Comparative Genomics Reveals a Burst of Homoplasy-Free *Numt* Insertions

Bin Liang,*[,†,1,2,3] Ning Wang,[†,2] Nan Li,[4] Rebecca T. Kimball,[1] and Edward L. Braun[1,*]

[1]Department of Biology, University of Florida, Gainesville, FL

[2]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI

[3]Forestry Research Institute of Hainan Province, Haikou, Hainan, P. R. China

[4]Department of Chemistry and Biochemistry, UC San Diego, La Jolla, CA

[†]These authors contributed equally to this work.

***Corresponding authors:** E-mails: woodybir_d@hotmail.com; ebraun68@ufl.edu.

Associate editor: Anne Yoder

Abstract

Mitochondrial DNA sequences are frequently transferred into the nuclear genome, giving rise to *numts* (*nu*clear *mi*tochondrial DNA segments). In the absence of whole genomes, avian *numts* have been suggested to be rare and relatively short. We examined 64 bird genomes to test hypotheses regarding *numt* frequency, distribution among taxa, and likelihood of homoplasy. We discovered 100-fold variation in *numt* number across species. Two songbirds, *Geospiza fortis* (Darwin's finch) and *Zonotrichia albicollis* (white-throated sparrow) had the largest number of *numts*. Ancestral state reconstruction of 957 *numt* insertions in these two species and their close relatives indicated a remarkable acceleration of *numt* insertion in the ancestor of *Geospiza* and *Zonotrichia* followed by slower, continued accumulation in each lineage. These *numts* appear to result primarily from de novo insertion with the duplication of existing *numts* representing a secondary pathway. Insertion events were essentially homoplasy-free and *numts* appear to represent perfect rare genomic changes.

Key words: **rare genomic changes, avian genome, *numts*, homoplasy-free, rate burst.**

Letter

*Nu*clear *mi*tochondrial DNA segments (*Numts*), which result from insertion of mitochondrial DNA (mtDNA) into the nuclear genome, have been described in numerous eukaryotes (Lopez et al. 1994; Hazkani-Covo et al. 2010). The genomics era has led to an explosion of information about the *numt* repertoire of taxa ranging from fungi and plants to mammals (Hazkani-Covo et al. 2010). Those studies revealed substantial heterogeneity in numbers of *numts* across taxa (Bensasson et al. 2001; Hazkani-Covo et al. 2010), with some species (e.g., humans) having relatively large numbers (e.g., Mourier et al. 2001; Tourmen et al. 2002). The factors that result in variation in *numt* numbers among genomes are unclear, but genome size may be important (Hazkani-Covo et al. 2010).

Historically, systematists focused on avoiding PCR amplification of *numts* because they can mislead barcoding, phylogenetic, and phylogeographic studies (Bensasson et al. 2001; Bertheau et al. 2011). In the genomic era, however, *numts* are of intrinsic evolutionary interest for several reasons (Bensasson et al. 2001; Tourmen et al. 2002; Leister 2005). First, the rate of nuclear sequence evolution is lower than the mitochondrial rate in vertebrates, so *numts* can provide information about ancestral mitochondrial sequences (Hu and Thilly 1994; Zischler et al. 1995; Bensasson et al. 2001; Tourmen et al. 2002). Second, *numts* could represent a novel type of rare genomic changes (RGC) for systematic studies (Hazkani-Covo 2009). Ideal RGCs are homoplasy-free and can accurately reconstruct a single bipartition in their associated gene tree. Therefore, identifying ideal RGCs is valuable for phylogenomic studies.

Avian *numts* have been suggested to be uncommon and relatively short (Pereira and Baker 2004; but see Nacer and do Amaral 2017 regarding falcons). If genome size correlates with *numt* numbers, the low variation in avian genome sizes (Zhang et al. 2014) suggests limited variation in *numt* numbers among species. However, avian studies have rarely leveraged complete genomes. Using 64 complete avian genomes, we tested the following hypotheses: 1) *numts* are uncommon in avian genomes; 2) numbers of *numts* do not vary substantially among taxa; and 3) *numts* fit an RGC model and exhibit little or no homoplasy.

