3′ end consisting of ⩾2 repeats of an 8-bp sequence. We estimate that there are approximately 100,000 CR1 elements in the chicken genome. Twelve CR1 sequences from avian species other than chicken were identified. Some of these sequences grouped into different subfamilies, demonstrating that multiple subfamilies existed early in avian evolution. Reptilian CR1 sequences were also identified, demonstrating that the CR1 element arose before the divergence of birds and reptiles.

Introduction

Interspersed repetitive elements constitute a significant fraction of vertebrate genomes and contribute to genome evolution in several ways. Insertion of new elements can disrupt genes (Kazazian et al. 1988), and recombination between repeated sequences can result in duplications, rearrangements, and deletions (Lehrman et al. 1985). Interspersed repetitive elements have been divided into two types, those that do not encode their means of retroposition (such as SINEs, short interspersed nucleotide elements) and those that do. The latter group contains retrovirus-like elements flanked by long terminal repeats (LTRs) and elements without such repeats (non-LTR retrotransposons). Vertebrate non-LTR retrotransposon families include L1 (or LINE-1, long interspersed nucleotide element-1) in mammals, Tx1 in *Xenopus* (Garrett et al. 1989), and chicken repeat 1 (CR1) in Aves.

The L1 repetitive element is one of the best-characterized non-LTR repetitive elements (reviewed in Hutchison et al. 1989; Martin 1991). The full-length L1 is 6–7 kb and contains two long open reading frames (ORFs), the second of which encodes a reverse transcriptase (Mathias et al. 1991). Propagation of L1 begins with transcription of a master element and translation of this RNA. The mechanistic details of the next step have not been defined, but the working hypothesis is that the reverse transcriptase recognizes its mRNA and uses it as a template for reverse transcription, priming from a nick in the chromosomal DNA (Eickbush 1992; Luan et al. 1993). Usually only a fraction of the retrotransposon RNA is inserted, resulting in truncated elements that extend a variable distance 5′ from a common 3′ end.

The CR1 element is a common repetitive sequence in the chicken, *Gallus gallus* (Stumpf et al. 1981). These elements are found in at least nine widely divergent orders in the class Aves (Chen et al. 1991). Like other repetitive elements (Korenberg and Rykowski 1988; Moyzis et al. 1989; Sainz et al. 1992), CR1 sequences are not randomly distributed. An overrepresentation of CR1s was found in a G+C-rich fraction of chicken DNA (Olofsson and Bernardi 1983), and CR1 elements account for 16% of the chicken β-globin gene cluster (Reitman et al. 1993). The 3′ end of CR1 elements consists of repeats (usually two or three) of an 8-bp sequence

Key words: non-LTR retrotransposition, evolution, repetitive DNA, phylogenetic analysis, chicken (*Gallus gallus*), CR1.

Address for correspondence and reprints: Marc Reitman, Diabetes Branch, Building 10, Room 8S-239, National Institutes of Health, Bethesda, Maryland 20892–1770.

Mol. Biol. Evol. 11(6):886–898. 1994.

0737-4038/94/1106-0007$02.00

and, unlike most non-LTR retrotransposons, does not contain an A- or AT-rich region (Silva and Burch 1989). The 5′ ends are heterogeneous, extending a variable distance upstream, with most elements <400 bp in length, as compared with the predicted full length of at least 5 kb. To date, the longest CR1 known is 2.3 kb and includes part of an ORF that is homologous to the reverse transcriptase of other non-LTR retrotransposons (Burch et al. 1993).

The current paradigm for the propagation of non-LTR interspersed repetitive elements is that a limited number of master elements exist (reviewed in Deininger et al. 1992). Daughter elements are incorporated into the genome intermittently, sometimes in large bursts (Pascale et al. 1993). These progeny are likely to be nonfunctional and to evolve without selective pressure. Within a single repetitive element family, three types of sequence differences will exist between individual elements: those due to differences between master elements, those due to sequence changes over time in a single master element, and those due to mutations in the progeny elements after retroposition (Deininger et al. 1992). By analyzing the sequences of the elements scattered throughout the genome, one can determine the sequence of the master elements, estimate their number, and gauge their age and tempo of retrotransposition.

We examine the sequences of a group of CR1 elements and determine their subfamily structure. We found that multiple subfamilies exist and determined the relationships among the subfamilies. Distinct master elements were responsible for at least four of the subfamilies. The data suggest that the CR1 family originated before the divergence of Aves and reptiles and that some CR1 subfamilies arose before speciation of the chicken.

Methods

The CR1 elements were identified in sequence searches using FASTA (Pearson 1990) from the Genetics Computer Group (GCG) sequence-analysis package (Devereux et al. 1984), in the other vertebrate subsection of the GenBank database (release 70.0 for the initial search) with a search word size of six bases and an Opt score cut off of 100. Templates for initial searches included the noncoding regions of the chicken β-globin locus (containing 19 CR1 elements; Reitman et al. 1993) and the CR1 sequence from the vitellogenin III pseudogene (GenBank no. Y00324; Silva and Burch 1989). The CR1 elements are listed in table 1 and are identified by their GenBank file designation. When more than one element was found in a file, the CR1s were given letter suffixes (e.g., X60547a). Additional searches were performed in GenBank release 78.0 using CR1 sequences from each subfamily as templates (K02907, X13894, J00907, M31321, M17627, M17964, J02714a, M17963, X61001a, J00906, and M59362).

The CR1 sequences were aligned first to the Y00324 CR1 using the GCG program BESTFIT, then optimized by inspection. The 3′ end was identified as the region containing the 8-bp direct repeats. The 5′ end was chosen by comparison with the other sequences. Most elements possessed multiple insertions, deletions, and base changes that hampered the alignment process. As observed previously, this variability was not uniformly distributed (Stumpf et al. 1984; van het Schip et al. 1987). Unique insertions within individual elements were removed from the matrix and grouped into separate characters (with each insertion given a different character state). Deletions larger than one base position were treated as a unique character state at one base position within the deleted region and the other deleted bases were treated as missing. The data matrix contained 95 aligned CR1 elements covering 298 homologous base positions. Since most of the elements were missing 5′ sequence, our analyses used the 217 positions at the 3′ end. Seven elements were <60 bp in length, 36 were 60–151 bp in length, and 52 were >151 bp in length. The seven CR1 elements shorter than 60 bp were not used in the analyses. The data matrix is available on request.

PAUP was used to infer phylogenetic relationships by the maximum-parsimony method (version 3.1.1; Swofford 1993). A heuristic search strategy was employed with initial trees constructed by random stepwise addition and branch swapping using the tree bisection-reconnection algorithm. All data reported are the result of analyses using multiple starting trees. Equal weight was given to each character. Character state changes were treated as unordered and unweighted. Trees were computed unrooted but were rooted to subfamily F for presentation.

PHYLIP was used to infer phylogenetic relationships by the neighbor-joining method (version 3.5p; Felsenstein 1993). Distances were calculated using the Jukes-Cantor method with gaps treated as missing on a data set from which the insertions had been removed. This data set was 205 characters in length. Bootstrap analyses were performed on the PAUP and PHYLIP data sets to assess the statistical significance of the tree groupings (Felsenstein 1985; Hedges 1992; Hillis and Bull 1993).

Nucleotide consensus sequences were decided by majority rule (>50% identity). In the case of a tie between two bases (each = 50%), the two-base ambiguity code was used. When more than two bases were present and each had a frequency of <50%, an *N* was used.

