BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: implications for molecular inference of population history and phylogenetics

Sérgio L Pereira*[1] and Allan J Baker[2]

Address: [1]Centre for Biodiversity and Conservation Biology – Royal Ontario Museum, 100 Queen's Park, Toronto, ON, M5S 2C6 Canada and [2]Department of Zoology, University of Toronto, Toronto ON, M5S 1A1, Canada

Email: Sérgio L Pereira* - sergio.pereira@utoronto.ca; Allan J Baker - allanb@rom.on.ca

* Corresponding author

Published: 25 June 2004

BMC Evolutionary Biology 2004, **4**:17 doi:10.1186/1471-2148-4-17

Received: 21 April 2004

Accepted: 25 June 2004

This article is available from: http://www.biomedcentral.com/1471-2148/4/17

Abstract

Background: Mitochondrial DNA has been detected in the nuclear genome of eukaryotes as pseudogenes, or *Numts*. Human and plant genomes harbor a large number of *Numts*, some of which have high similarity to mitochondrial fragments and thus may have been inadvertently included in population genetic and phylogenetic studies using mitochondrial DNA. Birds have smaller genomes relative to mammals, and the genome-wide frequency and distribution of *Numts* is still unknown. The release of a preliminary version of the chicken (*Gallus gallus*) genome by the Genome Sequencing Center at Washington University, St. Louis provided an opportunity to search this first avian genome for the frequency and characteristics of *Numts* relative to those in human and plants.

Results: We detected at least 13 *Numts* in the chicken nuclear genome. Identities between *Numts* and mitochondrial sequences varied from 58.6 to 88.8%. Fragments ranged from 131 to 1,733 nucleotides, collectively representing only 0.00078% of the nuclear genome. Because fewer *Numts* were detected in the chicken nuclear genome, they do not represent all regions of the mitochondrial genome and are not widespread in all chromosomes. Nuclear integrations in chicken seem to occur by a DNA intermediate and in regions of low gene density, especially in macrochromosomes.

Conclusion: The number of *Numts* in chicken is low compared to those in human and plant genomes, and is within the range found for most sequenced eukaryotic genomes. For chicken, PCR amplifications of fragments of about 1.5 kilobases are highly likely to represent true mitochondrial amplification. Sequencing of these fragments should expose the presence of unusual features typical of pseudogenes, unless the nuclear integration is very recent and has not yet been mutated. Metabolic selection for compact genomes with reduced repetitive DNA and gene-poor regions where *Numts* occur may explain their low incidence in birds.

Background

The establishment of the mitochondrion as a cellular organelle by endosymbiosis [1] changed the fate of the ancestral genome that free-living eubacterial ancestors possessed. Mitochondria have reduced genome size as a result of the interaction between them and their host cells. Genes once needed to support life as a free-living organism were lost or transferred to the nuclear genome of the

host eukaryote. One of the reasons why mitochondrial genes would benefit from being located in the nuclear genome is reduction in the accumulation of deleterious mutations. Asexually propagated genomes tend to build up their genetic load quicker than sexually propagated genomes, a principle known in population genetics as Muller's ratchet. Additionally, the formation of reactive oxygen species within mitochondria as a result of the process of respiratory electron transport increases the frequency of mutations, exacerbating the effects of the Muller's ratchet [2].

Mode of gene regulation, special properties of gene products, mechanisms of import of proteins into the mitochondrion, and other as yet unknown features may be acting against the complete transfer of all mitochondrial genes to the nucleus [reviewed in [3]]. Although essential genes are still located in the mitochondrial genome, amplification of nuclear copies of mitochondrial genes has been detected occasionally in several taxonomic groups [reviewed in [4]]. Non-functional nuclear copies or pseudogenes have been termed *Numts* (pronounced 'new-mights', for NUclear MiTochondrial DNA segments) by Lopez and collaborators [5], who found tandem-duplicated mitochondrial copies of a 7.9 kilobases (kb) fragment in the nuclear genome of cats. Subsequently, caution has been recommended when attempting to amplify authentic mitochondrial fragments by polymerase chain reaction (PCR) techniques, as the nuclear copies might be amplified in preference to mitochondrial ones, especially using conserved primers designed on gene sequences of organisms of different taxonomic levels [4,6,7].

With the completion of the sequencing of the human genome [8], extensive genomic analyses have found hundreds of *Numts* in the human nuclear genome [9-11]. These analyses indicate that nuclear copies are widespread in all human chromosomes and involve all mitochondrial genes and the control region. Some of these integrations encompass about 80% of the complete mitochondrial genome. Similarity between human *Numts* and their mitochondrial counterparts is as high as 99%, raising concerns for the fields of molecular population genetics and phylogenetics because PCR amplification and sequencing of mitochondrial DNA segments are major tools used to address many biological questions in ecology and evolution. High similarity of *Numts* with mitochondrial genes not only increases the chance of accidentally amplifying the nuclear copy but also lessens any suspicion that the fragment isolated is not of mitochondrial origin, and thus has the potential to invalidate the conclusions of many studies.

