Analysis of CR1 repeats in the zebra finch genome

George E. Liu, Yali Hou* and Twain Brown

Bovine Functional Genomics Laboratory, ANRI, ARS, USDA,
Beltsville, Maryland 20705, USA
***Also affiliated with Department of Animal and Avian Sciences, University of Maryland,**
College Park, Maryland 20742, USA

ABSTRACT

Most bird species have smaller genomes and fewer repeats than mammals. Chicken Repeat 1 (CR1) repeat is one of the most abundant families of repeats, ranging from ~133,000 to ~187,000 copies accounting for ~50 to ~80% of the interspersed repeats in the zebra finch and chicken genomes, respectively. CR1 repeats are believed to have arisen from the retrotransposition of a small number of master elements, which gave rise to multiple CR1 subfamilies in the chicken. In this study, we performed a global assessment of the divergence distributions, phylogenies, and consensus sequences of CR1 repeats in the zebra finch genome. We identified and validated 34 CR1 subfamilies and further analyzed the correlation between these subfamilies. We also discovered 4 novel lineage-specific CR1 subfamilies in the zebra finch when compared to the chicken genome. We built various evolutionary trees of these subfamilies and concluded that CR1 repeats may play an important role in reshaping the structure of bird genomes.

Keywords: CR1 repeats, comparative genomics, zebra finch, genome.

1. INTRODUCTION

The zebra finch (*Taeniopygia guttata*) is a songbird belonging to the large avian order Passeriformes. It is an important model for studying neuroscience, development, and evolution of learned vocalizations and communication. Although overall genome structures are similar in the zebra finch and chicken, they differ in chromosomal rearrangements, lineage-specific gene family expansions and other aspects [1]. In this study, we performed a global analysis of CR1 repeats by comparing the zebra finch and chicken genomes.

Most bird species have smaller genomes and fewer repeats than mammals. The genome size of these birds (~1,200 Mb) is approximately 40% of the size of the human genome. Within the repeatmaskable regions, repetitive elements make up only 9-10%, as compared to the 45% in the human genome [1-3]. As a non-LTR (long terminal repeat) retrotransposon, CR1 is one of the most abundant repeat families, belonging to long interspersed nuclear elements (LINEs). There are over 187,000 copies of CR1 repeats in the chicken genome, accounting for ~74% of its interspersed repeats [2]. On the other hand, there are over 133,000 CR1 repeats in the zebra finch genome, making up 48% of its interspersed repeats. Recent work increasingly recognizes that CR1 elements have a greater impact than expected on the evolution of both the chicken and zebra finch genomes [1,4]. It has been suggested that the relatively small genome size of birds in general may reflect selective pressure to optimize metabolism and to minimize the amount of repetitive DNA [3] .

A full-length CR1 is estimated to be ~4.5 kb and contains a (G+C)-rich internal promoter region, followed by two protein-coding sequences [2]. The exact function of ORF-1 is unknown. ORF-2 encodes endonuclease and reverse transcriptase domains and catalyzes the critical step of the retrotransposition process. The high specificity of ORF2 reverse transcriptase activity may explain the lack or lower numbers of other nonautonomous elements, including SINEs and pseudogenes in the chicken or zebra finch genomes, respectively [1,2]. Due to the truncation at their 5' ends, most CR1 fragments are left with a few hundred base pairs at their 3' ends, suggesting the premature termination of reverse transcription [4]. Unlike mammalian L1 elements, CR1 elements do not create target site duplications. Although their 5'-UTR are divergent, CR1's 3' UTR are well conserved, ending with 2-4 copies of 8bp repeat (ATTCTRTG) and lacking a polyadenylic acid (poly A) tail, in all chicken CR1 subfamilies, as well as in the turtle CR1 and the ancient L3 element [2].

CR1 elements are divided into subfamilies based on the extent of sequence diversity. The RECON analysis of the chicken genome generated a total of 22 CR1 subfamilies, including 11 full-length (4.1- 4.8 kb) and 11 additional (3' end 1.0 - 1.1 kb) CR1 subfamilies when only 3' end sequences were considered [2]. The evolutionary ages of chicken CR1 subfamilies have also been determined by a transposon-interruption analysis [4,5]. We carried out a phylogenetic analysis of the ORF2 sequences at a fine scale and identified 57 chicken CR1 subfamilies in the chicken genome [6]. The combined evidence indicated that several remarkably divergent CR1 elements have been existing and active in chickens, whereas in mammals a single lineage of L1 has been dominant [2]. The mixing of turtle and chicken CR1 elements in this ORF2-based phylogenetic tree also suggested that the oldest CR1 elements may predate the reptile-bird speciation [2]. Based on CR1 subfamily sequence diversity, a major burst in CR1 amplification was estimated to occur approximately 45 mya and since then gradually declined [4]. However, it is not clear whether these CR1s are still active in these birds at present.

To date, characterization of CR1 repeats has been mainly focused on the chicken [2]. For other birds, most studies have been based on PCR cross-amplification among diverse bird taxa and, therefore, are potentially biased to either conserved regions or limited to closely related species. Due to their unidirectional mode of evolution, CR1 insertions have been used as largely homoplasy-free character states in cladistic analyses of reptiles and birds like chicken, geese and penguins. CR1 insertion loci have also been used to clarify relationships among rockfowls, crows, and ravens.