To estimate the time since duplication events, the observed number of nucleotide differences was corrected

Table 1
Identified CR1 Elements

GenBank Number	5′ End	3′ End	Aligned Length	Total Length	Subfamily
J02714a	9,705	9,909	205	276[a]	A
J00905	4	206	203	203	A
X61001a	394	595	202	285	A
X60547a	20,516	20,317	200	665	A
M17963	305	503	199	503	A
M58749	273	82	192	276	A
M17966	34	208	175	175	A
Z12128	207	370	164	220	A
M17967	70	231	162	162	A
M77375	223	384	162	162	A
M17965	60	214	155	155	A
D13432	186	33	154	154	A
J04028a	4,905	4,770	136	136	A
L17432p	21,353	21,222	132	132	A
M75031	436	557	122	327	A
X12347	707	612	96	96	A
M59364	194	281	88	88	A
K02904	1,397	1,314	84	92	A
M21226b	1,985	2,065	81	81	A
K02907	270	52	219	310	B
J00904	1	213	213	213[a]	B
Y00324	2,459	2,671	213	836	B
K02905	204	72	133	409	B
L17432b	6,992	6,925	68	68	B
K02906	158	107	52	456	B
X03517	658	451	208	213	C
L17432j	10,390	10,184	207	255	C
L17432s	23,652	23,448	205	205	C
X13894	1,377	1,174	204	477	C
Y00407	2,900	2,699	202	267	C
D10167	1,052	1,251	200	200	C
X14617	593	788	196	684	C
L17432d	4,105	4,300	196	196	C
J00907	18	201	184	184	C
L17432q	22,205	22,365	161	161	C
L17432k	11,474	11,677	204	300	D
X64113	202	400	199	293	D
L17432i	7,467	7,663	197	228	D
X61192	971	787	185	244	D
M17964	104	286	183	247	D
L17432r	22,766	22,603	164	164	D
L17432e	6,828	6,725	104	104	D
M15861	1,620	1,715	96	96	D
L17432g	2,227	2,315	89	89	D
M14681	3	86	84	92	D
V00436	691	774	84	84	D
M12439	1	77	77	77[a]	D
L17432l	14,581	14,641	61	61	D
M32730	1,780	1,574	207	211	E
J00906	1	192	192	192[a]	E
M10946	11,188	11,020	169	169	E
J00922	4,844	4,966	123	123	E
M87298b	351	261	91	91	E
M32732	2,492	2,406	87	87	E
X56659	7,220	7,290	71	71	E
L17432o	19,855	20,070	216	246	F
M31321	10,100	10,305	206	280	F

Table 1 (*Continued*)

GenBank Number	5′ End	3′ End	Aligned Length	Total Length	Subfamily
J02714b	24,021	24,225	205	228	F
D00702	3,232	3,428	197	227	F
X51627	499	306	194	214	F
L17432n	19,323	19,132	192	222	F
X60547b	983	802	192	202	F
M17627	1,628	1,448	181	263	F
X61197	2,250	2,426	177	177	F
X52708	6,261	6,435	175	178	F
L02537	534	361	174	174	F
M32728	92	263	172	244	F
L17432c	2,389	2,559	171	227	F
D10484	2,569	2,734	166	172	F
L17432a	1,282	1,128	155	155	F
X57998	3,755	3,907	153	153	F
X66286	4,210	4,062	149	149	F
L00062	1	138	138	138[a]	F
M24403	1	132	132	132[a]	F
M35369	2,059	2,186	128	128	F
X59080	1,102	1,221	120	298	F
J04028b	4,242	4,132	111	256	F
L17432h	7,166	7,080	87	95	F
M84460	2,729	2,646	84	147	F
M21226a	124	42	83	245	F
M21225	124	43	82	228	F
X56595	870	791	80	264	F
L17432f	7,013	7,081	69	69	F
L10366	129	65	65	65	F
M87298a	6,059	6,122	64	64	F
M59362	251	457	207	448	...
X54093	1,325	1,511	187	1,156	...
M95725	1,872	1,965	94	188	...
X04479	1,436	1,498	63	175[a]	...
J05475	1,174	1,230	57	57	...
X13607	17,544	17,598	55	55	...
D90071	4,404	4,452	49	57	...
M29448	2,003	2,045	43	43	...
L17432m	18,781	18,823	43	43	...
X61001b	3,220	3,193	28	309	...
S50878	736	425	194	224	*Coturnix coturnix* (quail)
X16232	409	590	182	247	*Coturnix coturnix* (quail)
X57379	250	425	176	425[a]	*Anas platyrhynchos* (duck)
M68975	42	213	172	213[a]	*Dromaius novahollandiae* (emu)
M68958	42	213	172	213[a]	*Grus antigone* (crane)
J00956	400	510	111	111	*Phasianus colchicus* (pheasant)
X14380	250	142	109	109	*Phylloscopus trochilus* (warbler)
X68810	263	171	93	93	*Anas platyrhynchos* (duck)
S45624	1,388	1;315	74	74	*Anser anser* (goose)
M36973	266	210	57	57	*Columba livia* (pigeon)
Z27412	154	167	13	167[a]	*Falco peregrinus* (falcon)
Z30334	152	165	13	165[a]	*Falco peregrinus* (falcon)

NOTE.—CR1 elements used in this study are identified by their GenBank file designation. Listed are the base numbers of the 5′ and 3′ ends of the analyzed regions, the number of nucleotides within the analyzed region, the total length of the CR1 element, and the subfamily designations. The last twelve elements are avian CR1 homologs, with the species of origin indicated. Subfamily designations for elements <152 bp were determined by repeated parsimony analyses with a representative sample of elements from each subfamily. Ellipses indicate that an element had an uncertain subfamily affinity.

[a] Element continues to the end of the GenBank file.

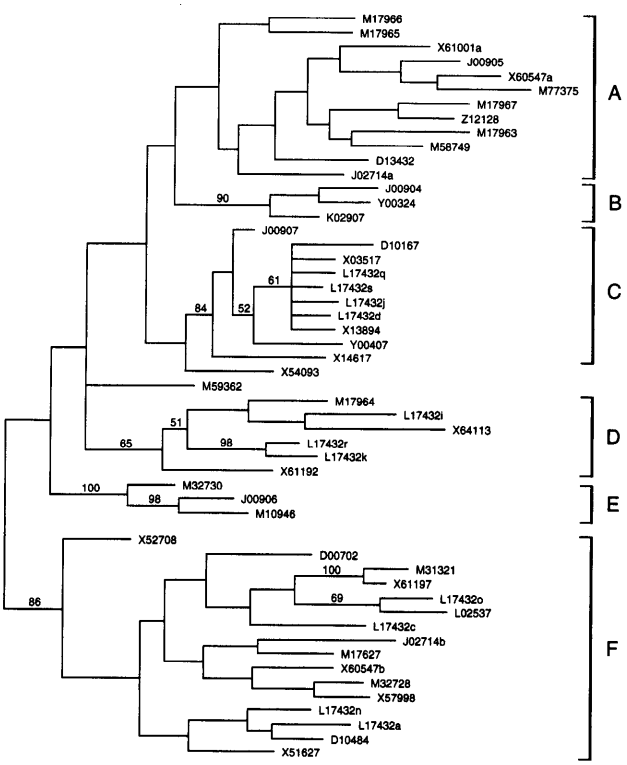


FIG. 1.—Parsimony analysis of the 52 CR1 elements containing >151 bp. The strict consensus of the 30 minimum length trees of 1576 steps from 30 PAUP searches is shown. The consensus tree has $CI_1 = 0.86$ and $CI_2 = 1.13\ e^{-76}$ (Rohlf 1982). The strict consensus tree was contained within the 50% majority-rule consensus tree from parsimony analysis of 50 bootstrap replicates, except in subfamily B, where the bootstrap analysis showed 52% support for a J00904/K02907 clade, which then joined Y00324. The percent branch support (when >50%) from the bootstrap analysis is shown. Branch lengths are proportional to evolutionary distance. Elements are named as in table 1 and the proposed subfamily groupings (A–F) are indicated.

for multiple substitutions using the Jukes-Cantor method, and the standard errors were calculated according to Kimura and Ohta (see Li and Graur 1991, pp. 50–51). The numbers of synonymous nucleotide substitutions per site were calculated using the program LWL91 (Li 1993). Since the unselected substitution rate for chicken nuclear DNA has not been reported, the substitution frequencies were converted to time using 4.61×10^{-9} synonymous substitutions per site per year (which was calculated from human-rodent comparisons; Li and Graur 1991, p. 70).