Different Avian Genomes Exhibit Remarkable Variation in *Numt* Content

We discovered remarkable variation across species in *numt* number, ranging from four to >600 *numts* (fig. 1A), the highest values approaching the upper estimates from the human genome (Hazkani-Covo et al. 2010). Thus, we can reject our first two hypotheses (*numts* are uncommon and numbers of *numts* do not vary much among taxa). We examined <1% of extant bird species (Brown et al. 2017), thus, it is reasonable to expect broader surveys to reveal additional variation in *numt* number. One clade was a clear outlier: the White-throated sparrow (*Zonotrichia albicollis*, a New world sparrow) and the Medium ground finch (*Geospiza fortis*, a tanager and one of

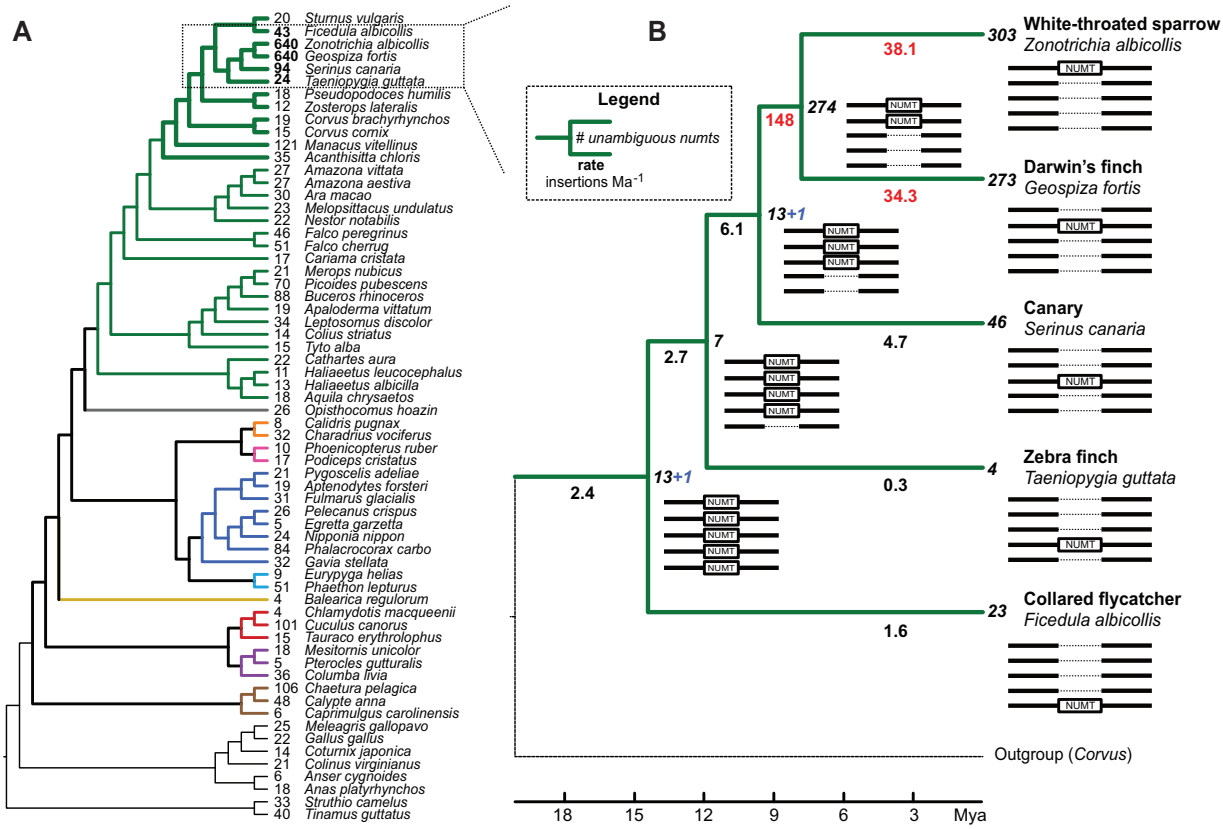

Fig. 1. *Numt* repertoire of avian genomes. (*A*) Number of *numts* in each bird genome mapped on the Brown et al. (2017) synthetic tree topology, with branches shaded to indicate the major clades described in Reddy et al. (2017). The five species examined in more detail are bolded. The number of *numt* insertions we identified in falcons is similar to Nacer and do Amaral (2017). (*B*) Time-calibrated tree for the five focal passerines indicating the number of *numt* loci where presence/absence could be scored in all taxa. Character states are shown graphically and the number of *numts* with each character state is indicted using italic numbers, either at the internal nodes (for *numt* insertions found in multiple taxa) or adjacent to the taxon names (for *numts* in a single taxon). *Numt* insertions where the taxonomic distribution could not be unambiguously assigned were omitted from part B (see supplementary fig. S7, Supplementary Material online). Estimates of *numt* insertion rates are presented below branches (the elevated rates in *Geospiza*, *Zonotrichia*, and their ancestral lineage are highlighted). "+1" indicates a potentially homoplastic locus.