Pevzner and colleagues identified more human *Alu* subfamilies at a much finer resolution than previously recognized using a novel

method (Alucode) [7]. We have successfully adapted it to analyze primate Alu repeats [8] and chicken CR1 repeats [6]. This method first splits repeat subfamilies based on "biprofiles", i.e. linkage of pairs of nucleotide values and then used the calibration of mutation rates to split subfamilies containing overrepresented individual mutations. In this study, we applied this method to further characterize the zebra finch CR1 elements and identified 34 CR1 subfamilies of which 22 are novel. In addition, we discovered 4 lineage specific CR1 repeat elements in the zebra finch. Considering the zebra finch diverged from the chicken approximately 100 million years ago (mya), our comparative analysis revealed that the activities of CR1 vary in different bird lineages. The new classification of the zebra finch CR1 repeats will provide insight into their diversity and biology.

2. MATERIALS AND METHODS

CR1 element identification: The zebra finch genome assembly (taeGut1) and repeat annotations were downloaded from the UCSC genome browser (http://genome.ucsc.edu/). To investigate the relationship between CR1 subfamilies, repeats were detected as previously described [6]. Briefly, we utilized 57 chicken [6] and 21 zebra finch consensus sequences of the previously described subfamilies from Repbase (http://www.girinst.org/, version 14.08 and [2]). We detected CR1 repeat elements using the slow search option (-s) of RepeatMasker (version open-3.1.0) [9]. For this study, only the 3' terminal region of ORF-2 was used because most CR1 elements are found as short fragments of the 3' region less than 1,000 bp [2]. The default zebra finch CR1 consensus sequences were trimmed to 465 bp from nucleotide positions 3,944 to 4,408 (accession no. U88211), corresponding to amino acid positions 818 to 972 of the consensus protein for ORF-2 (accession no. AAC60281) [3]. We selected all zebra finch CR1 repeats (6,759) with at least 98% length of the 465 bp consensus segments after excluding those containing ambiguous bases (i.e. Ns). Sequence divergences of CR1 elements from the consensus sequences were computed by RepeatMasker as described before. Divergence levels reported by RepeatMasker were corrected for the CpG content of each repeat by $D_{CpG} = D/(1+9F_{CpG})$, where F_{CpG} is the frequency of CpG dinucleotides in the consensus, and D_{CpG} is further corrected with the Jukes–Cantor formula for multiple substitutions [4]. We calculated the means and standard deviations of the divergence distributions. We used the mean of 9.0 substitutions/site (%) as the threshold to define "young" or "ancient" subfamilies.

Phylogenetic analyses: For major branches within phylogenetic trees, multiple sequence alignments were performed with ClustalW at default settings. MEGA [10] was used to construct neighbor-joining (NJ) trees using the Kimura 2-parameter model. The minimum spanning trees of zebra finch CR1 subfamilies, i.e. the trees with CR1 subfamilies as nodes that minimize the sum of edge distances, were constructed using the Alucode modified specifically for CR1 (i.e. Length = 465). We tested multiple subfamilies as the consensus sequence including CR1-X1_Pass, J2_Pass and E_pass. Under the null hypothesis of uniformity, the P-value for the linkage was calculated using the nonparametric computation as described by Price et al [7]. Since this code can run on a wide range of resolutions, it can split a CR1 population into multiple subfamilies. Based on the size of our data (6,759 zebra finch CR1 elements or 24,198 elements extracted from both zebra finch and chicken genomes), we chose MINCOUNT = 60 or 150, respectively. We used CR1-X1_Pass as the consensus sequence with all other default parameters. Under this setting, MS trees had similar stable topologies and numbers of CR1 subfamilies as the conventional NJ method.

3. RESULTS

CR1 Repeat Identification and Sequence divergence distribution. We utilized RepeatMasker to identify CR1 elements on the zebra finch genome assembly. We then extracted all nearly-full-length CR1 elements whose insert length was ≥ 98% of the corresponding consensus sequence length (465 bp). Within the repeatmaskable genomic regions, compared to the chicken genome (119.0 repeats/Mb, 16.7 nearly-full-length repeats/Mb), the zebra finch genome shows a significant lower density of CR1 repeats (70.5 repeats/Mb, 6.0 nearly-full-length repeats/Mb). This is in contrast with the three times of enrichment of retrovirus-derived long terminal repeat (LTR) element copies in the zebra finch as compared to the chicken [1].

We performed a CR1 divergence distribution analysis of the zebra finch genome using the 21 previously known CR1 subfamilies. The divergence levels reported by RepeatMasker were corrected by the CpG content of each repeat and multiple hits. We plotted the divergence (i.e. substitution from consensus) distribution by summing all 21 subfamilies (data not shown). In the stacking plot, two peaks of bursts in CR1 amplification was detected (at 0.10 and 0.17) and estimated to occur approximately 28 and 48 mya assuming a substitution rate of 3.6×10^{-9} substitutions/site/year [4]. Notable differences among the distributions were observed when each CR1 subfamily was considered: 1) L1_Tgu, L2_Tgu, K4_Tgu, K3_Tgu, K1_Tgu, and K2_Tgu subfamilies show a dominant "young" divergence profile with a mode less than 0.09 substitutions/site (Table 1, "Y" type); 2) Other subfamilies show a dominant "ancient" divergence profile with a mode greater than 0.09 substitutions/site (Table 1, "A" type. Within them, X1_Pass, J2_Pass and E_pass subfamilies have more than 800 elements).

