Results and Discussion

Reannotation of gene orders

Extraordinarily long first control regions and specific gene orders in *Prioniturus luconensis*, *Forpus passerines*, *Psittacus erithacus* and *Poicephalus gulielmi* prompted us to search for potentially hidden pseudogenes in these regions. The originally annotated gene rearrangement in *Prioniturus luconensis* (Eberhard and Wright 2016), looked like a degenerated variant of the fully duplicated gene order in Nestoridae and Cacatuidae (GO-FD in Figure 1D) but with a pseudogenization of the first *ND6* and *tRNA-Glu* copies as well as the second *tRNA-Thr* and *tRNA-Pro* copies, and the loss of the second copy of cytochrome b. The absence of this copy and much longer CR1 (1321 bp) than the typical paralogous CR2 (1237 bp) strongly suggested that a hidden *cytb* pseudogene can exist within the 3' end of CR1. In fact, the comparison of *Prioniturus luconensis* CR1 sequence with the functional *cytb* sequence of this species revealed a very significant similarity (E-value = 0; 87.4% identity) between these sequences along the 143-bp alignment (Figure S3A), which supports the presence of the *cytb* pseudogene in the *Prioniturus* mitogenome in the position between 17,657 and 17,801 bp. After reannotation of this pseudogene, the length of CR1 reduced to 1,137 bp. The newly annotated *Prioniturus* gene order was defined as GO-2 in Figure 1D.

Similarly, we also found *cytb* pseudogenes within the last 178 nucleotides of *Psittacus erithacus* CR1 (Figure S3D) and the last 170 nucleotides of *Poicephalus gulielmi* CR1 (Figure S3E), with E-value = 7.9E-9 and 1.7E-6, and the percent identity of 63.3% and 61.9%, respectively. In the case of *Psittacus erithacus*, we also noticed some traits of the *ND6* pseudogene within the first 68 nucleotides of CR1 with E-value=0.024 and 64.6% identity (Figure S3C). Thereby, these closely related taxa belonging to the Psittacinae subfamily share the same gene order ascribed as GO-3 in Figure 1D. Thorough homology analyses of *Forpus passerines*, CR1 revealed also the presence of the *ND6* pseudogene within the first 123 nucleotides (Figure S3B). The similarity between the *ND6* gene and its pseudogene is very significant with E-value=9.6E-32 and 76.4% identity. This gene order is very similar to that in other members of Arinae, *Amazona* and *Pionus* (GO-4 in Figure 1D) and was defined as GO-5 (Figure 1D).

Interestingly, *Neophema chrysogaster* is the only parrot species that contains one control region (Miller et al. 2013) but has two identical poly-C sequences within. One of them is localized near the 5' end (positions 73 - 87 bp) of CR and the second is in the central part of CR (positions 1078 - 1092 bp). Since the poly-C sequence is commonly considered the beginning of control regions, we decided to analyse the corresponding sequences that precede and follow the two poly-C motifs. The second poly-C sequence turned out to be the beginning of the second complete control

region (CR2) located from 1028 to 2504 bp, while the first poly-C sequence starts the first partial control region (CR1) from 13 to 631 bp (Figure S4A). A comparison of the two control regions (Figure S4C) showed that CR1 is deprived of a 826-bp sequence, which constitutes the central part of CR2. The remaining parts of the first control region are identical in 96.1% to the second control region (Figure S4C). Since the mitogenome of *Neophema chrysogaster* was assembled *de novo* and both control regions differ in their surroundings, we believe that these regions are actually present in this mitogenome. Thereby, the *Neophema* mitogenome can be the first example with a remnant variant of CR1, as so far only vestigial CR2s have been reported so far. Moreover, detailed analyses of the sequence between these two *Neophema* control regions revealed that it shows a significant similarity to *cytb* gene with E-value = 6.2E-6 and 67.5% identity on the 351-bp alignment (Figure S4C). This region could be subjected to a gene conversion because it is composed of seven parts, which show mutual similarity (Figure S4A and S4B). Because of its uniqueness, we marked the gene order in *Neophema chrysogaster* as GO-7 (Figure 1D).

As a consequence of these reannotations, the first control region became shorter in the mitogenomes of *Prioniturus luconensis*, *Forpus passerines*, *Psittacus erithacus* and *Poicephalus gulielmi*. These results suggest that more degenerated variants of mitochondrial gene order can be present also in other taxa belonging to the Psittacoidea superfamily.

Comparison of length and general structure of duplicated control regions

Sequencing of seven mitogenomes with duplication and reannotation of five others increased the number of such genomes to 22 and enabled their comparison to draw some general conclusions about control regions (Table S2). All control regions in parrots have been identified with a short spacer followed by a poly-C sequence at the 5' end. The poly-C motifs contain T, TA, TT or TTA in the middle, which are flanked by six to nine cytosine tracks. The 5' track can be one cytosine longer (12 cases) or two Cs longer as in *Calyptorhynchus baudinii*. Two paralogous motifs in a given genomes always include the same internal nucleotide(s) but the number of adjacent Cs may vary. The 5' spacer can be from 5 bp to 69 bp long and its median length is 15 bp. The paralogous spacers in the same genome can have the same or different length without any clear bias toward the first or the second copy. However, we found that the second control regions are longer than the first ones in 16 out of the 21 cases, after exclusion of *Neophema chrysogaster* because of the remnant CR1. The greater length of CR2 results from longer microsatellites at the 3' end in five *Amazona* species (*aestiva*, *auropalliata*, *barbadensis*, *ochrocephala* and *oratrix*). In *Forpus passerinus*, the CR1 is longer than CR2 because of elongated microsatellites.