Darwin's Galápagos finches). The *numts* in these taxa are unlikely to reflect genome assembly errors because the genomes for both taxa reflect high-coverage sequencing (>63×) and the assemblies were generated using different software. Moreover, the *Geospiza* and *Zonotrichia numt* insertions met the stringent criteria described in our supplementary methods, Supplementary Material online. Finally, these genomes are similar in size to other avian species (Zhang et al. 2014), indicating the large numbers of *numt* insertions in these taxa do not reflect a simple correlation with genome size.

A Burst of *Numt* Insertions in Tanagers and New World Sparrows

We focused additional analyses on a clade of five passerines that included *Geospiza* and *Zonotrichia* (fig. 1*A*). We identified 957 orthologous loci with *numts* present in at least one taxon, and scored the *numts* as present or absent in all five taxa. An additional 74 orthologous loci were found in some, but not all taxa; this could reflect the loss of the orthologous sequence in some taxa or a failure of some genome assemblies to include the orthologous region. The insertions in *Geospiza* or *Zonotrichia* totaled over 160 kilobases (kb) and covered the entire mitochondrial genome (fig. 2). It appears that all mtDNA regions have a similar propensity to integrate into the nuclear genome, with specific regions having a modest overrepresentation at most. Most *numt* loci included a single mtDNA sequence, although some *numt* loci comprised two or three mtDNA segments (supplementary methods and tables S1 and S2, Supplementary Material online).

Examination of the 957 loci without missing data revealed 649 *numts* unique to a single taxon (fig. 1*B*). Of the shared insertions, 33 *numts* appeared to reflect insertions prior to the divergence of *Serinus canaria* (canary) from the *Geospiza-Zonotrichia* clade; 13 of those *numts* appear to have been present in the common ancestor of all five passerines (fig. 1*B*). A much larger number of *numts* were present in the common ancestor of *Geospiza* and *Zonotrichia*; the estimated insertion rate is approximately two orders of magnitude greater than the rate outside of these two species (fig. 1*B*). The rate for the *Geospiza* and *Zonotrichia* terminal branches was lower, though it still exceeded the ancestral insertion rate. These results suggest a remarkable pulse of *numt* insertions followed by a decrease within the *Geospiza-Zonotrichia* clade.