Characterization of zebra finch CR1 repeat elements and their relationships at a fine resolution

We categorized the zebra finch CR1 subfamilies using the custom program modified from Alucode [7]. Based on our analysis of 6,759 CR1 repeats from the zebra finch genome, we identified 34 distinct subfamilies: the subfamily composition

Table 1. Divergences of 21 previously described CR1 elements in the zebra finch genome

Subfamily	Average Divergence	Standard Deviation	Count	Type
CR1-L1_Tgu	4.26	3.07	63	Y
CR1-L2_Tgu	5.58	2.51	77	Y
CR1-K4_Tgu	6.63	1.91	223	Y
CR1-K3_Tgu	7.34	3.79	74	Y
CR1-K1_Tgu	7.43	4.67	13	Y
CR1-K2_Tgu	7.66	4.82	34	Y
CR1-X1_Pass	9.72	1.95	1044	A
CR1-YB1_Tgu	10.48	4.37	47	A
CR1-YB2_Pass	11.30	2.56	445	A
CR1-J3_Pass	11.37	2.17	184	A
CR1-Y_Pass	11.83	2.75	171	A
CR1-X2_Pass	12.21	2.35	63	A
CR1-J2_Pass	12.29	1.64	836	A
CR1-I_Tgu	12.30	2.89	67	A
CR1-J1_Pass	12.37	1.72	582	A
CR1-X3_Pass	12.46	1.62	109	A
CR1-Z1_Pass	14.10	1.94	84	A
CR1-Z2_Pass	14.34	1.90	38	A
CR1-Y1_Aves	16.94	3.19	246	A
CR1-E_Pass	17.20	2.45	2104	A
CR1-Y2_Aves	20.25	3.40	277	A

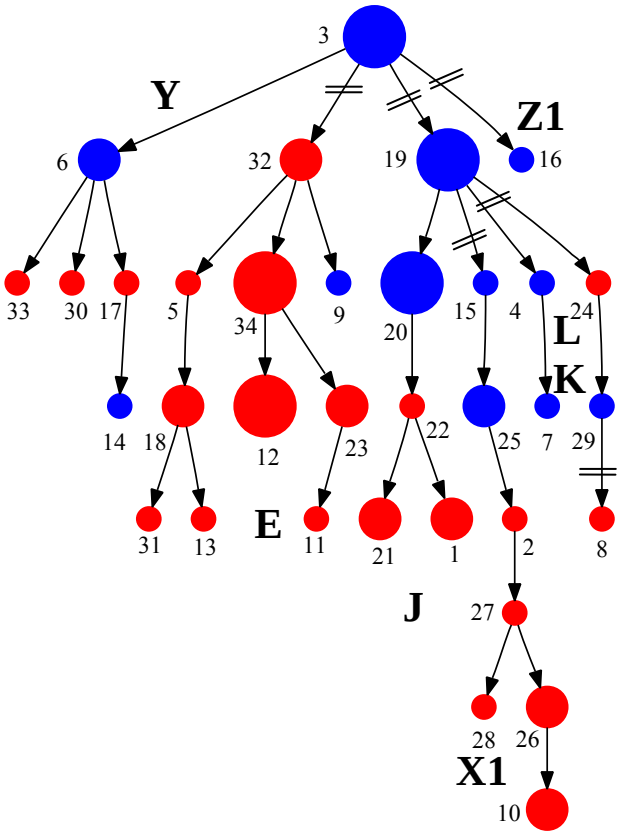

Figure 1. The minimum spanning tree of zebra finch CR1 subfamilies. This tree is based on an analysis of 6,759 zebra finch CR1 repeats. Previously known CR1 subfamilies are labeled in blue while new putative CR1 subfamilies are labeled in red. Large nodes: Subfamilies with more than 300 elements; medium nodes: 200 to 300 elements; small nodes: less than 200 elements. The subfamily number, name, type, count, P-value and sequence divergence within group are: 1. CR1-J2_Pass_2, new, 291, $5e^{-131}$, 0.26; 2. CR1-X1_Pass_2, new, 153, $9e^{-42}$, 0.198; 3. CR1-Y1_Aves, old, 503, $5e^{-131}$, 0.362; 4. CR1-K4_Tgu, old, 122, $5e^{-59}$, 0.126; 5. CR1-E_Pass_3, new, 179, $3e^{-66}$, 0.258; 6. CR1-YB2_Pass, old, 233, $1e^{-52}$, 0.235; 7. CR1-L2_Tgu, old, 136, $8e^{-77}$, 0.101; 8. CR1-J2_Pass_3, new, 102, $1e^{-39}$, 0.302; 9. CR1-E_Pass, old, 136, $6e^{-28}$, 0.251; 10. CR1-X1_Pass_3, new, 246, $7e^{-24}$, 0.198; 11. CR1-E_Pass_4, new, 122, $3e^{-35}$, 0.304; 12. CR1-E_Pass_5, new, 306, $3e^{-21}$, 0.298; 13. CR1-E_Pass_6, new, 131, $4e^{-32}$, 0.262; 14. CR1-Y_Pass, old, 115, $6e^{-56}$, 0.239; 15. CR1-X3_Pass, old, 111, $5e^{-60}$, 0.248; 16. CR1-Z1_Pass, old, 122, $4e^{-26}$, 0.322; 17. CR1-Y_Pass_2, new, 59, $2e^{-106}$, 0.242; 18. CR1-E_Pass_7, new, 245, $4e^{-48}$, 0.255; 19. CR1-J1_Pass, old, 518, $2e^{-60}$, 0.238; 20. CR1-J2_Pass, old, 375, $5e^{-139}$, 0.239; 21. CR1-J2_Pass_4, new, 236, $1e^{-377}$, 0.24; 22. CR1-J2_Pass_5, new, 155, $3e^{-344}$, 0.23; 23. CR1-E_Pass_8, new, 210, $7e^{-40}$, 0.283; 24. CR1-K4_Tgu_2, new, 140, $1e^{-90}$, 0.138; 25. CR1-X1_Pass, old, 229, $8e^{-49}$, 0.191; 26. CR1-X1_Pass_4, new, 202, $2e^{-59}$, 0.186; 27. CR1-X1_Pass_5, new, 191, $1e^{-152}$, 0.195; 28. CR1-X1_Pass_6, new, 89, $1e^{-147}$, 0.183; 29. CR1-K3_Tgu, old, 78, $2e^{-199}$, 0.135; 30. CR1-YB2_Pass_2, new, 174, $7e^{-99}$, 0.24; 31. CR1-E_Pass_9, new, 149, $1e^{-63}$, 0.241; 32. CR1-E_Pass_10, new, 252, $3e^{-53}$, 0.301; 33. CR1-YB2_Pass_3, new, 76, $3e^{-51}$, 0.228; and 34. CR1-E_Pass_2, new, 373, $1e^{-50}$, 0.302.