Although birds are well studied with over 815,000 sequences deposited in GenBank as of June 4, 2004, *Numts* have been reported for only four different avian orders (Table 1). In most cases, the nuclear integration involved the control region and cytochrome b gene (cyt *b*). Sequence divergence between these *Numts* and the corresponding mitochondrial segment varied from 2 to 31%. However, the extent and details of the *Numt* fraction in avian genomes will only be adduced when more nuclear genome sequences become available. The recent sequencing of the chicken (*Gallus gallus*) nuclear genome by the Genome Sequencing Center at Washington University, St. Louis and its availability for public access at the Ensembl [12] website provides an opportunity to check whether the high incidence and occasionally large size of *Numts* in the human genome also occur in this avian genome. Contrasting with humans, avian chromosomes are classified in macrochromosomes and microchromosomes, according to whether or not they are cytogenetically identifiable by conventional banding techniques. Consequently, chicken has a diploid number of 78 chromosomes classified in eight pairs of macrochromosomes, 30 pairs of

Table 1: Reported avian *Numts*. Genes name as in Figure 1.

Order	Species or group	Gene	Similarity to mtDNA	References
Anseriformes	*Anser caerulescens*	Control region	88.2 – 91%	37
Anseriformes	Aythyini	Control region	90.3 – 92.8%	38
Anseriformes	*Dendrocygna arcuata*	COI	88.4%	39
Anseriformes	*Somateria mollissima*	Control region	80.3%	40
Charadriformes	*Cepphus*	Control region	50%	41
Falconiformes	*Aquila*	Control region	72 – 95%	42
Falconiformes	*Buteo*	Control region	68.7 – 98.4%	43
Passeriformes	*Motacilla cinerea cinerea*	ND2	n.a.	44
Passeriformes	*Oeromanes*, *Conirostrum*	ND5	84 – 99%	15
Passeriformes	*Parus*	Cyt *b*	63%	45
Passeriformes	*Passer*	Cyt *b*	88.8 – 98%	46
Passeriformes	*Scytalopus* and *Myornis senilis*	Cyt *b*	81 – 84%	47
Passeriformes	Several species of Darwin's finches	Control region and cyt *b*	Substitution rate 2 – 4 times lower compared to mtDNA	48

Table 2: *Numts* detected in the chicken nuclear genome, and parameters of alignments returned on BLASTN searches. Start and end indicates positions of alignments in the chicken mitochondrial (mtDNA) and chromosomal (chrom) sequences. Orientation corresponds to whether integration in the nuclear genome is 5' > 3' (+) or 3' > 5' (-). E-val and % ID are respectively expected value and % of identity for each returned alignment. Some *Numts* were identified by more than one alignment.

Numt #	Genes included	start mtDNA	End mtDNA	Orientation	chromosome	Start on chrom.	End on chrom.	Orientation	Blast score	E-val	%ID	Length of alignment	Length of Numt
1	ND4	12341	12585	+	1	13059619	13059865	-	610	5.9e-32	74.10	251	848
	ND4	11738	12327	+	1	13059865	13060448	-	536	5.9e-32	58.65	607	-
2	$tRNA^{Glu}$ – CR – $tRNA^{Phe}$ – 12S	1063	1342	+	1	18250179	18250453	-	370	1.9e-05	61.97	284	1536
	$tRNA^{Glu}$ – CR – $tRNA^{Phe}$ – 12S	563	1139	+	1	18250273	18250834	-	743	2.4e-22	62.29	586	-
	$tRNA^{Glu}$ – CR – $tRNA^{Phe}$ – 12S	16582	16748	+	1	18250802	18250966	-	364	3.5e-05	69.46	167	-
3	$tRNA^{Ser}$ – $tRNA^{Leu}$	12941	13069	-	1	132949128	132949256	+	397	3.8e-05	80.15	131	131
4	ND4 – $tRNA^{His}$ – $tRNA^{Ser}$	12572	12980	-	2	48703806	48704212	+	1598	5.8e-61	88.83	412	412
5	ND5	13569	13804	+	2	88002064	88002298	+	460	2.2e-08	67.78	239	239
6	$tRNA^{Trp}$ – $tRNA^{Ala}$	6274	6404	-	4	1975588	1975719	-	381	6.2e-06	79.26	135	135
7	16S – $tRNA^{Leu}$ – ND1	3314	3389	+	4	22850352	22850426	+	257	5.5e-70	82.89	76	1182
	16S – $tRNA^{Leu}$ – ND1	3695	4169	+	4	22850419	22850882	+	1101	5.5e-70	73.50	483	-
	16S – $tRNA^{Leu}$ – ND1	4200	4498	+	4	22850879	22851175	+	737	5.5e-70	74.01	304	-
8a	ND5 – cyt *b*	14548	14983	-	4	47771664	47772104	+	1005	3.8e-34	72.67	450	1733
8b	ND5	13251	13612	+	4	47772159	47772520	+	426	5.8e-08	63.52	381	-
9	CR	476	842	-	4	72432412	72432775	+	572	1.1e-14	65.96	379	782
	CR	141	608	-	4	72432642	72433107	+	405	4.3e-07	60.37	492	-
	CR	61	526	-	4	72432662	72433124	+	458	1.7e-09	60.33	484	-
10	ND4	11603	11748	+	9	21658995	21659139	-	436	1.7e-07	78.08	146	146
11	CR	309	505	+	15	7657960	7658157	+	479	1.9e-10	73.76	202	202
12	Cyt *b*	15247	15565	-	19	5161850	5162161	-	797	7.6e-25	73.67	319	319
13	CO2 – $tRNA^{Lys}$ – ATP8/6	9192	9883	+	27	2160604	2161286	-	2022	3.1e-81	77.92	693	1204
	CO2 – $tRNA^{Lys}$ – ATP8/6	8680	9630	+	27	2160837	2161782	-	1393	4.0e-52	65.21	986	-