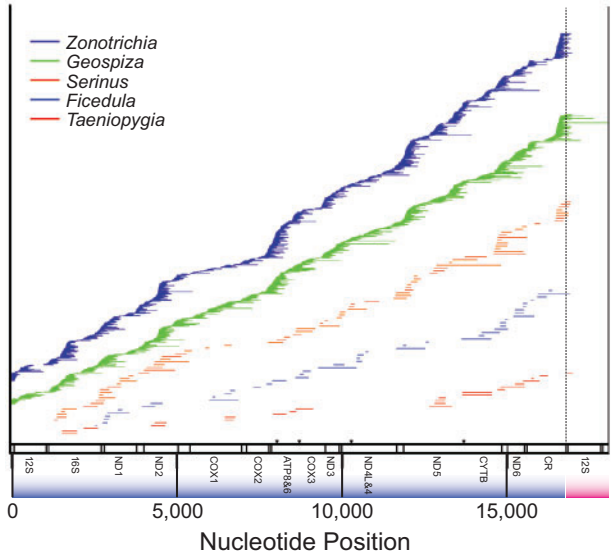

Fig. 2. Positions in the mitochondrial genome matching *numts* in each focal taxon (form top to bottom are: *Zonotrichia, Geospiza, Serinus, Ficedula, and Taeniopygia*). Ribosomal RNAs, protein-coding genes, and the control region are defined on the map, transfer RNA genes are indicated with lines. Because the mitogenome is circular, this map begins at the tRNA-Phe, immediately before the 12S ribosomal RNA. *Numts* that span the (arbitrarily chosen) start point of our map are shown at the end; this necessitated repeating the 12S ribosomal RNA (we indicate this using a dotted line and different shades below the gene map).

There are three potential models for *numt* insertions across our focal taxa: 1) insertions occurred at a constant rate (continuous); 2) a single rate shift in the common ancestor of *Geospiza* and *Zonotrichia*; and 3) two rate shifts occurred (an increase in their common ancestor followed by a decrease). This third model can be viewed as a punctuated model; such patterns have been documented in both primates and plants (e.g., Adams et al. 2002; Gunbin et al. 2017). Our insertion rate estimates (fig. 1*B*) are most consistent with the third model. However, there is evidence of recent insertions in both *Geospiza* and *Zonotrichia*. Recent *numt* insertions should exhibit a higher degree of identity to functional mitochondrial sequences and be longer. The former is due to the accumulation of point mutations while the latter is that *numt* length tends to decrease after insertion (Hazkani-Covo and Martin 2017) because small deletions outnumber small insertions (Johnson 2004). Some long (>1,000 bp) *numts* are present in either *Geospiza* or *Zonotrichia* but the shared insertions are shorter (fig. 3). The longest *numts* (>750 bp) unique to *Geospiza* exhibited a higher degree of identity to the *Geospiza* mitogenome (mean >91%) than the longest *numts* present in three or more taxa (mean <75%, see supplementary table S2, Supplementary Material online); the *Zonotrichia* mitogenome sequence is unavailable so we only conducted this analysis in *Geospiza*. Taken as a whole, these data suggest some recent *numt* insertions in *Geospiza* and *Zonotrichia* but the rate of *numt* accumulation was clearly higher in their common ancestor.

The *numt* presence/absence pattern for our five focal species appeared to be essentially homoplasy-free (fig. 1*B*), supporting our hypothesis that *numts* fit the RGC model. Indeed, the apparent absence of homoplasy seems surprising given extensive discordance among avian gene trees (e.g., Jarvis et al. 2014). Discordance among gene trees can drive the appearance of homoplasy even for characters that are actually homoplasy-free (supplementary text S1 and figs. S1 and S2, Supplementary Material online). We did find two *numts* with unclear insertion boundaries that appeared homoplastic (supplementary alignment, Supplementary Material online). However, even if we include those two loci the consistency index (Kluge and Farris 1969) was 0.999. Indeed, if we relax the no homoplasy assumption and estimate ancestral states using maximum likelihood, *numt* insertion rate estimates are essentially unchanged (supplementary fig. S3, Supplementary Material online).

What Drives *Numt* Insertions?