ranges from 59 to 518 with most subfamilies containing 200-300 elements (P-values for subfamily partition ranges from $1e^{-377}$ to $3e^{-21}$, see Price et al. [7] for the P-value definition and calculation). We next constructed a minimum spanning (MS) tree for these 34 CR1 subfamilies to summarize their evolutionary relationship (Figure 1, sequences available upon request from the authors). We identified approximately 22 new subfamilies (Figure 1, red dots) besides most of the previously known CR1 subfamilies (Figure 1, blue dots). Generally, we found a good agreement between the divergence distributions and this MS tree. Subfamily Y2_Aves is the most ancient one. Subfamilies YB2_Pass, Y_Pass, E_pass, J1_Pass, J2_Pass and Z1_Pass are derived from Y2_Aves. Subfamilies X1_Pass and X2_Pass are derived from J1_pass. Subfamilies Ls and Ks are the youngest subfamilies and they are directly derived from J1_Pass.

Characterization of lineage-specific CR1 repeat elements from turkey sequences

We used two distinct approaches to study lineage-specific CR1 subfamilies in the zebra finch-chicken comparison. First, we categorized CR1 subfamilies using the program Alucode [7]. Based on our analysis of 24,198 CR1 elements (6,759 from zebra finch and 17,439 from chicken), we identified 79 distinct subfamilies: the subfamily composition ranges from 102 to 1,065 with most subfamilies containing 500-800 elements (P-value for subfamily partition ranges from $3e^{-1068}$ to $2e^{-17}$). We next constructed a MS tree for these 79 CR1 subfamilies to summarize their evolutionary relationship (Figure 2). The topology of this tree is similar to the MS tree derived from the zebra finch only (Figure 1). We identified 1) 18 subfamilies shared between chicken and zebra finch ("cz"); 2) 29 subfamilies mainly in the chicken (≥95% by element count, "c"); 3) 10 subfamilies mainly in the zebra finch (≥95% by element count, "z"); 4) 18 subfamilies **only** in the chicken (labeled as "c*"); and 5) 4 subfamilies **only** in the zebra finch (labeled as "z*").

As a second method, we constructed a neighbor-joining (NJ) tree independently for 21 zebra finch CR1 consensus sequences (red symbols), 57 chicken consensus sequences (black symbols) as well as randomly selected 245 zebra finch CR1 repeats (Figure 3). The random samplings of zebra finch CR1 repeats were repeated multiple times and all replicates produced constant results. The zebra finch CR1 lineages include both ancestral and young elements: ancestral ones (Y2_Aves, E_Pass, X3_Pass, and etc) may be dead on arrival, while young ones (L1_Tgu, L2_Tgu, and K4_Tgu) may be still active more recently agreeing well with their divergences (Table 1) and MS tree results (Figures 1 and 2). This tree has several major branches: 1) on the left top are zebra finch ancestral subfamilies X1_Pass, X2_Pass and X3_Pass, which were old and not supported by bootstrapping. Among them, X1_Pass is interleaved together with chicken Y3, Y4 and Y4_2, which were supported by bootstrapping. 2) On the left bottom are chicken E subfamilies mixed with zebra finch E_Pass elements. These E_Pass subfamilies might represent degenerated copies of ancestral events. On their right are chicken subfamilies of Ds, Cs and recently derived Bs. 3) On the right bottom is the zebra finch YB2_Pass lineage, which is related to but distinct from chicken subfamilies Gs. Young chicken-specific subfamilies Fs and Y are derived from Gs. Ancient zebra finch Y1_Aves, Y2_Aves, Z1_Pass and Z2_pass are closely related to the chicken Gs. 4) On the right side are the zebra finch subfamilies Js and Is, which have a long branch length (ancient) and do not mix with any chicken consensus sequences. 5) On the right top are the most recent lineage-specific subfamilies: such as L1_Tgu, L2_Tgu and K4_Tgu elements in zebra finch and recent chicken-specific subfamilies X1s, X2s and Hs. Subfamilies K4_Tgu, and K4_Tgu_2 only contain zebra finch