The number of CR1s per genome was estimated using the nonredundant chicken GenBank DNA files longer than 10,000 bp (J02714, X60547, X13607, M10946, M13756, M10806, M31321, and Y00407). Only the β-globin file was not included because of its anomalously high number of CR1s (19 CR1s in 21,387 [nonexonic] bp; Reitman et al. 1993). The CR1 elements were identified with BESTFIT using bp 2375–2671 from Y00324 and bp 10100–10305 from M31321 as the search strings. An element was registered if the quality score was >6 standard deviations (SDs) larger than the average best score from multiple randomizations of the input sequence. Eleven CR1s were identified in the 114,685 bp of nonexonic DNA. Using a genome size of 1.2×10^9 bp (Fasman 1975), assuming that 90% of the genome is nonexonic, and assuming that these files are representative, we estimate that there are (mean ± SD = $(11 \pm \sqrt{11})(1.2 \times 10^9 \times 0.9/114{,}685) =$) 104,000 ± 34,000 CR1s in the chicken genome.

Results

Identification of CR1 Subfamilies

Our data-base searches revealed 95 CR1 elements (table 1). To determine the phylogenetic relationship among the CR1 elements, the aligned sequences were examined by parsimony analysis (Methods). The analyses were limited to the 52 elements containing >151 bp of sequence information. This avoided generation of a huge number of possible trees due to the ambiguity caused by missing characters. The consensus of the 30 minimum-length trees found in repeated searches is shown in figure 1. We interpret the branching pattern as clustering into at least six subfamilies (denoted A–F). A bootstrap analysis (limited to 50 replicates by computer time constraints) showed support for subfamilies B, C, D, E, and F in 90%, 84%, 65%, 100%, and 86% of the replicates, respectively (fig. 1). The SDs of the bootstrap proportions, calculated according to Hedges (1992), were 4.2%, 5.2%, 6.7%, and 4.9% for subfamilies B, C, D, and F. While the elements in subfamily A were not resolved into a separate group in this analysis, they were excluded from the other five groups.

As an independent analysis of the phylogeny, we used the neighbor-joining method. The element groupings in the neighbor-joining tree (fig. 2) were very similar to the groupings from the parsimony analysis. In a bootstrap analysis of this data set, the B, C, D, E, and F subfamilies were supported in 85%, 88%, 77%, 100%, and 91% of the replicates, respectively (fig. 2). The SDs were 1.1%, 1.0%, 1.3%, and 0.9% for subfamilies B, C, D, and F. Again, all the A subfamily elements were excluded from the other subfamilies.

In both the parsimony and neighbor-joining analyses, only the M59362 and X54093 elements did not consistently group with the same elements. These two elements may be old, having lost distinct subfamily character, or they may be members of currently unidentified subfamilies. Taken together, our data support the classification of the CR1 elements into at least six subfamilies.

Evidence for Multiple Master CR1 Elements

Multiple CR1 subfamilies could arise by two mechanisms. A single progenitor element could spawn multiple subfamilies of daughter elements periodically during evolution. Alternatively, multiple progenitor elements each could produce one subfamily. Multiple subfamilies also could arise through a combination of these two mechanisms.

To distinguish among these possibilities, we searched for putative master elements in the subfamilies. Using sequences that differed from each other by <20%, we derived the "active" consensus sequences shown in figure 3. No elements in subfamilies A or E were >80% identical. Subfamily C contains six elements that differed by 5.1 to 8.4% (mean ± SD = 6.6 ± 1.1%, determined from all 15 possible comparisons), demonstrating the existence of a master element in this subfamily (C*). The similar divergences of the six elements suggests that a burst of retrotransposition occurred in this subfamily. In subfamily B, a group of elements closely related to Y00324 was recently identified (<4% sequence differences among six sequences; Burch et al. 1993). These data demonstrate the existence of another recently active master element (B*). The small amount of variation among the sequences used to derive the B* and C* consensuses (<4% and 5%–8%, respectively) and the large difference (23%) between the B* and C* sequences conclusively demonstrates that the B and C subfamilies result from propagation by two different master CR1 elements, not a single element that mutated between spawning the two subfamilies.

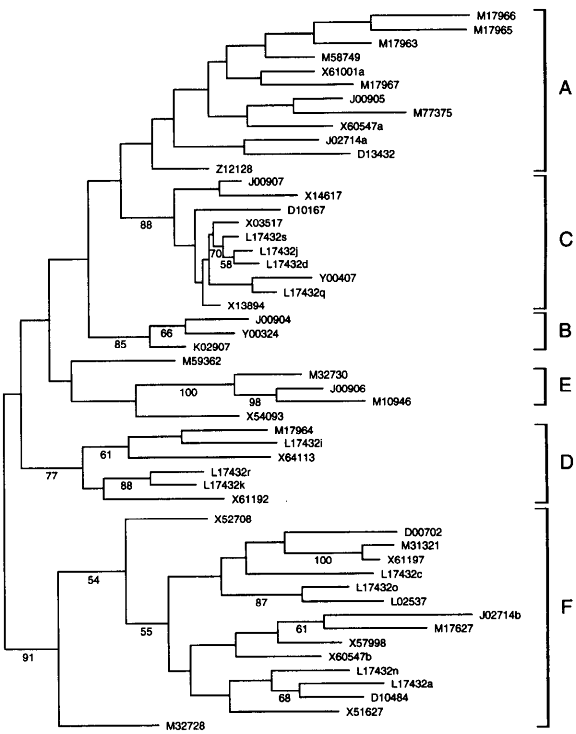
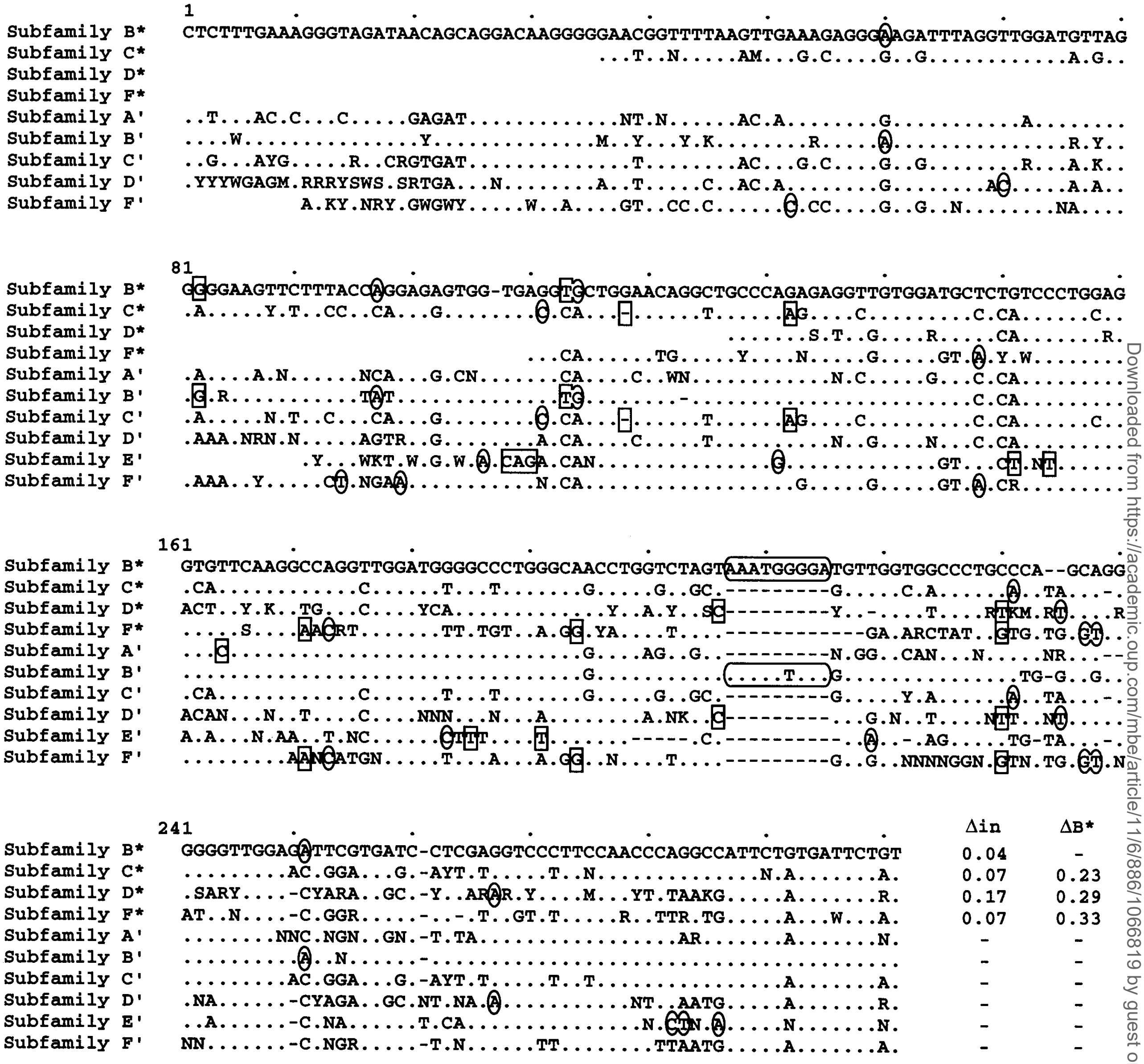