microchromosomes, and one pair of sex chromosomes ZW. We searched for mitochondrial pseudogenes in the chicken genome, and provide a descriptive characterization of *Numts* found. We also compare our results to other similar studies on sequenced eukaryotic genomes and show that the frequency and amplification of *Numts* varies from species to species, and that the numerous *Numts* found in the human and plant genomes may be the exception to the general rule in eukaryotes.

Results

Results of the BLAST search for sequences in the chicken nuclear genome that have homology with chicken mitochondrial DNA revealed 22 alignments that seemed to be biologically significant as defined by our threshold of 10^{-4} (Table 2). Moreover, size of alignments and similarity between legitimate mitochondrial sequences and their homologues in the nuclear genome indicate that the nuclear homologues could represent ancient degenerate mitochondrial sequences. Careful inspection of returned alignments led us to infer the presence of at least 13 mitochondrial fragments into the nuclear genome. They are numbered 1–7, 8a, 8b, 9–13. Identities between *Numts* and corresponding mitochondrial sequences varied from 58.6 to 88.8%.

Regarding size of mitochondrial pseudogenes, we found six *Numts* ranging from 782 – 1,733 bp that were recovered by two or more alignments, and seven *Numts* of 131 – 412 bp recovered by a single alignment. Considered together, *Numts* contributed 8,869 bp or 0.00078% of the nuclear genome of the chicken.

Ten protein-coding genes (except for ND2, ND3, ND4L, ND6, CO1 and CO3), ribosomal genes, the control region (CR), and 10 of 22 tRNAs were found in *Numts*. The mitochondrial control region, ND5 and ND4 were each found in three different *Numts*, followed by cyt *b*, which was detected in two *Numts*. All other genes included in a *Numt* were present only once (Table 2; Fig. 1).

Most genes found in chicken *Numts* correspond to partial mitochondrial sequences. Complete sequences for the control region, $tRNA^{Phe}$, $tRNA^{His}$, $tRNA^{Leu}$, $tRNA^{Lys}$, and ATPase 8 were found in *Numts* 2, 4, 7, and 13. Regardless of the completeness of the mitochondrial pseudogene in the nuclear genome, all protein-coding genes had internal stop codons and/or frame-shift mutations. No tRNA *Numt* could be perfectly folded in its predicted secondary structure, with the exception of the $tRNA^{His}$ that had only one substitution compared to the mitochondrial counterpart, and that did not interfere with its secondary structure (Fig. 2). However, this tRNA is part of a bigger nuclear fragment (containing partial sequences for $tRNA^{Ser}$ and ND4) that has high similarity with the chicken mitochondrial fragment, indicating that it may be a recent *Numt*.

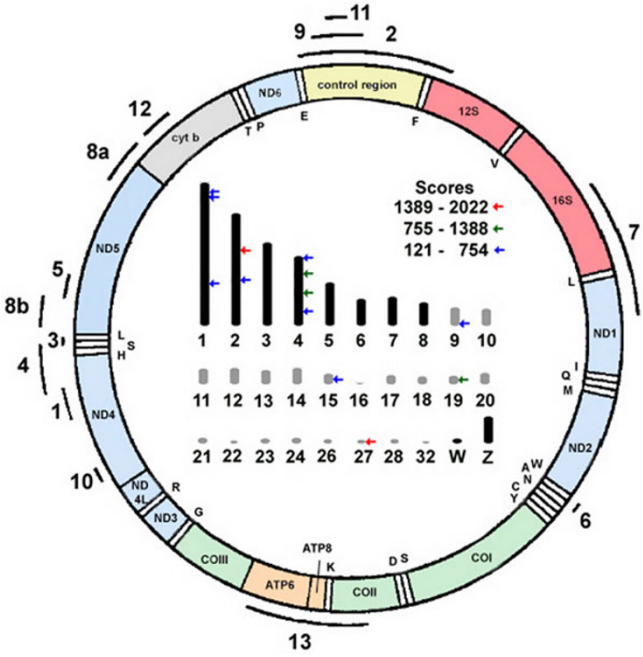

Figure 1
Representation of the chicken mitochondrial genome and chicken karyotype. Gene names are as follows: cyt *b* – cytochrome *b*; COI, COII and COIII – subunits I, II and III of cytochrome oxidase; ND1-6 – subunits 1 to 6 of NADH reductase; tRNAs are represented by their IUPAC one-letter amino acid abbreviations; ribosomal gene subunits are represented by 12S and 16S. Relative position of each *Numt*, and their numbers as in Table 2, are shown outside the circular mitochondrial genome. A karyotype representation for chicken is shown inside the circular mitochondrial genome. Chromosomes 1–8 are macrochromosomes, W and Z are sex chromosomes, and all others are microchromosomes. Not all chicken microchromosomes can be unambiguously identified by conventional banding techniques, and they are not represented here. Range for BLAST scores is also shown.