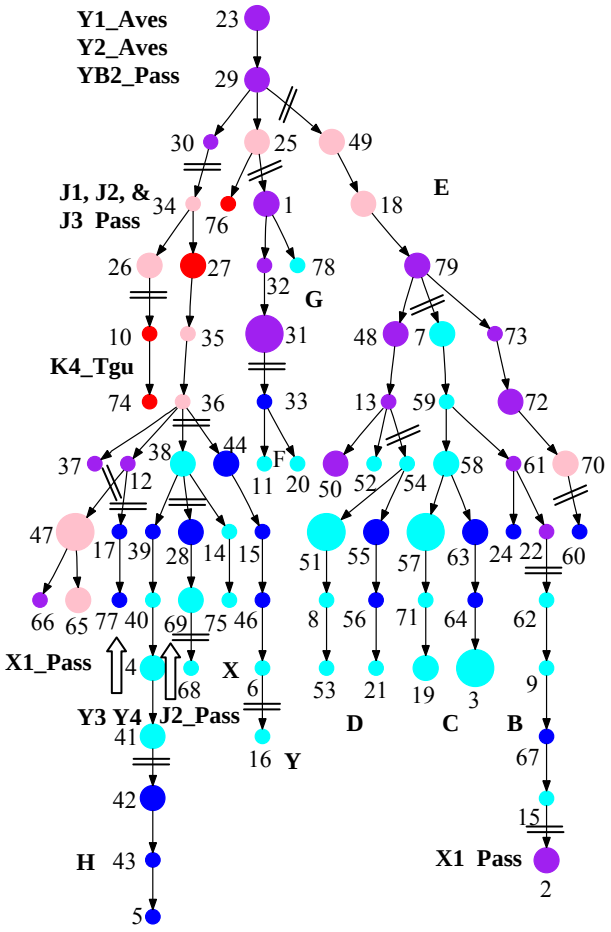

Figure 2. The minimum spanning tree of zebra finch and chicken CR1 subfamilies. This tree is based on an analysis of 6,759 zebra finch and 17,439 chicken CR1 repeats. Lineage-specific subfamilies are labeled red (zebra finch only) or blue (chicken only). Subfamilies with more than 95% elements coming from either bird are labeled pink (zebra finch) or cyan (chicken). Shared subfamilies between 2 bird species are labeled in purple. Large nodes: Subfamilies with more than 500 elements; medium nodes: 300 to 500 elements; small nodes: less than 300 elements. The subfamily number, name, type, count, P-value and sequence divergence within group, chicken element% and zebra finch element% are: 1. G_2, cz, 331, $1e^{-156}$, 0.233, 77.0%, 23.0%; 2. X1_Pass_2, cz, 400, $2e^{-208}$, 0.288, 37.5%, 62.5%; 3. C, c, 1065, $2e^{-216}$, 0.087, 99.6%, 0.4%; 4. X_2, c, 313, $4e^{-174}$, 0.162, 99.7%, 0.3%; 5. H_2, c*, 295, $7e^{-324}$, 0.038, 100.0%, 0.0%; 6. X2_2, c, 227, $1e^{-238}$, 0.091, 99.1%, 0.9%; 7. C4_2, c, 366, $8e^{-211}$, 0.249, 97.8%, 2.2%; 8. D2_2, c, 296, $2e^{-192}$, 0.207, 96.6%, 3.4%; 9. B2, c, 290, $2e^{-92}$, 0.098, 99.3%, 0.7%; 10. K4_Tgu_2, z*, 272, $9e^{-73}$, 0.13, 0.0%, 100.0%; 11. F2, c, 320, $9e^{-200}$, 0.176, 99.7%, 0.3%; 12. J2_Pass_2, cz, 239, $2e^{-17}$, 0.278, 63.2%, 36.8%; 13. E_3, cz, 282, $3e^{-301}$, 0.187, 83.7%, 16.3%; 14. X1, c, 276, $3e^{-197}$, 0.094, 97.8%, 2.2%; 15. B, c, 204, $3e^{-93}$, 0.049, 98.5%, 1.5%; 16. Y, c, 277, $8e^{-201}$, 0.055, 99.6%, 0.4%; 17. Y4, c*, 219, $6e^{-99}$, 0.26, 100.0%, 0.0%; 18. E_4, z, 431, $1e^{-126}$, 0.276, 0.2%, 99.8%; 19. C4_3, c, 304, $7e^{-70}$, 0.256, 98.0%, 2.0%; 20. F, c, 258, $3e^{-130}$, 0.149, 98.4%, 1.6%; 21. D_2, c, 254, $2e^{-135}$, 0.161, 99.2%, 0.8%; 22. C3_2, cz, 195, $6e^{-68}$, 0.171, 92.3%, 7.7%; 23. Y2_Aves, cz, 380, $2e^{-192}$, 0.379, 38.2%, 61.8%; 24. C3_3, c*, 286, $2e^{-30}$, 0.168, 100.0%, 0.0%; 25. YB2_Pass, z, 397, $7e^{-110}$, 0.249, 3.8%, 96.2%; 26. J3_Pass, z, 327, $1e^{-179}$, 0.241, 2.1%, 97.9%; 27. J2_Pass_3, z*, 399, $1e^{-29}$, 0.249, 0.0%, 100.0%; 28. J2_Pass_4, c*, 343, $1e^{-29}$, 0.203, 100.0%, 0.0%; 29. Y1_Aves, cz, 448, $2e^{-172}$, 0.312, 53.8%, 46.2%; 30. Y1_Aves_2, cz, 257, $2e^{-670}$, 0.328, 70.0%, 30.0%; 31. G_3, cz, 508, $4e^{-106}$, 0.235, 94.3%, 5.7%; 32. G, cz, 252, $5e^{-267}$, 0.221, 93.3%, 6.7%; 33. F2_2, c*, 