The transposition of mtDNA into the nuclear genome has occurred continuously over time (Mourier et al. 2001; Hazkani-Covo et al. 2010). Work in yeast suggests *numt* insertion reflects passive capture of mtDNA into nuclear double-stranded DNA breaks (DSBs) by the nonhomologous end joining (NHEJ) repair machinery (Ricchetti et al. 1999); this is probably the major *numt* insertion mechanism in many taxa (Hazkani-Covo and Covo 2008). *Numt* loci can reflect independent mtDNA insertions due to DSBs or postinsertion duplications (Bensasson et al. 2003; Hazkani-Covo et al. 2003). The de novo insertion hypothesis implies *numts* would correspond to many different mtDNA segments, as we found (fig. 2). However, a few mitogenomic regions appeared overrepresented as *numts* (supplementary fig. S4, Supplementary Material online). This could indicate a role for postinsertion duplications or hotspots for transfer. We identified some duplicated flanking sequences, as predicted by the postinsertion duplication model (supplementary fig. S5, Supplementary Material online), though we cannot rule out a contribution of hotspots. Regardless, de novo insertion appears to explain most of the *numts* in *Geospiza* and *Zonotrichia*.

What might have led to the burst of insertions in the ancestor of *Geospiza* and *Zonotrichia*? First, a substantial increase in the rate of DSBs in the ancestor could drive the observed pattern. The primary endogenous source of DSBs appears to be the blockage or pausing of replication forks (Mehta and Haber 2014). The basis of replication fork blockage is complex, and is likely to be very difficult to reconstruct for ancestral lineages like the *Geospiza-Zonotrichia* ancestor. Second, a mutation that reduced the fidelity of NHEJ during DSB repair could also increase the rate of *numt* insertion. Finally, increased leakage of DNA from the mitochondrion could also drive an increased *numt* insertion rate. Mutations that reduce the integrity of the mitochondrion and lead to an increased insertion of *numts* in somatic tissue are known (Srinivasainagendra et al. 2017); assuming this also occurs in the germline it would lead to an increased *numt* insertion rate. We examined the best-characterized gene that affects

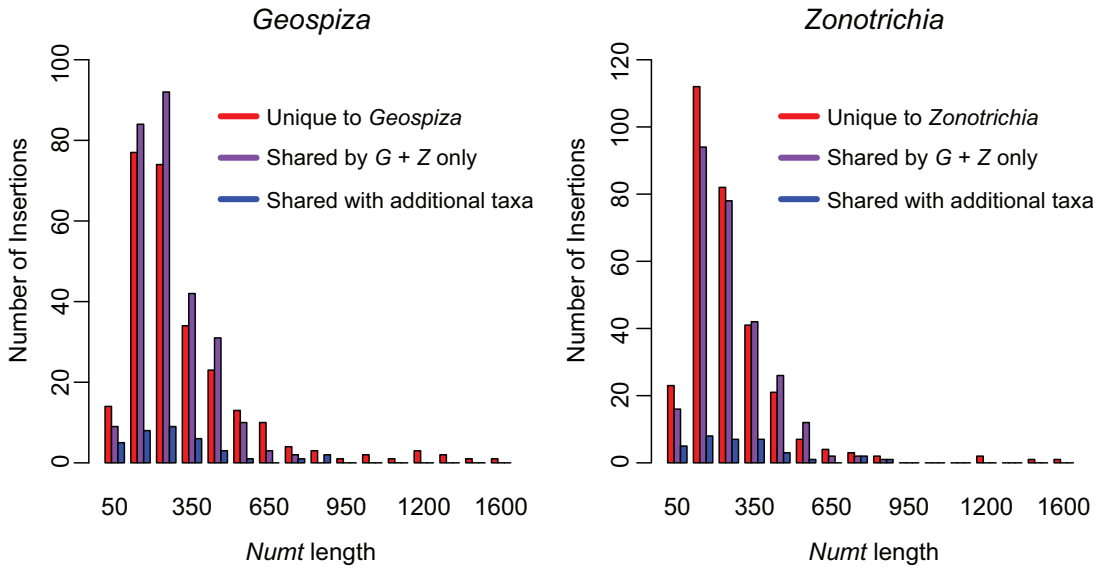

Fig. 3. Histogram of *numt* length distribution for *Geospiza fortis* and *Zonotrichia albicollis*. *Numts* are split into those unique to each taxon (the leftmost bars in the *numt* length categories), those only present in both *Geospiza* and *Zonotrichia* (the middle bars), and those shared by *Geospiza*, *Zonotrichia*, and at least one additional species (the rightmost bars).