FIG. 2.—Neighbor-joining analysis of the 52 CR1 elements containing >151 bp. The neighbor-joining tree was determined using the PHYLIP programs dnadist81 and neighbor81 as described in Methods. The neighbor-joining tree was contained within the 50% majority-rule consensus tree from neighbor-joining analysis of 1,000 bootstrap replicates. The percent branch support (when >50%) from the bootstrap analysis is shown. Branch lengths are proportional to evolutionary distance; however, for calculation of the distance matrix, gaps were treated as missing and insertions were ignored. Elements are named as in table 1 and the proposed subfamily groupings (A–F) are indicated.

The frequency of synonymous nucleotide substitutions between B* and C* in the 3′ end of the putative coding region (positions 37–202 in fig. 3) was 0.53 ± 0.17. The nonsynonymous substitution frequency was threefold less (0.18 ± 0.04). The average substitution frequencies in the coding and 3′ regions (positions 206–285) were the same (0.26 ± 0.04 and 0.25 ± 0.07). Thus, both the coding and 3′ untranslated regions have been under selective pressure in the progenitor elements. From the synonymous substitution frequency, we estimate that these master elements have been diverging for roughly ((0.53 ± 0.17)/(2 × 0.00461) =) 57 ± 18 Myr (see Methods).

Data suggesting the existence of a third master element come from the subfamily D elements L17432k and L17432r, which are 17% different. The divergence



```
                   1       .         .         .         .         .         .         .         .
Subfamily B*    CTCTTTGAAAGGGTAGATAACAGCAGGACAAGGGGGAACGGTTTTAAGTTGAAAGAGGGAAGATTTAGGTTGGATGTTAG
Subfamily C*                                       ...T..N.....AM...G.C....G..G............A.G..
Subfamily D*
Subfamily F*
Subfamily A'    ..T...AC.C...C.....GAGAT.............NT.N......AC.A........G...........A........
Subfamily B'    ....W..............Y..............M..Y...Y.K........R....A...............R.Y..
Subfamily C'    ..G...AYG.....R..CRGTGAT.............T........AC...G.C....G..G........R...A.K..
Subfamily D'    .YYYWGAGM.RRRYSWS.SRTGA...N........A..T.....C..AC.A........G........AC.....A.A..
Subfamily F'              A.KY.NRY.GWGWY.....W..A....GT..CC.C......C.CC....G..G..N........NA....

                  81       .         .         .         .         .         .         .         .
Subfamily B*    GGGGAAGTTCTTTACCAGGAGAGTGG-TGAGGTGCTGGAACAGGCTGCCCAGAGAGGTTGTGGATGCTCTGTCCCTGGAG
Subfamily C*    .A.....Y.T..CC..CA...G........C.CA...-......T......AG....C..........C.CA......C..
Subfamily D*                                                  .......S.T..G....R.....CA.......R.
Subfamily F*                                   ...CA......TG.....Y....N.....G.....GT.A.Y.W.........
Subfamily A'    .A....A.N......NCA...G.CN.......CA....C..WN.............N.C.....G...C.CA........
Subfamily B'    .G.R..........TAT..............TG........-...................C.CA.........
Subfamily C'    .A.....N.T..C...CA...G........C.CA...-......T......AG....C..........C.CA......C..
Subfamily D'    .AAA.NRN.N.....AGTR..G........A.CA....C.....T............N..G....N...C.CA........
Subfamily E'              .Y...WKT.W.G.W.A.CAGA.CAN.................G...............GT...CT.NT......
Subfamily F'    .AAA..Y.....CT.NGAA.............N.CA.....................G.....G.....GT.A.CR.........

                 161       .         .         .         .         .         .         .         .
Subfamily B*    GTGTTCAAGGCCAGGTTGGATGGGGCCCTGGGCAACCTGGTCTAGTAAATGGGGATGTTGGTGGCCCTGCCCA--GCAGG
Subfamily C*    .CA............C......T...T.......G.....G..GC.---------G.....C.A......A..TA...-.
Subfamily D*    ACT..Y.K..TG...C....YCA..............Y..A.Y..SC---------Y.-....T....RTKM.RT....R
Subfamily F*    .....S....AACRT.......TT.TGT..A.GG.YA....T...-------------GA.ARCTAT..GTG.TG.GT..
Subfamily A'    ...C..............................G....AG..G..---------N.GG..CAN...N.....NR...--
Subfamily B'    ..................................G...............T...G...............TG-G..G..
Subfamily C'    .CA............C......T...T.......G.....G..GC.---------G.....Y.A......A..TA...-.
Subfamily D'    ACAN...N..T....C....NNN...N...A.........A.NK..C---------...G.N..T....NTT..NT.....
Subfamily E'    A.A...N.AA..T.NC......CTTT....T.......------.C.----------..A..-.AG.....TG-TA...-.
Subfamily F'    ..........AANCATGN.....T...A....A.GG..N....T...---------G..G..NNNNGGN.GTN.TG.GT.N

                 241       .         .         .         .         .         .              Δin     ΔB*
Subfamily B*    GGGGTTGGAGATTCGTGATC-CTCGAGGTCCCTTCCAACCCAGGCCATTCTGTGATTCTGT            0.04     -
Subfamily C*    .........AC.GGA...G.-AYT.T.....T..N..............N.A.......A.            0.07    0.23
Subfamily D*    .SARY....-CYARA..GC.-Y..ARAR.Y....M...YT.TAAKG.....A.......R.            0.17    0.29
Subfamily F*    AT..N....-C.GGR.....-..-.T..GT.T.....R..TTR.TG.....A...W...A.            0.07    0.33
Subfamily A'    ........NNC.NGN..GN.-T.TA..................AR.......A.......N.             -       -
Subfamily B'    .........A..N.....-..........................................             -       -
Subfamily C'    .........AC.GGA...G.-AYT.T.....T..T..................A.......A.           -       -
Subfamily D'    .NA......-CYAGA..GC.NT.NA.A...........NT..AATG.....A.......R.             -       -
Subfamily E'    ..A......-C.NA......T.CA................N.CTN.A.....N.......N.             -       -
Subfamily F'    NN.......-C.NGR.....-T.N......TT.........TTAATG.....A.......A.             -       -
```