272, $2e^{-309}$, 0.168, 100.0%, 0.0%; 34. J1_Pass, z, 251, $5e^{-63}$, 0.26, 2.8%, 97.2%; 35. J2_Pass, z, 226, $2e^{-125}$, 0.23, 4.4%, 95.6%; 36. J2_Pass_5, z, 279, $4e^{-240}$, 0.248, 2.9%, 97.1%; 37. J2_Pass_6, cz, 165, $2e^{-722}$, 0.258, 9.7%, 90.3%; 38. X_3, c, 477, $3e^{-78}$, 0.186, 99.4%, 0.6%; 39. X_4, c*, 218, $1e^{-47}$, 0.172, 100.0%, 0.0%; 40. X_5, c, 243, $5e^{-72}$, 0.18, 96.7%, 3.3%; 41. X, c, 354, $2e^{-1063}$, 0.071, 99.4%, 0.6%; 42. H_3, c*, 493, $9e^{-820}$, 0.045, 100.0%, 0.0%; 43. H, c*, 256, $4e^{-196}$, 0.046, 100.0%, 0.0%; 44. X2_3, c*, 477, $3e^{-1068}$, 0.097, 100.0%, 0.0%; 45. X2_4, c*, 264, $6e^{-545}$, 0.083, 100.0%, 0.0%; 46. X2, c*, 102, $1e^{-576}$, 0.084, 100.0%, 0.0%; 47. X1_Pass_3, z, 508, $8e^{-136}$, 0.193, 3.3%, 96.7%; 48. E_5, cz, 367, $3e^{-69}$, 0.185, 89.6%, 10.4%; 49. E_6, z, 325, $5e^{-69}$, 0.252, 0.3%, 99.7%; 50. E, cz, 360, $6e^{-71}$, 0.171, 91.9%, 8.1%; 51. D2, c, 550, $2e^{-125}$, 0.194, 97.6%, 2.4%; 52. E_7, c, 153, $2e^{-68}$, 0.155, 96.7%, 3.3%; 53. D2_3, c, 244, $1e^{-64}$, 0.21, 96.3%, 3.7%; 54. D2_4, c, 236, $2e^{-218}$, 0.205, 98.3%, 1.7%; 55. D, c*, 336, $5e^{-112}$, 0.164, 100.0%, 0.0%; 56. D_3, c*, 164, $2e^{-407}$, 0.165, 100.0%, 0.0%; 57. C4_4, c, 566, $2e^{-115}$, 0.218, 99.5%, 0.5%; 58. C4, c, 336, $8e^{-88}$, 0.228, 99.1%, 0.9%; 59. C4_5, c, 194, $1e^{-177}$, 0.242, 95.9%, 4.1%; 60. C3_4, c*, 161, $4e^{-151}$, 0.204, 100.0%, 0.0%; 61. C3, cz, 156, $2e^{-91}$, 0.208, 94.9%, 5.1%; 62. B2_2, c, 248, $4e^{-244}$, 0.111, 99.2%, 0.8%; 63. C2, c*, 445, $1e^{-366}$, 0.108, 100.0%, 0.0%; 64. C_2, c*, 248, $1e^{-70}$, 0.098, 100.0%, 0.0%; 65. X1_Pass, z, 301, $3e^{-1041}$, 0.196, 1.3%, 98.7%; 66. X1_Pass_4, cz, 224, $1e^{-290}$, 0.247, 17.4%, 82.6%; 67. B_2, c*, 150, $7e^{-313}$, 0.052, 100.0%, 0.0%; 68. X_6, c, 237, $3e^{-226}$, 0.223, 99.2%, 0.8%; 69. J2_Pass_7, c, 337, $1e^{-152}$, 0.199, 99.4%, 0.6%; 70. E_8, z, 370, $5e^{-149}$, 0.296, 2.4%, 97.6%; 71. C4_6, c, 216, $5e^{-69}$, 0.243, 99.1%, 0.9%; 72. E_9, cz, 396, $1e^{-117}$, 0.305, 5.3%, 94.7%; 73. E_10, cz, 219, $9e^{-49}$, 0.308, 14.2%, 85.8%; 74. K4_Tgu, z*, 193, $5e^{-115}$, 0.148, 0.0%, 100.0%; 75. X1_2, c, 190, $1e^{-345}$, 0.088, 99.5%, 0.5%; 76. YB2_Pass_2, z*, 161, $2e^{-135}$, 0.232, 0.0%, 100.0%; 77. Y3, c*, 186, $6e^{-124}$, 0.224, 100.0%, 0.0%; 78. G_4, c, 151, $5e^{-44}$, 0.223, 98.0%, 2.0%; 79. E_2, cz, 482, $3e^{-72}$, 0.307, 44.8%, 55.2%.