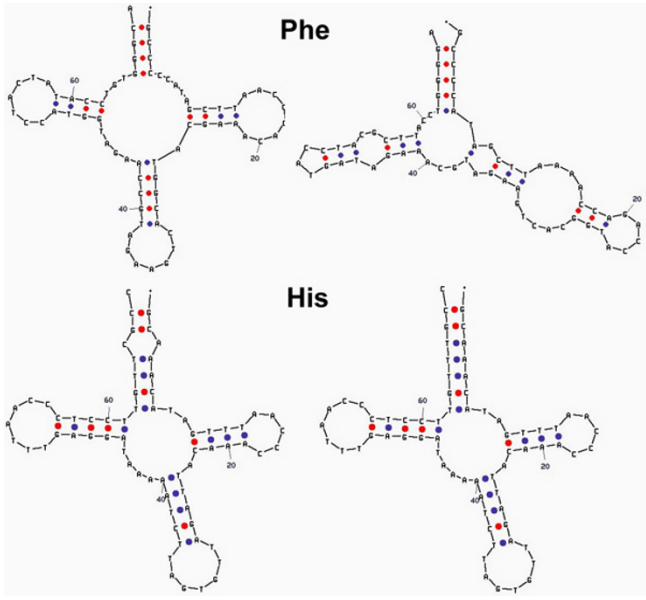

Figure 2
Prediction of secondary structure for tRNAs. Secondary structure for legitimate mitochondrial $tRNA^{Phe}$ and $tRNA^{His}$ are shown to the left, and their corresponding nuclear pseudogenes to the right.

The site of *Numt* integration was further analyzed using the graphic interface available at the Ensembl website [12], using the mitochondrial region and chromosomal positions given in Table 2. We found that integrations occurred in regions where no known or predicted genes were located. These regions were also rich in repeat elements like LINES, microsatellites and low complexity repeats, but with no apparent association between them and *Numts*. Ten *Numts* were localized in three macrochromosomes, and the remaining four in different microchromosomes (Fig. 1). No *Numts* were identified in the sex chromosomes W and Z, or in contigs not yet assigned to chromosomes. Chromosome 4 (GGA4) had three of the five largest *Numts* detected in the nuclear genome.

Two *Numts* deserve more consideration. *Numt* 7 in GGA4 was located at the very 3' end of the contig, and it may be longer than 1,182 bp. However, it is necessary to close the gap between the contig where it was found and the adjacent contig to check the extension of this integration. *Numt* 8a and 8b, also found in GGA4, would be considered to represent two independent integrations according to our criteria (e.g. they are inserted in opposite directions and they do not overlap). However, because the region between them spans 816 bp in the nuclear genome and this is similar to the missing fragment of 936 bp in the mitochondrial genome, they may have been part of one transfer event that was later involved in a chromosomal rearrangement, leading to change of orientation of one of the fragments. Moreover, the presence of an intercalated microsatellite at the 3' end of *Numt* 8a and 5' end of *Numt* 8b indicates that the rearrangement is more plausible than two independent transfers.

The mechanism of mitochondrial integration in the nuclear genome may be via RNA [13] or DNA [5], and can be identified by checking the 5' and 3' ends of the genes involved. Integration via a DNA intermediate is the most common mechanism in the human genome [10]. In chicken, *Numts* 2, 9, and 11 contain the CR and the integration is clearly by a DNA intermediate, as the CR is not transcribed. Polycistronic mitochondrial RNA transcripts

are quickly processed as they are transcribed, and mRNAs do not have polyadenylation signals at the 5' end [14]. Therefore *Numts* 3, 4, 6, 7, 8a, 8b and 13 were also integrated via DNA as these signals of processing were not present. Mode of integration cannot be inferred for *Numts* 1, 5, 10 and 12 as they represent integrations of partial fragments of protein-coding genes with no associated neighboring gene in the same integration, and no end is present in the *Numt* to check for these processing signatures. We also discarded the possibility that any of these *Numts* found in the chicken nuclear genome originated from a duplication of another *Numt*. Furthermore, we found no evidence of tandem repeats around the site of integration.

Discussion

Numts in the chicken and other eukaryotic genomes

Our search of the chicken nuclear genome indicates the presence of 13 apparently independent integrations of mitochondrial DNA genes. Two of these *Numts* may actually represent a single integration that underwent rearrangement resulting in loss of an intermediate region and change of the orientation of one of the remaining fragment. Such rearrangements of *Numts* have been detected previously in birds and humans [9,15]. No correlation seems to exist between the size of a nuclear genome and number of *Numts*, although bigger genomes and larger chromosomes can bear more integrations [11]. Although no clear site for integration of *Numts* has been recognized so far, regions with low gene content are more prone to integrations [10], which probably avoid disruption of well-organized gene complexes in gene-rich regions, and therefore survival of the integration in the nuclear genome. In chicken, most insertions were detected in macrochromosomes that are low in gene content compared to microchromosomes [16,17]. Two mitochondrial DNA regions were identified as hotspots for insertions into the nuclear genome, one at the control region and the other encompassing the intervening sequence between ND4 and cyt *b*. Although most *Numts* detected in PCR products in birds are examples of the integration of the CR or cyt *b* genes (Table 1), this is a consequence of these genes being the most targeted for amplification in ecological and evolutionary studies compared to other regions of the mitochondrial genome. As our analysis was performed in a pre-assembled version of the chicken genome, other *Numts* may be found when the complete assembly is released. However, the conclusions of our study should still hold as the genomic assembly we searched included contigs not yet assigned to chromosomes.