FIG. 3.—Nucleotide consensus sequences of the CR1 subfamilies. The "active" consensus sequences (B*, C*, D*, and F*) were constructed using the rules given in Methods except for B*, which is GenBank no. L22152 from Burch et al. (1993). The C* is the consensus of X03517, L17432s, L17432j, L17532d, L17432q, and X13894, D* is the consensus of L17432k and L17432r, and F* is the consensus of M31321 and X61197. "Historical" consensuses (A′ to F′) were constructed from all elements in the subfamily >151 bp (table 1). Matches to the B* sequence are indicated by (.), gaps by (-), and missing data by ( ). Diagnostic and characteristic bases (table 2) are in rectangles or ovals, respectively; Δ_{in} is the average difference between the elements used to derive the active subfamily consensus; Δ_{B*} is the difference between the active subfamily consensus and the B* sequence. A "CCGT" insert between positions 148 and 149 in C′ is not shown. Positions 1 to 81 were not used in the phylogenetic analyses.

between the L17432k/L17432r consensus (D*) and the B* and C* sequences is 29% and 35%, respectively. Evidence for a fourth progenitor comes from subfamily F. Subfamily F elements M31321 and X61197 show 7% sequence difference. The consensus of M31321 and X61197 (F*) differs from the B* and C* consensuses by 33% and 38%, respectively. These data suggest the presence of distinct D and F subfamily progenitors. Sequence information from more elements will increase the robustness of this conclusion and the quality of the D* and F* sequences.

We examined the sequences flanking similar CR1 elements to determine whether the elements could have been produced by duplication events other than retrotransposition. None of the flanking sequences showed recognizable sequence similarity beyond the 3′ end,

```
                                                                                            %
                                                                                        Identity

Subfamily B*   LF-ERVDNSRTRGNGFKLKEGRFRLDVRGKFFTRRVVRCWNRLPREVVDA-LSLEVFKARLDGALGNLV*         -

Subfamily C*   LF-TRVDGDRTRGNGFKLRQGRFRLDMRRKFFPQRVVTH*NRLPKEAVDA-PSLQAFKARLDVALGSLGC       66.7
                         I      M            L        ^
                         M

Subfamily A'   LF-TRADRDRTRGNGFKLKEGRFRLDVRRKFFTQRAVRHWHXLPREAVGA-PSLEVLKARLDGALGSLSW       71.3
                            SA               S              D
                            ID               Y

Subfamily B'   LF-ERVDNSRTRGNGFKLKEGRFRLDIRGKFFTMRVVRCWN-LPREVVDARPSLEVFKARLDGALGSLV*       90.6
                V             L          V              ^         ^

Subfamily C'   LF-TRVDGDRTRGNGFKLRQGRFRLDIRRKXFTQRVVTH*NRLPKEAVDA-PSLQAFKARLDVALGSLGC       69.1
                  M   S                  M            ^

Subfamily D'   ........DRTRGNGFKLKEGRFRLDIRKXXFTVRVVKHWHRLPREVVDA-PSLETFKVRLDXALSNLI*       69.1
                        S                                   D  G       LM   E      ME
                                                                       VR

Subfamily E'   ...............................FTVRVTEHWNRLPGEVVES-PSLEIFKTWLDAFLCNLL*       53.9
                                              S LWE  Q             Y    * C

Subfamily F'   .......RDRTRGSGLKLHQGRFRLDIRKNFFSERVVRHWNRLPREVVES-PSLEVFKKHXDVALRDLV*       62.4
                      SV             L  E           M                    N        M
                                                    K                             V

Ray            LFPLRVGKIQTRWHGLRLKGEKFRGNMRGNFFTQRVVGGWNELPAEVVDA-GSILMFKEKLD........       54.1
Lizard         LF-SAALQTRTRNNGFKLQERRFHLNIRKNFLTVRAVRQWNSLPGAVVEV-PSLEAFKQRLDGHLSGVL*       54.4
Snake          LP-QRRGSQAIFQSRWKLTKERSNLELRRNFLTVRTINQWNNLPPEAVNA-PTLEVFKKRLDSHFSEMV*       39.7
```

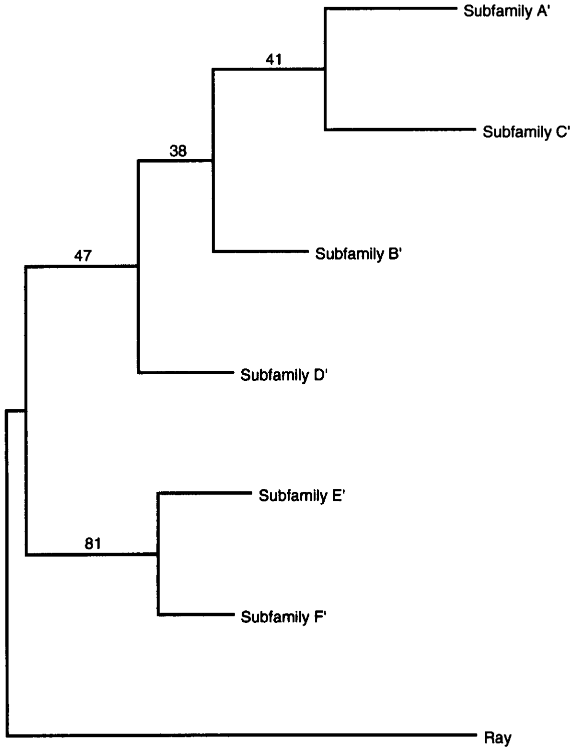


FIG. 4.—Conceptual translation of the ORF at the 3′ end of the CR1 consensuses. *Above,* Subfamily nucleotide sequences (B*, C*, A′–F′, from fig. 3) were translated and aligned. Residues are indicated by single-letter amino acid codes, with gaps indicated by a dash (-) and termination codons indicated by an asterisk (*). For positions with one, two, or three possible residues, all are shown, while those having more than three alternatives are indicated by an X. Amino acids present in ⩾50% of the CR1 subfamilies are shown in bold. The percent sequence identity relative to the B* consensus is listed on the right. Arrows mark the positions of frame shifts introduced to increase similarity. The ray, lizard, and snake sequences are from X56517, L31503, and D13384, respectively. *Right,* Aligned sequences used for a bootstrapped PAUP analysis using the branch and bound method. The data in the panel above were used, except that residue positions with any ambiguity were treated as missing. The percent branch support in 1,000 replicates is shown. The ray sequence is used as an outgroup.

consistent with the elements' creation by retrotransposition. In summary, there is strong evidence for the relatively recent activity of two CR1 progenitor elements and suggestive evidence for two other previously active progenitors.

A "historical" nucleotide consensus sequence for each subfamily was derived from all elements >151 bp (A′ to F′ in fig. 3). These consensuses represent the subfamily averages and do not necessarily correspond to ancestral sequences. The historical consensus sequences were conceptually translated and aligned (fig. 4, *top panel*). In this region, corresponding to the C terminus of the reverse transcriptase ORF, the CR1 subfamilies are 54%–71% identical to B*.

Table 2
Characteristic and Diagnostic Bases of the CR1 Subfamilies

Subfamily	Characteristic Base[a] (position/base)	Diagnostic Base (position/base)
A	. . .	164/C
B	60/A	82/G
	97/A	113/T
	114/G	
	207-215/9-bp insert	
	251/A	
C	111/C	132/A
	118/Δ	
	231/A	
D	70/C	206/C(or A)
	235/T	230/T
	267/A	
E	106/A	108/C
	131/G	109/A
	183/C	110/G
	219/A	151/T
	282-3/CT	154/T
	286/A	185/T
		191/T
F	52/C	171/A(or G)
	94/T	194/G
	99/A	230/G
	148/A	
	173/C	
	237-8/GT	

NOTE.—Characteristic bases are those found in >50% of the elements in a subfamily but <50% of elements in all other subfamilies. Diagnostic bases are those found in ⩾80% of the elements in a subfamily, but ⩽20% of elements in all other subfamilies. Numbers correspond to the aligned positions in fig. 3.

