Biology Direct

BioMed Central

Discovery notes

Open Access

Endogenous retroviruses of the chicken genome

Ahsan Huda, Nalini Polavarapu, I King Jordan* and John F McDonald

Address: School of Biology, Georgia Institute of Technology, Atlanta, GA, USA

Email: Ahsan Huda - ahsan.huda@gatech.edu; Nalini Polavarapu - nalini@gatech.edu; I King Jordan* - king.jordan@biology.gatech.edu; John F McDonald - john.mcdonald@biology.gatech.edu

* Corresponding author

Published: 24 March 2008

Biology Direct 2008, **3**:9 doi:10.1186/1745-6150-3-9

Received: 6 March 2008
Accepted: 24 March 2008

This article is available from: http://www.biology-direct.com/content/3/1/9

Abstract

We analyzed the chicken (*Gallus gallus*) genome sequence to search for previously uncharacterized endogenous retrovirus (ERV) sequences using *ab initio* and combined evidence approaches. We discovered 11 novel families of ERVs that occupy more than 21 million base pairs, approximately 2%, of the chicken genome. These novel families include a number of recently active full-length elements possessing identical long terminal repeats (LTRs) as well as intact *gag* and *pol* open reading frames. The abundance and diversity of chicken ERVs we discovered underscore the utility of an approach that combines multiple methods for the identification of interspersed repeats in vertebrate genomes.

Reviewers: This article was reviewed by Igor Zhulin and Itai Yanai.

Findings

Chicken, a modern descendant of the dinosaurs, is the first avian to have its genome sequenced [1]. Phylogenetically, its position between fish and mammals provides valuable insight into the evolution of vertebrates. The chicken genome has a size of 1.2 billion bases, approximately one third of the size of the human genome.

The overall interspersed repeat, *i.e.* transposable element (TE), content of the chicken genome was determined to be less than 9% [1]. This fraction is considerably lower than that of mammalian genomes, where transposable elements (TEs) account for 40–50% of genomic sequences [2-4]. While chicken has long been a model system for the study of retroviruses [5], a mere 1.3% of the chicken genome can be classified as endogenous retroviruses (ERVs) compared to about 5% in humans [3]. Nevertheless, protein coding sequences still make up only a minor fraction of the chicken genome leaving a substantial quotient that has yet to be been accounted for. The authors of the initial analysis of the chicken genome posited that much of the uncharacterized sequence was likely to be derived from unrecognized TEs [1]. Indeed, novel or previously uncharacterized TE sequences may be missed by homology-based methods for the detection of repeats, such as the widely used RepeatMasker program [6], which rely on the comparison of genomic sequences to libraries of known repeat consensus sequences. *Ab initio* methods, on the other hand, identify repeats by virtue of their structural characteristics without regard to any sequence similarity to known elements. We used a combination of *ab initio* detection, sequence similarity searches, motif identification and evaluation of element structural (repeat) features to search for novel ERVs that may have been missed in the initial analysis of the chicken genome.

LTR_STRUC was the first *ab initio* program designed to detect long terminal repeat (LTR) containing elements, such as ERVs, in genomic sequence [7]. Briefly, LTR_STRUC works by sliding a window along genomic

sequence and looking for direct repeats that are spaced apart within a specified size range (*e.g.* 5–10 kb). After identifying putative LTRs in this way, it searches for other characteristic features of LTR elements such as target site duplications, inverted repeats at LTR termini, primer binding sites and poly purine tracts. Based on these features, it predicts the direction of the LTR element and provides the corresponding three frame translation of the reverse transcriptase (RT) sequence in the internal region of the element. LTR_STRUC has proven effective at identifying novel LTR elements, including ERVs, in chimpanzee [8], mouse [9] and rice genome sequences [10].

LTR_STRUC was run on the 2004 build of the chicken genome sequence, *i.e.* the v1.0 draft assembly from the Washington University Genome Sequencing Center [1] distributed on the UCSC Genome Browser [11], resulting in the detection of 39 putative full-length LTR elements. RT homologous sequences were identified in these elements and used as queries in TBLASTN [12] searches against the chicken genome sequence. The BLAST output and flanking genomic sequences were visually inspected to look for ERV characteristic features such as LTRs, target site repeats and terminal inverted repeats. LTRs are direct repeats at the 5' and 3' termini of the ERVs that are ~200–350 bp in length. Characteristic dinucleotide terminal inverted repeats are found at the beginning (TG) and ends (CA) of ERV LTRs. Target site repeats are short (4–6 bp) direct repeats found immediately upstream and downstream of ERV insertions that result from resolution of a staggered break that is made when the elements integrate in the genome. We identified a total of 89 putative ERVs in the genome using the combined *ab initio*, sequence similarity and element feature detection approach. The presence of intact open reading frames that encode sequences that have significant sequence similarity to RT along with the canonical RT catalytic motif [13] were used to validate 61 of these cases as intact full-length ERVs.