The number of *Numts* found in chicken is within the range found in most sequenced eukaryotic genomes (Fig. 3). That is, mitochondrial pseudogenes do not seem to represent a large portion of eukaryotic genomes, and with the exception of human, mouse and plants, they number less than 100. Our results are consistent with the observation that avian genomes harbor less repetitive elements and other non-coding sequences [18,19]. Only 17% of the chicken genome is assumed to be composed of repetitive elements including LINEs, SINEs, microsatellites, minisatellites and simple repeats [20] compared to 40 – 50% of the genome of humans and rodents [8,21,22]. Flight has been claimed to impose constraints on the size of bird genomes, and there is a positive association between genome size and flying abilities: stronger fliers posses smaller genomes than weak fliers [23]. As flight demands a high metabolic rate and, and high metabolic rate in turn restricts cell size, genome content is expected to be reduced to fit a small cell. These same reasons appear to explain why bats have small genomes [23-25] and provide independent evidence for the association between flight and compact genomes in homoeothermic vertebrates. Because the number of genes in chicken is similar to those in human, small genome size in chicken has been achieved in part by loss of repetitive DNA and gene-poor chromosomal regions where most *Numts* occur. Metabolic selection for compact genomes could therefore explain the low incidence of *Numts* that have been observed in birds. Although sequenced plant genomes of *Arabidopsis* and rice are smaller than those of humans and chicken, they have a high number of *Numts*. The reasons for this discrepancy are not well understood, but plant genomes seem to be able to harbor a large number of repetitive elements and to transfer DNA bidirectionally between chloroplasts and mitochondria [28-30].

Implications for inference of population history and phylogenetics

As mitochondrial DNA is one of the main sources of information for population genetics and phylogenetics at several taxonomic depths, the inadvertent amplification of *Numts* via PCR technologies may seriously impact studies and lead to erroneous conclusions about phylogeography and taxon relationships. For example, in a recent study of great apes [7], *Numts* seem to have been preferentially amplified in gorillas, and similarity of these inserts with mitochondrial copies was high enough to avoid suspicion *Numts* were amplified. Also, the demonstration that the human genome has hundreds of *Numts* representing all mitochondrial regions including large portions of the mtDNA molecule, some of which have high similarity with their mitochondrial counterparts, has raised concerns that *Numts* may have gone undetected in many studies published in the last decade. This problem would be especially acute if *Numts* are a common feature of genomes. Fortunately, it seems that *Numts* are not as frequent in most sequenced eukaryotic genomes as they are in humans or plants (Fig. 3).