elements and do not mix with any chicken subfamilies. They have short length (young), multiple branches (active), suggesting these younger lineage-specific CR1 elements may still be active in the zebra finch.

Subfamily consensus sequences and phylogeny

We also performed a phylogenetic analysis (NJ tree) on 34 zebra finch (with postfix of "ms") identified by the current study and 21 previously known zebra finch CR1 consensus sequences. In the NJ tree shown in Figure 4, the relationship among known and new CR1 consensus sequences was recovered as expected. Out of 21 known subfamilies, 12 were confirmed and covered by new consensus sequences (labeled as black brackets). The sequence distances between known consensus sequences and their closest neighbors range from 0.011 to 0.052, with an average of 0.028 and standard deviation of 0.014. The few discrepancies between our consensus sequences and the consensus sequences reported in Repbase occur mostly at CpG dinucleotide positions, which are ill-determined because of frequent mutation. In spite of the above-mentioned ancestry sharing, 22 new consensus sequences were discovered (Figure 4, labeled by red brackets). The new subfamilies include CR1-J2_Pass (4), CR1-X1_Pass (5), CR1-E_Pass (9), CR1-YB2_Pass (2), CR1-Y_Pass (1), and CR1-K4_Tgu (1). Overwhelming majority of newly discovered

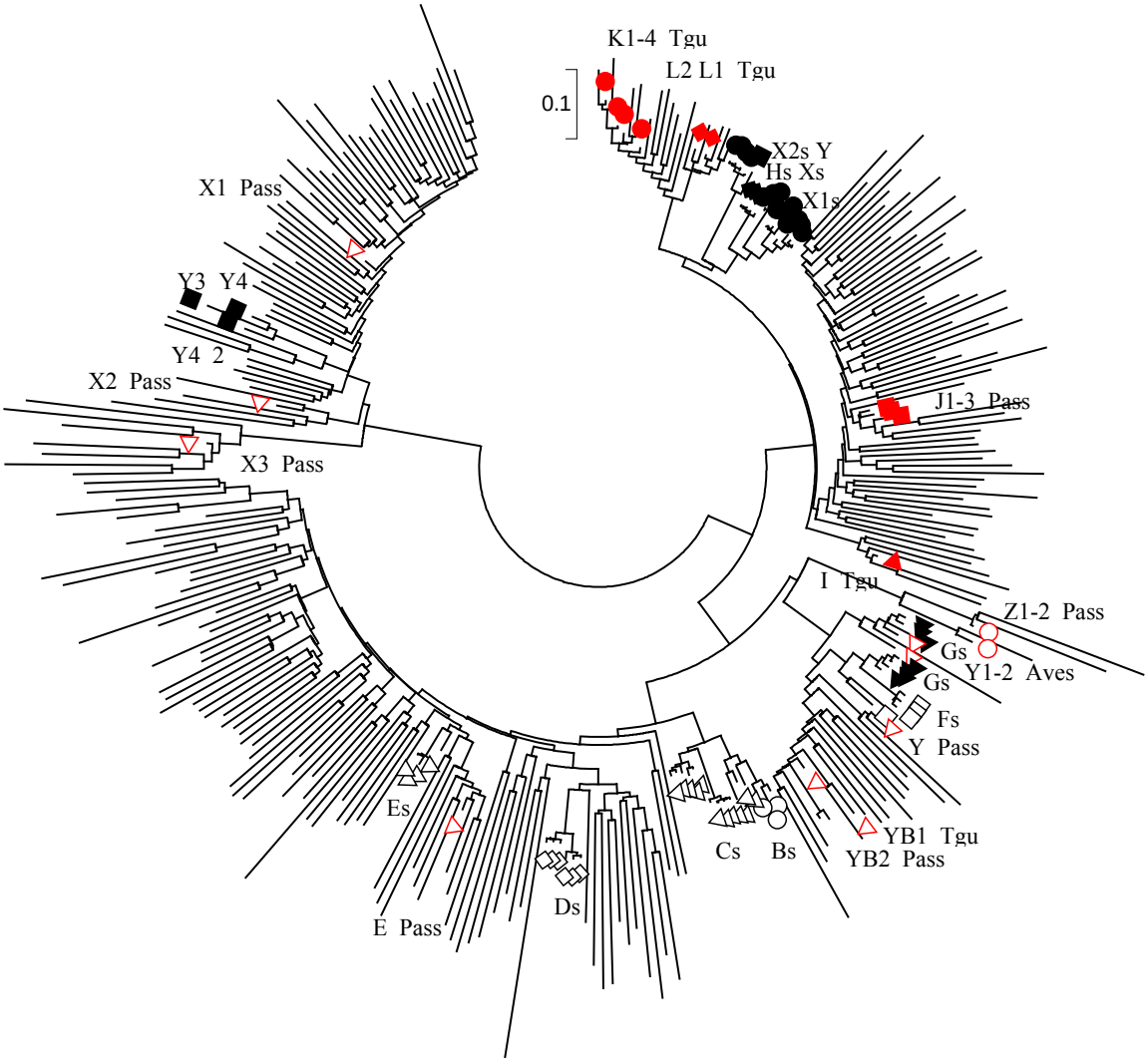

Figure 3. The neighbor-joining tree of zebra finch-chicken CR1 comparison. This neighbor-joining tree includes 21 previously known zebra finch CR1 consensus sequences (red symbols), 57 chicken consensus sequences (black symbols) as well as randomly selected 245 zebra finch CR1 repeats (lineages without dots). The major branches are labeled with subfamily names. See main texts for details.

consensus sequences come from those subfamilies with many elements, including subfamilies X1_Pass, J2_Pass and E_Pass. Those missed subfamilies (not bracketed) are likely due to the high threshold (MINCOUNT=60) of Alucode.

4. DISCUSSION

In this project, we performed a global characterization of CR1 elements in the zebra finch genome using an integrated approach combining two distinct phylogenetic methods: NJ and MS trees. We identified 34 zebra finch CR1 consensus sequences. Our analysis supports a model in which a burst of CR1 activities occurred between 28-48 mya, with multiple master CR1 genes involved in the zebra finch lineages. These observations generally support that CR1 subfamilies originated through the serial fixation of multiple master CR1 elements. We further identified 4 zebra finch specific CR1 subfamilies.

Our results have confirmed previous analysis [4] as well as provided new insights with respect to evolutionary relationships of the zebra finch CR1 subfamilies. The earlier results based on insertion order/rank analysis in chickens suggested that 1) X, X1, Y4, and C4 are the most ancient CR1 subfamilies, with C4 being the most common; 2) C, C3, D, D2, E, G, H, X2, Y and Y3 represent the major burst of CR1 elements and 3) B, B2, C, C2, F, F0, F2, H2 and Y2 are among the youngest subfamilies. On the other hand, our earlier data indicated that a subset of CR1-G belongs to the most ancient group and parts of CR1-H, X, X1, and X2 belong to the youngest group [6]. The current study further confirmed our earlier data: CR1_Y2_Aves are the most ancient repeats shared by the zebra finch and chicken. Chicken CR1-Gs derived from Y2_Aves via YB2_Pass. Younger lineage-specific CR1 elements like K4_Tgu may still be active in zebra finch.