[a] Δ indicates a 1-bp deletion.

To determine the relationships between the CR1 subfamilies, multiple bootstrapped parsimony analyses were performed on the translated sequences. While the basic phylogeny (fig. 4, *right panel*) did not change, the statistical support for individual clusters varied, depending on which of the nonchicken elements (if any) were included. Typically, subfamilies E′ and F′ grouped together in ~80% of the bootstrap replicates and an A′ + B′ + C′ clade was supported in ~50% of the replicates. When included, the lizard element grouped with the E′ subfamily in ~80% of the replicates.

We have determined a set of "characteristic" and "diagnostic" base identities for each subfamily (defined in table 2 and fig. 3). Subfamily assignment cannot be made on the basis of a single position of the aligned sequence but can be made using multiple positions. A characteristic change in the C* sequence is a deletion of a G (at position 118 in fig. 3), creating a termination codon and removing the final 29 amino acids of the ORF. This deletion occurred in the master element, before production of the seven most recent progeny (D10167–X13894; see fig. 2). This suggests that the 29 amino acids are not essential for retrotransposition, although any missing functions could have been supplied from another master element.

CR1s in Other Avian Species

Twelve CR1s were identified in avian species other than chicken (table 1). The eight longest elements covering the region analyzed were used in parsimony analyses to assess whether the CR1 subfamilies antedate speciation of the chicken (fig. 5). Two quail CR1s were more similar to their subfamilies (E and F) than to each other. The same was true of two duck CR1s (from subfamilies ABCD and EF). The most parsimonious explanation for these results is that the relevant subfamily progenitors existed in the common ancestor of chickens and quails, and chickens and ducks, respectively.

Related Sequences in Nonavian Species

In addition to the avian CR1s, our GenBank searches revealed other similar sequences. Most striking were two *Anolis carolinensis* (lizard) sequences (L31503, bases

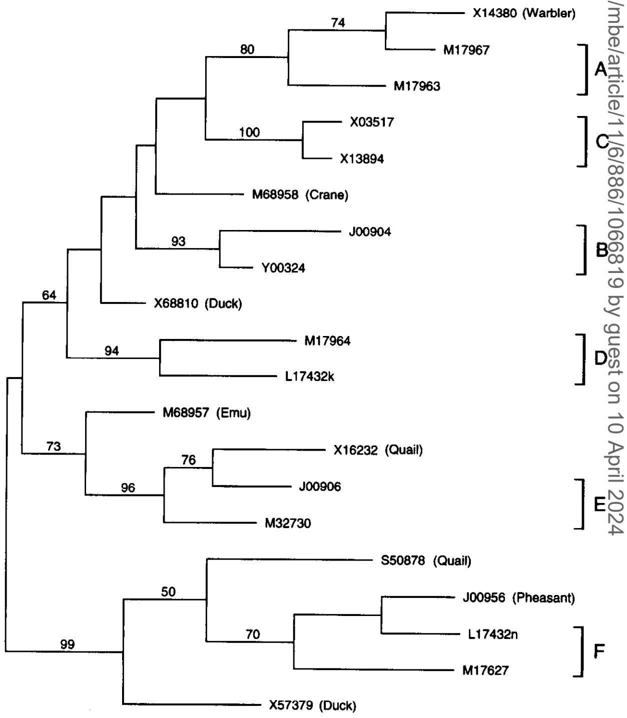


FIG. 5.—The single minimum length tree found for 20 avian CR1 elements (12 chicken and eight nonchicken) from 40 PAUP searches. The tree has a length of 816 steps with a CI = 0.633. The percent branch support (when >50%) from the bootstrap analysis (500 replicates) is shown. Branch lengths are proportional to evolutionary distance. Elements are named as in table 1 and the chicken elements are labeled with their CR1 subfamily (A–F).

9015–8640 and 9715–9645) containing characteristic CR1 3′ ends ((TATTCTAT)$_1$(GATTCTAT)$_{1\ \mathrm{or}\ 2}$). The overlapping regions were identical except for a 2-bp indel, suggesting a recent origin. A 266-bp region was 59% identical to the CR1 B* consensus sequence, with no indels. The C-terminal end of the ORF was 54% identical to the translated B* sequence (fig. 4, top panel).

Sequences with similarity to CR1 were also found in *Torpedo spp.* (ray) and *Trimeresurus flavovirides* (snake). These putative repetitive elements do not contain 8-bp repeats at their 3′ ends but do show an impressive similarity to CR1 (fig. 4, *top panel*).

Sequence similarity to multiple members of the *Geoclemys reevessi* (tortoise) PolIII/SINE family of repetitive elements (Endoh et al. 1990) was also found. The PolIII/SINE element is ~185 bp in length with the 5′-most 75 bp showing a ~70% similarity to tRNA$^{\mathrm{lys}}$ (Endoh et al. 1990). The next ~100 bp show sequence similarity (65%–70%) to the 3′ end of CR1.

Properties of the Chicken CR1 Elements

Having identified a large group of CR1 elements, we examined the properties of the 95 chicken elements in our data set. The mean and median element lengths were 206 and 184 bp, respectively. There were no apparent preferred regions for 5′ truncation. The number of 8-bp repeats at the 3′ end of the elements was one in 11%, two in 49%, three in 33%, and four in 6% of the elements. In the elements having ⩾2 repeats, the 5′-most repeats have the form 5′-SATTCTRTGATTCTRT-3′. The six subfamilies individually showed similar distributions of element lengths and numbers of 8-bp repeats. We were unable to consistently identify target-site duplications at the ends of the elements, even in the least diverged elements.

Discussion

Multiple CR1 Subfamilies and Progenitors

We propose that the chicken CR1 elements group into at least six subfamilies (A, B, C, D, E, and F). The bootstrap support for the B through F subfamilies was 65% to 100%. A recent analysis has shown that, under conditions similar to ours, bootstrap proportions of ⩾70% correspond to a ⩾95% probability that the corresponding grouping is found in the true phylogeny (Hillis and Bull 1993). We have also presented strong evidence for two discrete progenitors (B* and C*) and weaker evidence for two more (D* and F*).

Our results lead us to postulate the following model for the evolution of the CR1 family. The ancestral progenitor duplicated early in CR1 evolution, producing elements that in turn we ancestral to the ABCD and EF groups. Two more progenitor duplications occurred in the ABCD group. One possibility is that the ABCD ancestor duplicated to produce the ABC and D progenitors, and later the ABC ancestor duplicated to form the B and C masters. It is not clear if these multiple master elements arose by DNA duplications or by retrotransposition. Within the ABCD group, a distinct master element in each of the B, C, and D subfamilies produced all the elements in its subfamily. The lack of closely related elements within the A subfamily suggests that it consists of elements spawned from one or more ancestors of the B and/or C (or less likely, D) subfamilies. The A subfamily structure suggests that with more information it might resolve into two or more subfamilies. The F subfamily structure contains evidence of an active element and also suggests that this subfamily may consist of multiple subfamilies. To date, no closely related sequences have been found in subfamily E. Thus the E subfamily is probably derived from a master element, related to an ancestor of the F master element. Of course, in both the A and E subfamilies, the discovery of minimally divergent elements would demonstrate the existence of distinct progenitors.

Multiple subfamilies could arise from the activity of more than one master element or from sequential amplification from a single, evolving master. Like CR1s, L1s show subfamily structure (*Galago,* Lloyd and Potter 1988; rabbit, Price et al. 1992; mouse, Jubier-Maurin et al. 1992; rat, Pascale et al. 1990), but there is controversy about the number of L1 master elements. In *Galago,* strong evidence exists for two master elements (Stanhope et al. 1993), while the origin of the human L1 subfamilies has been attributed both to multiple master elements (Skowronski and Singer 1986; Scott et al. 1987) and to the evolution of a single master (Jurka 1989). The existence of multiple CR1 subfamilies, derived from multiple progenitors, suggests that multiple master elements may be common in non-LTR retrotransposon families.