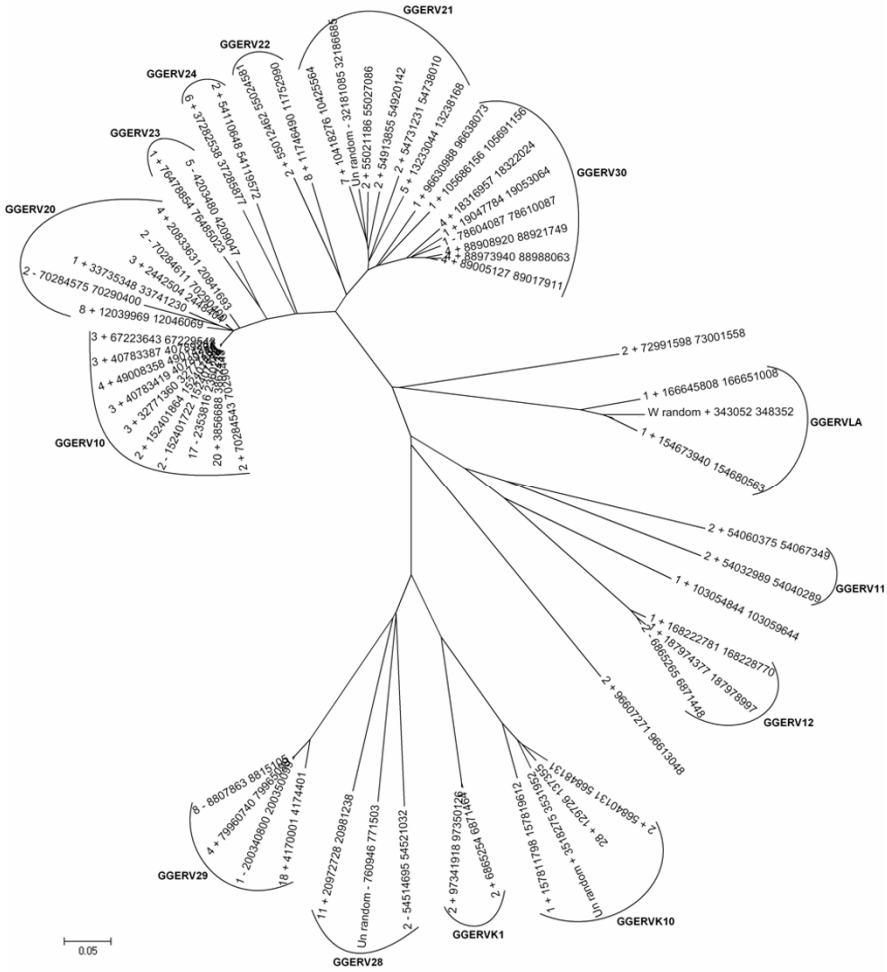
Phylogenetic analysis of an RT nucleotide sequence alignment was used to classify the chicken ERV sequences that we identified. ERV phylogenies were built using the neighbor-joining and maximum parsimony methods implemented in the program MEGA [14] and maximum likelihood using the program PhyML [15]. For neighbor-joining and maximum parsimony 1,000 bootstrap replicates were run to assess the support for internal branches on the phylogeny, and the approximate likelihood ratio test [16] was used to evaluate the support for branches along the maximum likelihood tree. The ERV phylogeny shows a number of well resolved groups that correspond for 14 distinct families of chicken ERVs, 11 of which are described here for the first time (Figure 1). In the absence of a standard naming convention for viral families in the chicken genome, we named the families using GGERVNN, for *Gallus gallus* endogenous retrovirus followed by the family number. We also reported the new families to Repbase [17] where they constitute nearly half (8 out of 17) of all the ERV families known for the chicken genome.