As discussed before [6], one source of these discrepancies may be that we limited our analyses to the 465 bp of the 3' terminus (155 amino acids) of ORF2. Other studies are based on longer 3' terminus (~1,000 bp) of or full length ORF2 [4]. Since the vast majority of CR1s are fragments shorter than 1,000 bp, filtering of RepeatMasker output with a shorter length requirement will preserve more CR1 copies, thus making our samples more representative. Another difference is that two

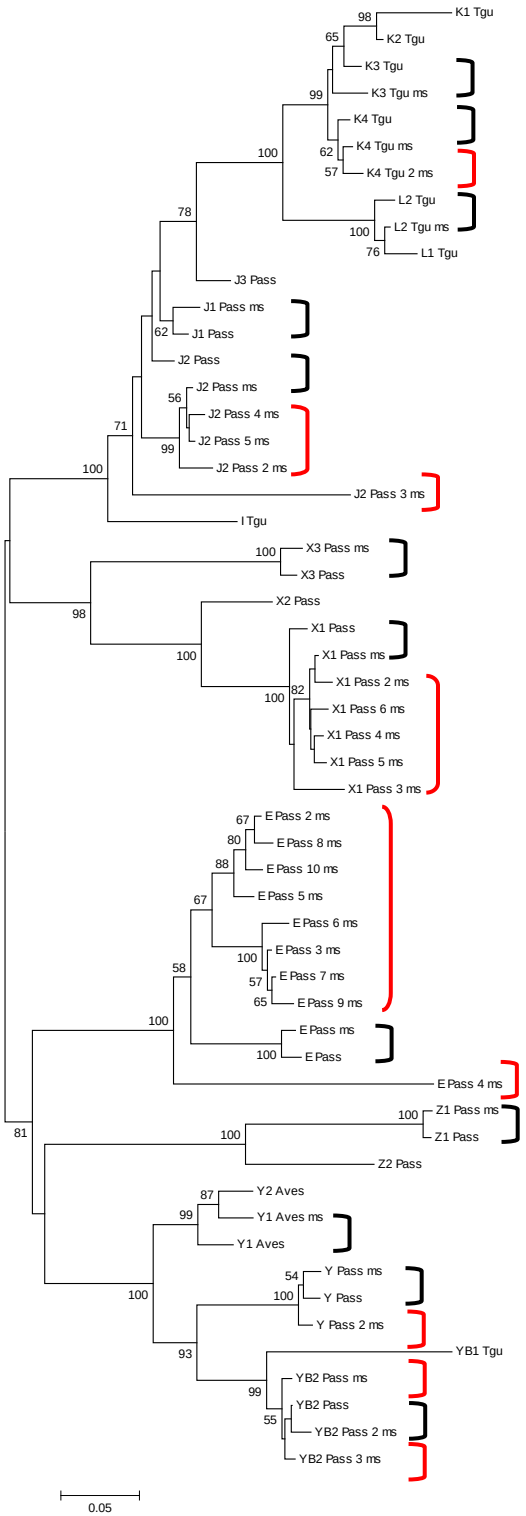

Figure 4. The neighbor-joining tree of previously known and newly discovered zebra finch CR1 consensus sequences. This neighbor-joining tree includes 34 zebra finch (with postfix of "ms") identified by the current study and 21 previously known zebra finch CR1 consensus sequences. The confirmations of previously known consensus sequences by the new zebra finch CR1 subfamilies are labeled by black brackets. The newly derived subfamilies are labeled by red brackets. All branches are labeled with the bootstrap values (>50%) with n=1,000 replicates.

distinct methods were used. The insertion order/rank method does not directly depend on sequence divergences but instead depends on the RepeatMasker program to properly assign repeat subfamily [5]. The accuracy of that method also depends on the repeat length and their connectedness with other repeats. The proper subfamily assignment of repeats by RepeatMasker depends on the fact that the consensus sequences are properly constructed and thoroughly verified. Therefore, our results of 34 zebra finch CR1 subfamilies offer a new refined prospective for CR1 classification and evolution.

In summary, our analysis has provided an evolutionary framework for further classification and refinement of the CR1 repeat phylogeny. These new CR1 subfamilies expand our understanding of CR1 evolution and their impacts on bird genome architecture. The differences in the distribution and rates of CR1 activity may play an important role in subtly reshaping the structure of the zebra finch genome. The structural and functional consequences of these changes among the bird genomes are an important area for future investigation.

5. REFERENCES

[1] Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A *et al.*: **The genome of a songbird.** *Nature* 2010, **464:** 757-762.

[2] International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432:** 695-716.

[3] Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA *et al.*: **The repetitive landscape of the chicken genome.** *Genome Res* 2005, **15:** 126-136.

[4] Abrusan G, Krambeck HJ, Junier T, Giordano J, Warburton PE: **Biased distributions and decay of long interspersed nuclear elements in the chicken genome.** *Genetics* 2008, **178:** 573-581.

[5] Giordano J, Ge Y, Gelfand Y, Abrusan G, Benson G, Warburton PE: **Evolutionary history of mammalian transposons determined by genome-wide defragmentation.** *PLoS Comput Biol* 2007, **3:** e137.

[6] Liu GE, Jiang L, Tian F, Zhu B, Song J: **Calibration of mutation rates reveals diverse subfamily structure of galliform CR1 repeats.** *Genome Biol Evol* 2009, **1:** 119-130.

[7] Price AL, Eskin E, Pevzner PA: **Whole-genome analysis of Alu repeat elements reveals complex evolutionary history.** *Genome Res* 2004, **14:** 2245-2252.

[8] Liu GE, Alkan C, Jiang L, Zhao S, Eichler EE: **Comparative analysis of Alu repeats in primate genomes.** *Genome Res* 2009, **19:** 876-885.

[9] Smit AF: **Interspersed repeats and other mementos of transposable elements in mammalian genomes.** *Current Opinion in Genetics & Development* 1999, **9:** 657-63.

[10] Kumar S, Tamura K, Jakobsen IB, Nei M: **MEGA2: molecular evolutionary genetics analysis software.** *Bioinformatics (Oxford)* 2001, **17:** 1244-5.