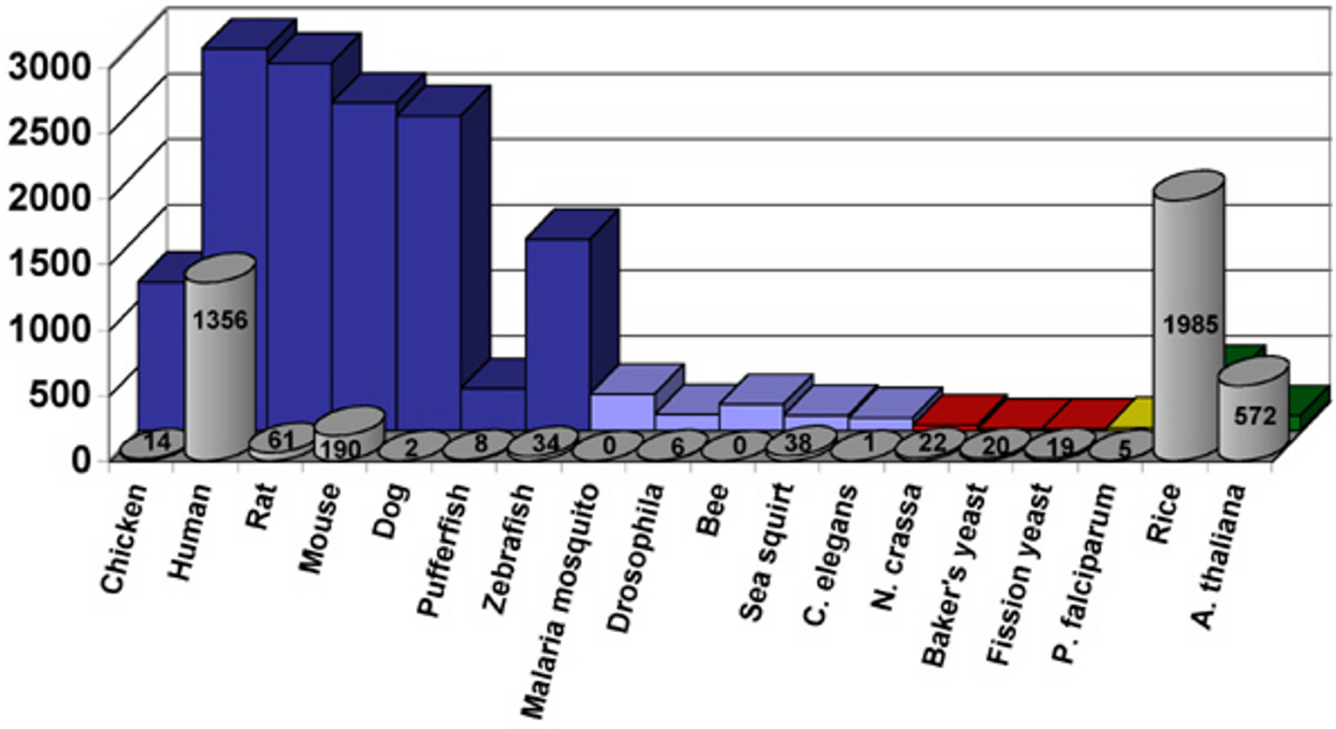

Figure 3
Size of nuclear genome for eukaryotes and number of *Numts* detected. Scale to the left is genome size in Megabases. Numbers of *Numts* are indicated on gray cylinders. Data is from [11], except for chicken, dog, zebrafish and bee. See material and methods for more details.

In birds, for example, mitochondrial pseudogenes have been occasionally detected (Table 1), but numbers reported may actually be underestimates as not all findings of *Numts* are formally published. Unfortunately, these studies do not provide information on the real extent of *Numts* because the experimental design was not aimed at a search for such elements. Our search indicates that the size of *Numts* detected in the chicken genome is often smaller than the usual size of the fragments isolated by PCR technology (>600 bp) in most published studies. Also, similarities between chicken *Numts* and their mitochondrial counterparts were below 89%, and the presence of indels, stop codons or frame-shift mutations would clearly indicate the amplification of a pseudogene instead of a fragment of the mitochondrial genome. In humans, some large *Numts* representing about 80% of the total mitochondrial genome have been found. However, most human mitochondrial pseudogenes are smaller than 500 bp [26]. Collectively these observations imply that the amplification of a *Numt* will be rare if mitochondrial fragments targeted for amplification are above the size range of most described *Numts*.

If the *Numts* in chicken are typical of those in other birds, amplifications of fragments of about 1.5 kb are highly likely to represent true mitochondrial amplification, and are economically more viable than performing amplification of very large segments (e.g. > 5 kb) of the mitochondrial genome, or by cloning PCR products. Moreover, sequencing of fragments of about 1.5 kb should easily detect the presence of unusual features of pseudogenes unless the nuclear integration is very recent. However, if PCR amplification results in more than one band, or sequence ambiguities or background signal are present, direct PCR amplification may not produce authentic mtDNA sequences. In this case other methodologies such as isolation of mitochondria from cells previous to DNA isolation, or isolation of DNA from mitochondria-enriched tissues may provide a solution. Use of conserved primers increases the chance that they might preferentially anneal to a *Numt*, as they are effectively molecular fossils because they have a slower rate of DNA substitution than does mitochondrial DNA [27,31].

Conclusions

We have shown that the numbers of *Numts* in the nuclear genome of the chicken is low compared to what was found in the genome of humans and plants. Although caution must still be exercised in PCR-based studies, the small size and sequence divergence of these chicken pesudogenes from mitochondrial copies indicates that they may be less of a concern in mtDNA-based studies of birds relative to primates and plants. However, we will only know to what extent these findings apply generally to avian genomes when sequences of more diverse bird taxa are completed.

Methods

Sequence analysis

The full-length mitochondrial genome for chicken [32] was retrieved from The National Center for Biotechnology Information database [33] under accession number NC_001323 and used to perform similarity searches against a database of the draft sequence of the chicken nuclear genome released by the Genome Sequencing Center at Washington University, St Louis (Build WASHUC1) and publicly available at the Ensembl Genome Browser [12,34] as of March 2004. BLAST [35] searches were used, with the whole mitochondrial genome sequence or mitochondrial genes individually as query. Results from both strategies were the same. We set the maximum expectation value in BLASTN searches to be $e = 10^{-4}$ to recover hits that are biologically significant. No filters were used during searches. Assuming e values in the range of 10^{-4} to 10 resulted in extra hits that have lower similarity with the query sequence, and shorter alignments, therefore indicating the randomness of these hits. Further analyses indicated that some recovered alignments represented short T-rich regions in the nuclear genome that aligned with a short T-stretch present at the beginning of the chicken mitochondrial control region.

Identification of mitochondrial integrations in the nuclear genome

Results from searches were analyzed via BLASTView, a graphic interface that displays the results after a BLAST search in the Ensembl website. For all recovered alignments that had similarity between mitochondrial and nuclear genomes above 50% and a significant e value, we downloaded the contigs where these BLAST hits were observed to investigate the characteristics of the mitochondrial pseudogene in the nuclear genome, also known as *Numt* [5]. For most *Numts*, contig and chromosomal position was obtained from BLASTView, and information on the region of integration was gathered by examining the maps and annotation provided in ContigView and ExportView links, respectively. When two alignments were returned for the same contig, they were merged and considered to be the same integration event if they were in the same orientation and had overlapping bases. When gaps between alignments in the same contig were observed they were considered to be the same if the gap was similar in size to the mitochondrial fragment expected to fill this gap. Also, for the later case, we checked the intervening sequence for the possible presence of insertion or deletion of nucleotides.