An Ancient Origin for the CR1 Element

Presuming that the duplicated CR1 elements are not under active selection, they will lose identifying characteristics as a result of random mutation. Assuming a mutation rate equal to a midrange vertebrate unselected substitution rate, two CR1 sequences would lose detectable similarity in roughly 40 Myr. Thus, the elements we studied are not this old. However, they contain information concerning the evolutionary conservation of their progenitors.

We conclude that the *A. carolinensis* elements are CR1s on the basis of their remarkable sequence similarity to the avian elements (including the characteristic 8-bp repeats). Thus, the CR1 family is presumed to have existed in the last common ancestor of Aves and *Anolis.*

The similarity between the avian CR1s and the snake and ray sequences suggests a common ancestor, although these sequences do not have 8-bp repeats at their 3′ end. Whether the snake and ray elements also represent CR1 family members or are just homologous non-LTR retrotransposons requires more sequence information and an understanding of CR1 biology.

CR1 Subfamilies Are Also Ancient

The high degree of divergence between the subfamilies suggests that their existence is ancient. For example, the divergence between the B and E subfamilies is similar to that between the B subfamily and the lizard and ray homologs (fig. 4, *top panel*). Comparison of avian CR1s indicated that some subfamilies were established before the speciation of chickens. Although we cannot rule out horizontal transfer as a mechanism for the dispersion of CR1 subfamilies among species, we consider this unlikely since it would have had to occur multiple times. We have no evidence for (or against) the existence of multiple subfamilies prior to the divergence of Aves, although this could be addressed by an analysis of more reptilian CR1s.

Tempo of CR1 Activity

Our data demonstrate that multiple progenitor elements were active over long and overlapping time periods. There was a relatively recent burst of activity from the C master, as evidenced by multiple elements with 5%–8% divergence. The B subfamily contains elements with even less divergence (Burch et al. 1993), suggesting even more recent transposition events. We do not have rigorous proof that the CR1 progenitors are currently active (e.g., by showing insertion at a site unoccupied in the previous generation). However, in view of their longevity and recent activity, it is likely that some CR1 progenitors are currently competent for retrotransposition. Interestingly, the other CR1 subfamilies apparently did not undergo a similar degree of amplification at the same time as the C subfamily. This suggests that the retroposition rates of the different master elements are independent of each other.

The L1 master elements also exhibit different rates of propagation, with periods of active transposition interspersed with periods of relative quiescence. For example, in rodent L1 evolution, few intermediates exist between the active L1 ancestor (Lx) and its active modern mouse and rat L1 descendants (Pascale et al. 1993). Similarly, the tempo of retroposition in voles is quite different from that in mice (Vanlerberghe et al. 1993). Punctuated amplification has also been proposed for two SINE families, the Alu family in humans (Deininger et al. 1992) and the C elements in rabbits (Krane et al. 1991). The mechanisms regulating the tempo of master-element propagation are not understood (Deininger et al. 1992).

Two specific examples of CR1 insertion from master elements have been identified, a B subfamily element found in the vitellogenin III pseudogene and an A subfamily element found in the second intron of the ε-globin gene. In the former case, the CR1 is not present in the vitellogenin III gene from which the pseudogene diverged ∼16 Mya (Burch et al. 1993). In the latter, the CR1 is not present in the duck ε-globin gene, from which the chicken gene diverged by speciation ∼70–90 million years ago (Cracraft and Mindel 1989).

Geoclemys reevessi PolIII/SINE elements and CR1

We found an intriguing sequence similarity between CR1 elements and a tortoise SINE repetitive element family. This observation and the presence of CR1 in *Anolis* suggest that a CR1-like retrotransposon will be found in *Geoclemys reevessi.* The most likely explanation for the PolIII/SINE sequence is that a non-LTR retrotransposon homolog of CR1 was inserted into the SINE master element, becoming a part of the SINE master. Alternatively, a master element was created by insertion of this homolog into a tRNA gene. Two other composite transposable elements have been noted in vertebrates: a *Galago* SINE family that is part alu and part L1 monomer (Daniels and Deininger 1991) and a mouse L1 subfamily that is a fusion of two distinct L1 subfamilies (Adey et al. 1991). It has been postulated that SINE duplication uses the retroposition machinery of non-LTR retrotransposons (Eickbush 1992). Thus a CR1-like reverse transcriptase could participate in *Geoclemys reevessi* PolIII/SINE retroposition through recognition of the homologous 3′ sequence.

Properties of the CR1 Elements

We have assembled the largest group of CR1s to date with 95 chicken elements, 12 from other avian species, and two from *Anolis.* Our data set confirmed that the CR1 elements are very short, that integration site duplications are frequently not detectable (but see Silva and Burch 1989), and that the 3′ end does not contain the A- or AT-rich regions found in the other vertebrate non-LTR retrotransposons. The 3′ end consists of 1–4 repeats of an 8-bp sequence. The elements with only one 8-bp repeat tended to be more diverged, suggesting that ⩾2 repeats is the rule. The above attributes were observed in all of the CR1s, implying that they also characterize the ancestral CR1 elements.

The number of CR1 elements in the chicken genome was estimated previously at 7,000 to 30,000 by hybridization using single CR1 elements as probes (Stumpf et al. 1981; Hache and Deeley 1988; Shapira et al. 1991; Burch et al. 1993). Our identification of multiple, divergent CR1 subfamilies suggests that these estimates are likely to be low. From sequence analysis, we estimate that there are ~100,000 CR1 elements per haploid genome (Methods), which account for ~2% of the genome (using the mean element length of 206 bp).

In summary, we show that multiple CR1 subfamilies exist and were derived from multiple progenitors. The subfamilies are ancient, antedating the speciation of the chicken. The existence of CR1s in both avian and reptilian species suggests an origin for this element before the divergence of these vertebrate classes. Thus, CR1 elements have a long history of influencing genome structure and evolution.

Acknowledgments

We thank K. Usdin for insightful suggestions, F. Ouellette for identifying the long chicken GenBank files, and A. Ginsberg for comments on the manuscript. M.R. is a Lucille P. Markey Scholar, and this work was supported in part by a grant from the Lucille P. Markey Charitable Trust.

LITERATURE CITED

ADEY, N. B., S. A. SCHICHMAN, C. A. HUTCHINSON, III and M. H. EDGELL. 1991. Composite of A and F-type 5′ terminal sequences defines a subfamily of mouse LINE-1 elements. J. Mol. Biol. **221**:367–373.

BURCH, J. B. E., D. L. DAVIS, and N. B. HAAS. 1993. Chicken repeat 1 elements contain a *pol*-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. Proc. Natl. Acad. Sci. USA **90**:8199–8203.

CHEN, Z.-Q., R. G. RITZEL, C. C. LIN, and R. B. HODGETTS. 1991. Sequence conservation in avian CR1: an interspersed repetitive DNA family evolving under functional constraints. Proc. Natl. Acad. Sci. USA **88**:5814–5818.

CRACRAFT, J., and D. P. MINDELL. 1989. The early history of modern birds: a comparison of molecular and morphological evidence. Pp. 389–403 *in* B. FERNHOLM et al., eds. The heirarchy of life. Elsevier, New York.

DANIELS, G. R., and P. L. DEININGER. 1991. Characterization of a third major SINE family of repetitive sequences in the galago genome. Nucleic Acids Res. **19**:1649–1656.

DEININGER, P. L., M. A. BATZER, C. A. HUTCHISON, III, and M. H. EDGELL. 1992. Master genes in mammalian repetitive DNA amplification. Trends Genet. **8**:307–311.

DEVEREUX, J., P. HAEBERLI, and O. SMITHIES. 1984. A comprehensive set of sequence analysis programs for VAX systems. Nucleic Acids Res. **12**:387–395.

EICKBUSH, T. H. 1992. Transposing without ends: the non-LTR retrotransposable elements. New Biol. **4**:430–440.