mitochondrial leakage (i.e., orthologs of yeast *YME1*, Thorsness and Fox 1993) and were unable to find a clear pattern to explain our observations (supplementary text S2 and fig. S6, Supplementary Material online). The reduced insertion rate on the terminal branches (fig. 1*B*) suggests the process(es) that drove the burst are no longer active, making it difficult to test these hypotheses. However, these hypotheses could be tested in taxa with an ongoing burst of *numt* insertions (we expect such a taxon to have many polymorphic *numts*) since the numbers of DSBs and mitochondrial integrity could be examined in those taxa.

The Importance of *Numts* for Understanding Evolution

One significant result is the observation that *numt* insertions are essentially homoplasy free, making them virtually perfect RGCs. The DSB mechanism is consistent with *numt* insertions being perfect RGCs because the locations of DSBs should be random throughout the genome rather than occurring independently at the same site in different taxa. This is in contrast to better-studied RGCs, like transposable element insertions or microinversions, which appear to exhibit some true homoplasy (Braun et al. 2011; Han et al. 2011). Perfect RGCs define gene tree bipartitions accurately and so can provide accurate estimates of the amount of incomplete lineage sorting deep in the tree (Matzke et al. 2012; Jarvis et al. 2014; Suh et al. 2015). A large number of true gene tree partitions would provide a unique source of information about ancient population processes; *numts* are most useful for divergences that occurred during the burst of insertions and they should outperform other RGCs during a burst. We have only begun to survey avian genomes so it seems reasonable to expect additional *numt* insertion bursts. In fact, the other lineages with moderately large numbers of *numts* (fig. 1) could have relatives with a much larger *numt* repertoire; sequencing of those related taxa could reveal additional bursts of *numt* insertions.

Methods

We searched 64 avian genomes using related mitogenomes as BLASTN (Camacho et al. 2009) queries. We then focused on five passerine species where we conducted BLASTN searches with the flanks of the 1,031 *numts* identified in those taxa; the *numt* was scored as present (1), absent (0), or unknown (?) for each of the five species (957 loci had data for all taxa). We estimated the time tree with 100 intron loci from Jarvis et al. (2014) using RAxML (Stamatakis 2014) and treePL (Smith and O'Meara 2012) assuming *Corvus* and *Geospiza* diverged 20.526 Ma based on Jarvis et al. (2014). Estimates of *numt* insertion rates were obtained by dividing the number of insertions on each branch by the length of that branch. For additional details see supplementary methods and fig. S7, Supplementary Material online.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to the help from Joseph Brown in reconstructing the distribution plot. Stephen Smith provided useful comments on the first version of the manuscript. We thank the constructive comments from the three anonymous reviewers and editors. This study is supported by Hainan Key Research and Development Program (No. ZDYF2018143 to B.L.) and the National Natural Science Foundation of China (No. 31301894 to B.L. and No. 31360510 to N.W.). N.W. is also supported by NSF DEB AVATOL 1207915; R.T.K. and E.L.B. are supported by NSF DEB 1655683.

References

Adams KL, Qiu YL, Stoutemyer M, Palmer JD. 2002. Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc Natl Acad Sci U S A*. 99(15):9905–9912.

Bensasson D, Zhang DX, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol*. 16(6):314–321.

Bensasson D, Feldman MW, Petrov DA. 2003. Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J Mol Evol*. 57(3):343–354.

Bertheau C, Schuler H, Krumbock S, Arthofer W, Stauffer C. 2011. Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. *Mol Ecol Resour*. 11(6):1056–1059.

Braun EL, Kimball RT, Han KL, Iuhasz-Velez NR, Bonilla AJ, Chojnowski JL, Smith JV, Bowie RCK, Braun MJ, Hackett SJ, et al. 2011. Homoplastic microinversions and the avian tree of life. *BMC Evol Biol*. 11:141.