The 11 new ERV families we discovered using LTR_STRUC and BLAST analysis include 48 full-length elements and 1,542 fragmented sequences, most of which are solo LTRs that result from intra-element LTR-LTR recombination. When representatives of the 11 novel families were used to search the chicken genome for homologous sequences using the RepeatMasker program [6], they hit ~21 megabases of ERV sequence, or 2.0% of the genome. Together, the previously characterized and newly characterized ERVs represent more than 30 megabases of sequence and 2.9% of the chicken genome, a substantial increase over the previous figure of 1.3% of ERV sequences.

GGERV21, GGERV22 and GGERV30 are the most abundant lineages and account for more than half of all the viral sequences in the genome. However, only a few full-length elements were found for these abundant families; most of their sequences exist as fragments or solo LTRs. These abundant families are most closely related to the Birdawg and Kronos LTR elements previously identified as high copy number elements using cot-based sequencing and analysis of the chicken genome [18]. However, we did not identify any full-length elements corresponding to the Hithcock or Soprano LTR elements identified in the same study.

The LTRs at the 5' and 3' ends of a full-length ERV genomic sequence are generated from a single template during reverse transcription of RNA into DNA [19]. Therefore, at the moment that a full-length ERV integrates into the genome, its 5' and 3' LTRs are expected to be identical in sequence, and intra-element sequence differences between LTRs can be used to estimate the time that has elapsed since an element was active [20]. The ages of chicken ERVs were estimated in this way using the formula $t = d/2r$, where t is the time since insertion, d is the nucleotide sequence divergence per site between 5' and 3' LTRs of a single element and r is the rate of nucleotide substitution per site per million years. The value of r used here, 7.5×10^{-4}, is based on comparisons of nuclear genes among four avian taxa [21].

Age ranges for the 11 novel ERV families we detected are shown in Figure 1. The youngest family of chicken ERVs is GGERV10, which includes 10 full-length elements with 5' and 3' LTRs that are either identical or differ by only 1 bp. The GGERV10 family of element sequences integrated from 0–3 million years ago. Full-length GGERV10 family



Family	LTR length	Inserted element length	Copy number	5' to 3' LTR Identity	Age (Million Years)		Support		
					Min	Max	Neighbour-Joining	Parsimony	Likelihood
GGERV10	220	5386	97	100	0	3	72	63	92
GGERV29	336	5568	36	97	0	17.9	99	58	80
GGERV12	240	7073	50	97	2.8	32.9	99	67	98
GGERV20	220	5150	79	96	0	59.1	83	63	95
GGERV28	351	5736	113	90	16.7	66.5	54	98	52
GGERV30	310	5308	152	92	24.6	77.9	56	68	79
GGERV21	322	7665	149	86	32.4	130.9	54	71	43
GGERV11	232	5328	86	90	63.2	102.9	99	96	98
GGERV23	244	4226	195	89	66.9	71	55	82	33
GGERV22	344	5714	626	84	102.7	131	97	83	84
GGERV24	318	7210	7	78	136.5	143.9	55	67	56

Figure 1

Chicken endogenous retrovirus families. Phylogenetic analysis of an RT multiple sequence alignment for all full-length elements was used to delineate chicken ERV families. The neighbor-joining phylogeny is shown; maximum parsimony and maximum likelihood trees were also reconstructed. The names of the taxa (ERV sequences) correspond to the chicken chromosome number, strand, start and end coordinates from the May 2006 build, v2.1 draft assembly from the Washington University Genome Sequencing Center, found on the UCSC Genome Browser. Family names and characteristics for the 11 novel ERV families discovered here are shown below the tree. Family copy numbers are indicated along with the family averages of intra-element percent identity between 5' and 3' LTRs and their age ranges (lower-to-upper bounds). For each family, percent support values are shown for the internal branch that subtends the family based on bootstrap analysis, for neighbor-joining and maximum parsimony, and the approximate likelihood ratio test for maximum likelihood.

members encode a ~1,600 base pair intact *gag* open reading frame (ORF) and a ~3,300 base pair *pol* ORF that encodes a polyprotein with homology to the protease, RT, RNAseH and integrase enzymes that catalyze reverse transcription. In other words, GGERV10 family members are potentially active ERVs that were integrated into the chicken genome very recently. Incidentally, the GGERV10 family is substantially younger than the GGERVLA (Figure 1) family that was previously described as the most recently active family in the genome [1].

The next youngest family is GGERV29, with elements that inserted 0–17.9 million years ago, and the oldest family we identified is GGERV24 at 136.5–143.9 million years old. This wide range of ages encompasses all newly discovered and previously characterized chicken ERVs. Even though the *ab initio* approach we used is best suited to find relatively young elements with readily identifiable structural elements (*i.e.* LTRs), it was able to detect families that were active hundreds-of-millions of years apart.