ENDOH, H., S. NAGAHASHI, and N. OKADA. 1990. A highly repetitive and transcribable sequence in the tortoise genome is probably a retroposon. Eur. J. Biochem. **189**:25–31.

FASMAN, G. D. 1975. Handbook of biochemistry and molecular biology. Vol. **2**: Nucleic Acids. P. 2. 3d. edition. CRC, Cleveland.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1993. PHYLIP—phylogeny inference package, version 3.5. University of Washington, Seattle.

GARRETT, J. E., D. S. KNUTZON, and D. CARROLL. 1989. Composite transposable elements in the *Xenopus laevis* genome. Mol. Cell. Biol. **9**:3018–3027.

HACHE, R. J. G., and R. G. DEELEY. 1988. Organization, sequence and nuclease hypersensitivity of repetitive elements flanking the chicken apoVLDLII gene: extended sequence similarity to elements flanking the chicken vitellogenin gene. Nucleic Acids Res. **16**:97–113.

HEDGES, S. B. 1992. The number of replications needed for accurate estimation of the bootstrap *P* value in phylogenetic studies. Mol. Biol. Evol. **9**:366–369.

HILLIS, D. M., and J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. **42**:182–192.

HUTCHISON, C. A., III, S. C. HARDIES, D. D. LOEB, W. R. SHEHEE, and M. H. EDGELL. 1989. Lines and related retrotransposons: long interspersed repeated sequences in the eucaryotic genome. Pp. 593–617 *in* D. E. BERG and M. M. HOWE, eds. Mobile DNA. American Society for Microbiology. Washington, D.C.

JUBIER-MAURIN, V., G. CUNY, A.-M. LAURENT, L. PAQUEREAU, and G. ROIZES. 1992. A new 5′ sequence associated with mouse L1 elements is representative of a major class of L1 termini. Mol. Biol. Evol. **9**:41–55.

JURKA, J. 1989. Subfamily structure and evolution of the human L1 family of repetitive sequences. J. Mol. Evol. **29**: 496–503.

KAZAZIAN, H. H., JR., C. WONG, H. YOUSSOUFIAN, A. F. SCOTT, D. G. PHILLIPS, and S. E. ANTONARAKIS. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature **332**:164–166.

KORENBERG, J. R., and M. C. RYKOWSKI. 1988. Human genome organization: Alu, Lines, and the molecular structure of metaphase chromosome bands. Cell **53**:391–400.

KRANE, D. E., A. G. CLARK, J.-F. CHENG, and R. C. HARDISON. 1991. Subfamily relationships and clustering of rabbit C repeats. Mol. Biol. Evol. **8**:1–30.

LEHRMAN, M. A., W. J. SCHNEIDER, T. C. SUDOHOF, M. S. BROWN, J. L. GOLDSTEIN, and D. W. RUSSELL. 1985. Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. Science **227**:140–146.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

LI, W.-H., and D. GRAUR. 1991. Fundamentals of molecular evolution. Sinauer, Sunderland, Mass.

LLOYD, J. A., and S. S. POTTER. 1988. Distinct subfamilies of primate L1Gg retroposons, with some elements carrying tandem repeats in the 5′ region. Nucleic Acids Res. **16:** 6147–6156.

Luan, D. D., M. H. Korman, J. L. Jakubczak, and T. H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell **72:**595–605.

Martin, S. L. 1991. LINEs. Curr. Opin. Genet. Dev. **1:**505–508.

Mathias, S. L., A. F. Scott, H. H., Kazazian, Jr., J. D. Boeke, and A. Gabriel. 1991. Reverse transcriptase encoded by a human transposable element. Science **254:**1808–1810.

Moyzis, R. K., D. C. Torney, J. Meyne, J. M. Buckingham, J.-R. Wu, C. Burks, K. M. Sirotkin, and W. B. Goad. 1989. The distribution of interspersed repetitive DNA sequences in the human genome. Genomics **4:**273–289.

Olofsson, B., and G. Bernardi. 1983. The distribution of CR1, an *Alu*-like family of interspersed repeats, in the chicken genome. Bioch. Biophys. Acta **740:**339–341.

Pascale, E., C. Liu, E. Valle, K. Usdin, and A. V. Furano. 1993. The evolution of long interspersed repeated DNA (L1, LINE 1) as revealed by the analysis of an ancient rodent L1 DNA family. J. Mol. Evol. **36:**9–20.

Pascale, E., E. Valle, and A. V. Furano. 1990. Amplification of an ancestral mammalian L1 family of long interspersed repeated DNA occurred just before the murine radiation. Proc. Natl. Acad. Sci. USA **87:**9481–9485.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. **183:**63–98.

Price, D. K., J. A. Ayres, D. Pasqualone, C. H. Cabell, W. Miller, and R. C. Hardison. 1992. The 5′ ends of LINE1 repeats in rabbit DNA define subfamilies and reveal a short sequence conserved between rabbits and humans. Genomics **14:**320–331.

Reitman, M., J. A. Grasso, R. Blumenthal, and P. Lewit. 1993. Primary sequence, evolution and repetitive elements of the *G. gallus* (chicken) β-globin cluster. Genomics **18:** 616–626.

Rohlf, F. J. 1982. Consensus indices for comparing classifications. Math. Biosci. **59:**131–144.

Sainz, J., L. Pevny, Y. Wu, C. R. Cantor, and C. L. Smith. 1992. Distribution of interspersed repeats (*Alu* and *Kpn*) on *Not*I restriction fragments of human chromosome 21. Proc. Natl. Acad. Sci. USA **89:**1080–1084.

Scott, A. F., B. J. Schmeckpeper, M. Abdelrazik, C. T. Comey, B. O'Hara, J. P. Rossiter, T. Cooley, P. Heath, K. D. Smith, and L. Margolet. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics **1:**113–125.

Shapira, E., S. Yarus, and A. Fainsod. 1991. Genomic organization and expression during embryogenesis of the chicken CR1 repeat. Genomics **10:**931–939.

Silva, R., and J. B. E. Burch. 1989. Evidence that chicken CR1 elements represent a novel family of retroposons. Mol. Cell. Biol. **9:**3563–3566.

Skowronski, J., and M. F. Singer. 1986. The abundant LINE-1 family of repeated DNA sequences in mammals: genes and pseudogenes. Cold Spring Harbor Symp. Quant. Biol. **51:**457–464.

Stanhope, M. J., D. A. Tagle, M. S. Shivji, M. Hattori, Y. Sakaki, J. L. Slightom, and M. Goodman. 1993. Multiple L1 progenitors in prosimian primates: phylogenetic evidence form ORF1 sequences. J. Mol. Evol. **37:**179–189.

Stumph, W. E., P. Kristo, M.-J. Tsai, and B. W. O'Malley. 1981. A chicken middle-repetitive DNA sequence which shares homology with mammalian ubiquitous repeats. Nucleic Acids Res. **9:**5383–5397.

Stumph, W. E., C. P. Hodgson, M.-J. Tsai, and B. W. O'Malley. 1984. Genomic structure and possible retroviral origin of the chicken CR1 repetitive DNA sequence family. Proc. Natl. Acad. Sci. USA **81:**6667–6671.

Swofford, D. L. 1993. PAUP: phylogenetic analysis using parsimony, version 3.1.1. Illinois Natural History Survey, Champaign.

van het Schip, F., J. Samallo, F. Meijlink, M. Gruber, and G. Ab. 1987. A new repetitive element of the CR1 family downstream of the chicken vitellogenin gene. Nucleic Acids Res. **10:**4193–4202.

Vanlerberghe, F., F. Bonhomme, C. A. Hutchison, III, and M. H. Edgell. 1993. A major difference between the divergence patterns within the Lines-1 families in mice and voles. Mol. Biol. Evol. **10:**719–731.

Thomas Eickbush, reviewing editor

Received April 6, 1994

Accepted June 22, 1994