Brown JW, Wang N, Smith SA. 2017. The development of scientific consensus: analyzing conflict and concordance among avian phylogenies. *Mol Phylogenet Evol*. 116:69–77.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Gunbin K, Peshkin L, Popadin K, Annis S, Ackermann RR, Khrapko K. 2017. Integration of mtDNA pseudogenes into the nuclear genome coincides with speciation of the human genus. A hypothesis. *Mitochondrion* 34:20–23.

Han KL, Braun EL, Kimball RT, Reddy S, Bowie RCK, Braun MJ, Chojnowski JL, Hackett SJ, Harshman J, Huddleston CJ, et al. 2011. Are transposable element insertions homoplasy free? An examination using the avian tree of life. *Syst Biol*. 60(3):375–386.

Hazkani-Covo E. 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. *Mol Biol Evol*. 26(10):2175–2179.

Hazkani-Covo E, Covo S. 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genet*. 4:e1000237.

Hazkani-Covo E, Martin WF. 2017. Quantifying the number of independent organelle DNA insertions in genome evolution and human health. *Genome Biol Evol*. 9(5):1190–1203.

Hazkani-Covo E, Sorek R, Graur D. 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. *J Mol Evol*. 56:169–174.

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet*. 6:e1000834.

Hu G, Thilly WG. 1994. Evolutionary trail of the mitochondrial genome as based on human 16S rDNA pseudogenes. *Gene* 147:197–204.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215):1320–1331.

Johnson KP. 2004. Deletion bias in avian introns over evolutionary timescales. *Mol Biol Evol*. 21(3):599–602.

Kluge AG, Farris JS. 1969. Quantitative phyletics and evolution of anurans. *Syst Zool*. 18(1):1–32.

Leister D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends Genet*. 21(12):655–663.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. 1994. Numt, a recent transfer and tandem amplification of mitochondrial-DNA to the nuclear genome of the domestic cat. *J Mol Evol*. 39(2):174–190.

Matzke A, Churakov G, Berkes P, Arms EM, Kelsey D, Brosius J, Kriegs JO, Schmitz J. 2012. Retroposon insertion patterns of neoavian birds: strong evidence for an extensive incomplete lineage sorting era. *Mol Biol Evol*. 29:1497–1501.

Mehta A, Haber JE. 2014. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb Perspect Biol*. 6(9):a016428.

Mourier T, Hansen AJ, Willerslev E, Arctander P. 2001. The human genome project reveals a continuous, transfer of large mitochondrial fragments to the nucleus. *Mol Biol Evol*. 18(9):1833–1837.

Nacer DF, do Amaral FR. 2017. Striking pseudogenization in avian phylogenetics: numts are large and common in falcons. *Mol Phylogenet Evol*. 115:1–6.

Pereira SL, Baker AJ. 2004. Low number of mitochondrial pseudogenes in the chicken (*Gallus gallus*) nuclear genome: implications for molecular inference of population history and phylogenetics. *BMC Evol Biol*. 4:17.

Reddy S, Kimball RT, Pandey A, Hosner PA, Braun MJ, Hackett SJ, Han K-L, Harshman J, Huddleston CJ, Kingston S, et al. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Syst Biol*. 66(5):857–879.

Ricchetti M, Fairhead C, Dujon B. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402(6757):96–100.

Smith SA, O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28(20):2689–2690.

Srinivasainagendra V, Sandel MW, Singh B, Sundaresan A, Mooga VP, Bajpai P, Tiwari HK, Singh KK. 2017. Migration of mitochondrial DNA in the nuclear genome of colorectal adenocarcinoma. *Genome Med*. 9(1):31.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.

Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. *PLoS Biol*. 13(8):e1002224.

Thorsness PE, Fox TD. 1993. Nuclear mutations in *Saccharomyces cerevisiae* that affect the escape of DNA from mitochondria to the nucleus. *Genetics* 134(1):21–28.

Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 80(1):71–77.

Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, et al. 2014. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* 346(6215):1311–1320.

Zischler H, Geisert H, von Haeseler A, Pääbo S. 1995. A nuclear "fossil" of the mitochondrial D-loop and the origin of modern humans. *Nature* 378(6556):489–492.