Secondary structure for tRNAs involved in Numts

Prediction of secondary structure for legitimate tRNAs and their *Numts* which had the complete tRNA sequence were obtained using the DNA mfold web server [36]. Folding temperature used was the default set to 37°C. For some tRNAs, some bases were forced to pair to obtain the expected mitochondrial tRNA structure as previously described [32].

Search for Numts in other available genomes

A recent study has summarized the distribution of *Numts* in a variety of organisms [11]. However, they did not include information on *Numts* for dog, zebrafish and bee genomes that have only recently became available in GenBank. Therefore, we performed an initial analysis for *Numts* in these genomes (GenBank Builds cra_dog_assembly, zebrafish_HTGS 1.1, Amel 1.1, respectively), using the corresponding full-length mitochondrial sequence. The database for these organisms is pre-draft assembly available in GenBank in late March, 2004. In these searches, we only recorded the number of alignments found by a BLASTN searches, and no further analysis was performed to evaluate the overlap between alignments. This same procedure was adopted in [11]. Therefore, caution is necessary in the interpretation of the number of *Numts* reported in those organisms and in chicken. Our goal was to have a rough estimate of number of *Numts* in these genomes for comparative purposes only.

Author's contributions

Both authors conceived and designed the study. SLP carried out the genomic analyses, drafted the manuscript and drew the figures. AJB assisted with drafting, revising and editing the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

The chicken nuclear genome sequence was determined by whole genome shotgun method at the Genome Sequencing Center at Washington University, St. Louis and is publicly available at Ensembl http://www.ensembl.org/Gallus_gallus/. We are grateful to Dr. Laila A. Nahum and two anonymous reviewers for helpful discussion and suggestions. Our work was supported by NSERC grant 200-02 to AJB.