Using a combined evidence approach that integrates *ab initio* element detection with sequence similarity searches, motif identification and evaluation of element features we detected 11 novel ERV families covering more than 21 megabases of previously uncharacterized chicken genome sequence. Several of these families were fairly ancient, consistent with the expectation that degenerated element sequences may be missed by homology-based detection methods. However, a number of the ERVs we identified are members of young families that have been active very recently in the chicken genome. These results underscore the importance of integrating multiple methods [22] for the detection of interspersed repeats in eukaryotic genomes.

Reviewers' comments

Reviewer's report 1

Igor Zhulin, University of Tennessee and Oak Ridge National Laboratory

This is an interesting discovery of novel viral families in the chicken genome, which accounts for more than 2% of the genome sequence. I do not have any major concerns regarding this paper and support its publication; however, I would like to offer some comments for authors' consideration, mainly regarding the clarity and presentation.

Authors' response

We are grateful to the reviewer Dr. Igor Zhulin for providing a number of very specific and constructive comments regarding the clarity and presentation of the manuscript. We revised the paper according to his suggestions.

Reviewer's report 2

Itai Yanai, Harvard University

I support publication of this manuscript.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

AH and NP implemented LTR_STRUC on the chicken genome. AH performed all other sequence analyses and the phylogenetic analysis under the supervision of JFM and IKJ. AH and IKJ drafted the manuscript. All authors reviewed and approved the final version of the manuscript.

Acknowledgements

AH, NP and IKJ are supported by the School of Biology, Georgia Institute of Technology. JFM and NP were supported by a grant from the Georgia Tech Research Foundation. We would like to thank members of the McDonald lab and Jordan lab for their support and technical assistance.

References

1. Consortium ICGS: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432:**695-716.
2. Consortium CSaA: **Initial sequence of the chimpanzee genome and comparison with the human genome.** *Nature* 2005, **437:**69-87.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
4. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al.*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
5. Rous P: **A transmissible avian neoplasm.** *J Exp Med* 1910, **12:**696-705.
6. **RepeatMasker Open-3.0** [http://www.repeatmasker.org]
7. McCarthy EM, McDonald JF: **LTR_STRUC: a novel search and identification program for LTR retrotransposons.** *Bioinformatics* 2003, **19:**362-367.
8. Polavarapu N, Bowen NJ, McDonald JF: **Identification, characterization and comparative genomics of chimpanzee endogenous retroviruses.** *Genome Biol* 2006, **7:**R51.
9. McCarthy EM, McDonald JF: **Long terminal repeat retrotransposons of Mus musculus.** *Genome Biol* 2004, **5:**R14.
10. McCarthy EM, Liu J, Lizhi G, McDonald JF: **Long terminal repeat retrotransposons of Oryza sativa.** *Genome Biol* 2002, **3:**RESEARCH0053.
11. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
12. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
13. Xiong Y, Eickbush TH: **Origin and evolution of retroelements based upon their reverse transcriptase sequences.** *Embo J* 1990, **9:**3353-3362.
14. Kumar S, Tamura K, Nei M: **MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment.** *Brief Bioinform* 2004, **5:**150-163.
15. Guindon S, Lethiec F, Duroux P, Gascuel O: **PHYML Online – a web server for fast maximum likelihood-based phylogenetic inference.** *Nucleic Acids Res* 2005, **33:**W557-559.

16. Anisimova M, Gascuel O: **Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative.** *Syst Biol* 2006, **55:**539-552.
17. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.
18. Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH, Ivarie R: **The repetitive landscape of the chicken genome.** *Genome Res* 2005, **15:**126-136.
19. Arkhipova IR, Mazo AM, Cherkasova VA, Gorelova TV, Schuppe NG, Llyin YV: **The steps of reverse transcription of Drosophila mobile dispersed genetic elements and U3-R-U5 structure of their LTRs.** *Cell* 1986, **44:**555-563.
20. SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL: **Nested retrotransposons in the intergenic regions of the maize genome.** *Science* 1996, **274:**765-768.
21. Hedges SB, Parker PH, Sibley CG, Kumar S: **Continental breakup and the ordinal diversification of birds and mammals.** *Nature* 1996, **381:**226-229.
22. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D: **Combined evidence annotation of transposable elements in genome sequences.** *PLoS Comput Biol* 2005, **1:**166-175.