References

1. Margulis L: *Origin of Eukaryotic Cells* New Haven: Yale University Press; 1970.

2. Race HL, Herrmann RG, Martin W: **Why have organelles retained genomes?** *Trends in Genetics* 1999, **15:**364-370.
3. Lang BF, Gray MW, Burger G: **Mitochondrial genome evolution and the origin of eukaryotes.** *Annual Reviews of Genetics* 1999, **33:**351-397.
4. Bensasson D, Zhang D-X, Hartl DL, Hewitt GM: **Mitochondrial pseudogenes: evolution's misplaced witness.** *Trends in Ecology and Evolution* 2001, **16:**314-321.
5. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ: **Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat.** *Journal of Molecular Evolution* 1994, **39:**174-190.
6. Zhang D-X, Hewitt GM: **Nuclear integrations: challenges for mitochondrial DNA markers.** *Trends in Ecology and Evolution* 1996, **11:**247-251.
7. Thalmann O, Hebler J, Poinar HN, Pääbo S, Vigilant L: **Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of human and other great apes.** *Molecular Ecology* 2004, **13:**321-335.
8. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
9. Tourmen Y, Baris O, Dessen P, Jacques C, Malthièry Y, Reynier P: **Structure and chromosomal distribution of human mitochondrial pseudogenes.** *Genomics* 2002, **80:**71-77.
10. Woischnik M, Moraes CT: **Pattern of organization of human mitochondrial pseudogenes in the nuclear genome.** *Genome Research* 2002, **12:**885-893.
11. Richly E, Leister D: ***Numts* in sequenced eukaryotic genomes.** *Molecular Biology and Evolution* 2004, **21:**1081-1084.
12. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slator G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M: *The Ensembl genome database project. Nucleic Acids Res* 2002, **30:**38-41.
13. Nugent JM, Palmer JD: **RNA-mediated transfer of the coxII from the mitochondrion to the nucleus during flowering plant evolution.** *Cell* 1991, **66:**473-481.
14. Taanman J-W: **The mitochondrial genome: structure, transcription, translation and replication.** *Biochimica et Biophysica Acta* 1999, **1410:**103-123.
15. Nielsen KK, Arctander P: **Recombination among multiple mitochondrial pseudogenes from a passerine genus.** *Molecular Phylogenetics and Evolution* 2001, **18:**362-369.
16. MacQueen HA, Siriaco G, Bird AP: **Chicken microchromosomes are hyperacetylated, early replicating, and gene rich.** *Genome Research* 1998, **8:**621-630.
17. Smith J, Bruley CK, Paton IR, Dunn I, Jones CT, Windsor D, Morrice DR, Law AS, Masabanda J, Sazanov A, Waddington D, Fries R, Burt DW: **Differences in gene density on chicken macrochromosomes and microchromosomes.** *Animal Genetics* 2000, **31:**96-103.
18. Holmquist GP: **Evolution of chromosome bands: molecular ecology of noncoding DNA.** *Journal of Molecular Evolution* 1989, **28:**469-486.
19. Primmer CR, Raudsepp T, Chowdhary BP, Moller AP, Ellegren H: **Low frequency of microsatellites in the avian genome.** *Genome Research* 1997, **7:**471-482.
20. Clark MS, Edwards YJK, McQueen HA, Meek SE, Smith S, Umrania Y, Warner S, Williams G, Elgar G: **Sequence scanning chicken cosmids: a methodology for genome screening.** *Gene* 1999, **227:**223-230.
21. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
22. Rat Genome Sequencing Project Consortium: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution.** *Nature* 2004, **428:**493-521.
23. Hughes AL: *Adaptive evolution of genes and genomes* Oxford: Oxford University Press; 1999.
24. Tiersch TR, Wachtel SS: **On the evolution of genome size of birds.** *Journal of Heredity* 1991, **82:**363-368.
25. Gregory TR: **A bird's-eye view of the C-value enigma: genome size, cell size, and metabolic rate in the Class Aves.** *Evolution* 2002, **56:**121-130.
26. Bensasson D, Feldman MW, Petrov DA: **Rates of DNA duplication and mitochondrial DNA insertion in the human genome.** *Journal of Molecular Evolution* 2003, **57:**343-354.
27. Fukuda M, Wakasugi S, Tsuzuki T, Nomiyama H, Shimada K, Miyata T: **Mitochondrial DNA-like sequences in the human nuclear genome.** *Journal of Molecular Biology* 1985, **186:**257-266.
28. Kubis S, Schmidt T, Heslop-Harrison JS: **Repetitive DNA elements as a major component of plant genomes.** *Annals of Botany* 1998, **82(Supplement A):**45-55.
29. Schmidt T: **LINEs, SINEs and repetitive DNA: non-retrotransposons in plant genomes.** *Plant Molecular Biology* 1999, **40:**903-910.
30. Bennetzen JL: **Transposable element contributions to plant gene and genome evolution.** *Plant Molecular Biology* 2000, **42:**251-269.
31. Perna NT, Kocher TD: **Mitochondrial DNA: molecular fossils in the nucleus.** *Current Biology* 1996, **6:**128-129.
32. Desjardins P, Moraes R: **Sequence and gene organization of the chicken mitochondrial genome.** *Journal of Molecular Biology* 1990, **212:**599-634.
33. **National Center for Biotechnology Information database** [http://www.ncbi.nlm.nih.gov/]
34. **Ensembl Chicken Genome Browser** [http://www.ensembl.org/Gallus_gallus/]
35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215:**403-410.
36. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Research* 2003, **31:**3406-3415.
37. Quinn TW, White BN: **Analysis of DNA sequence variation.** In *Avian Genetics* Edited by: Cooke F, Buckley PA. London: Academic Press; 1987:163-198.
38. Sorenson MD, Fleischer RC: **Multiple independent transpositions of mitochondrial DNA control region sequences to the nucleus.** *Proceedings of the National Academy of Science (U.S.A.)* 1996, **93:**15239-15243.
39. Sorenson MD, Quinn TW: ***Numts*: A challenge for avian systematics and population biology.** *The Auk* 1998, **115:**214-221.
40. Tiedemann R, von Kistowski KG: **Novel primers for the mitochondrial control region and its homologous nuclear pseudogene in the Eider duck *Somateria mollissima*.** *Animal Genetics* 1998, **29:**468.
41. Kidd MG, Friesen VL: **Sequence Variation in the Guillemot (Alcidae: *Cepphus* mitochondrial control region and its nuclear homolog.** *Molecular Biology and Evolution* 1998, **15:**61-70.
42. Väli Ü: **Mitochondrial pseudo-control region in old world eagles (genus *Aquila*).** *Molecular Ecology* 2002, **11:**2189-2194.
43. Riesing MJ, Kruckenhauser L, Gamauf A, Haring E: **Molecular phylogeny of the genus *Buteo* (Aves: Accipitridae) based on mitochondrial marker sequences.** *Molecular Phylogenetics and Evolution* 2003, **27:**328-42.
44. Ödeen A, Björklund M: **Dynamics in the evolution of sexual traits: losses and gains, radiation and convergence in yellow wagtails (*Motacilla flava*).** *Molecular Ecology* 2003, **12:**2113-2130.
45. Kvist L, Ruokonen M, Orell M, Lumme J: **Evolutionary patterns and phylogeny of tits and chikadees (genus *Parus*) based on the sequence of the mitochondrial cytochrome *b* gene.** *Ornis Fennica* 1996, **73:**145-156.
46. Allende LM, Rubio I, Ruíz-del-Valle V, Guillén J, Martínez-Laso J, Lowy E, Varela P, Zamora J, Arnaiz-Villena A: **The Old World sparrows (genus *Passer*) phylogeography and their relative abundance of nuclear mtDNA pseudogenes.** *Journal of Molecular Evolution* 2001, **53:**144-154.
47. Arctander P: **Comparison of a mitochondrial gene and a corresponding nuclear pseudogene.** *Proceedings of the Royal Society of London B* 1995, **262:**13-19.
48. Sato A, Tichy H, O'hUigin C, Grant PR, Grant BR, Klein J: **On the origin of Darwin's finches.** *Molecular Biology and Evolution* 2001, **18:**299-311.