To systematically study if the difference in length is a rule in other groups of birds, we compared the pairs of duplicated control regions in all available avian mitogenomes. We included only the pairs of control regions in which none of them was remnant (Table S3). We observed a strong bias in the length of these regions. Out of 121 pairs of such regions, the CR2 was longer in 109 cases (p=9.2E-21). Besides Psittaciformes (p=0.0259), the same statistically significant trend was also observed for three other avian orders with at least 13 annotated pairs of control regions. CR2 was longer than CR1 in all mitogenomes of Gruiformes (p=0.00021) and Pelecaniformes (p= 1.6E-6), and in almost 98% of Passeriformes mitogenomes (p=3.0E-13) (Table S3). The median lengths of CR1 and CR2 for the whole set were 1136 bp and 1265 bp, respectively, and the difference between them was statistically significant with p=3.2E-14. CR2 was also statistically longer than CR1 in individual avian orders: Gruiformes (p=0.0017), Passeriformes (p=1.1E-8), Pelecaniformes (p=1.9E-6) and Psittaciformes (p=0.046) (Table S3). The median difference between the length of CR2 and CR1 was from 57 bp (Psittaciformes) to 356 bp (Pelecaniformes) for avian orders with at least 5 annotated pairs of control regions. We also noticed significant differences in the length of control regions compared between orders of birds (Table S4). The mitogenomes of Gruiformes showed generally the shortest CR1 and Psittaciformes the longest, while CR2 had the shortest length usually for representatives of Procellariiformes and was the longest for Pelecaniformes.

Materials and Methods

Phylogenetic analyses

To infer phylogenetic relationships between as many parrot species as possible, we used the nucleotide sequences of NADH dehydrogenase subunit 2 (*ND2*) gene and to maximize the number of informative sites, we included all genes from complete mitochondrial genomes. The newly obtained sequences were compared with all available parrot homologous sequences available in the GenBank database (Table S7). After removing incomplete and fragmentary sequences, the final set of *ND2* included 245 sequences, while the genomic set consisted of 13 protein coding genes, 12S and 16S rRNA, as well as 22 tRNA sequences coming from 48 parrot species and five representatives of Passeriformes used as an outgroup. To obtain phylogenetic relationships between control region (CR) sequences, we analysed two sets of these regions. The first one included all available 67 sequences of CR and the second 18 sequences from *Amazona* and *Pionus*. The sequences were aligned in MAFFT using a slow and accurate algorithm L-INS-i with 1,000 cycles of iterative refinement (Katoh and Standley 2013). The alignments were then edited manually in JalView (Waterhouse et al. 2009) and sites suitable for phylogenetic study were selected in

GBlocks (Talavera and Castresana 2007). The *ND2* alignment was 1041 bp long, while the concatenated alignment of mitochondrial genes included 15,249 bp.

We applied three phylogenetic approaches to infer evolutionary relationships between parrots: the maximum likelihood method in IQ-TREE (Nguyen et al. 2015), as well as two Bayesian analyses in MrBayes (Ronquist et al. 2012) and PhyloBayes (Lartillot and Philippe 2004). In the case of *ND2* set, we checked the necessity of using separate nucleotide substitution models for three codon positions, while for the concatenated alignments of mitochondrial genes, we considered 63 potential partitions: three codon positions for each individual protein coding gene and separate partitions for each of the RNA genes (Table S8).

The ModelFinder program associated with IQ-TREE (Chernomor et al. 2016; Kalyaanamoorthy et al. 2017), provided one substitution model GTR+R7 for all three partitions of the *ND2* alignment and five substitution models for 37 mitochondrial markers (Table S8). In the case of CR alignments, the program proposed HKY+I+Γ4 for the first set and HKY+R3 for the second one. In IQ-TREE, we applied Shimodara-Hasegawa-like approximate likelihood ratio test (SH-aLRT) assuming 10,000 replicates and non-parametric bootstrap with 1,000 replicates.

In MrBayes analyses, we assumed separate substitution models for the three codon positions of the *ND2* alignment and 12 models for the appropriate partitions of the mtDNA genes according to the results of PartitionFinder (Lanfear et al. 2012) assuming BIC criterion for the model selection (Table S8). We applied mixed models rather than fixed ones to specify appropriate substitution models across the large parameter space (Huelsenbeck et al. 2004), but the models describing heterogeneity rate across sites were adopted according to PartitionFinder. In the case of CR sets, we assumed the mixed models and the heterogeneity rate across sites described by invariant and discrete gamma models according to results of jModelTest 2.1 (Darriba et al. 2012). Two independent runs starting from random trees, each using four Markov chains, were applied. The trees were sampled every 100 generations for 10,000,000 generations for *ND2* and 20,000,000 for the mtDNA genes and the CR sets. In the final analysis, we selected the last 3,723,000 to 12,608,000 (depending on the alignment set) trees that reached the stationary phase and convergence, *i.e.* when the standard deviation of split frequencies stabilized and was lower than 0.002.

In PhyloBayes, we applied the GTR+Γ model for the first CR set and CAT+GTR+Γ model for the second CR set as well as for the *ND2* and mtDNA genes. The number of components, weights and profiles of the applied models were inferred from the data. Two independent Markov chains were run for 100,000 generations with one tree sampled for each generation. The last 75,000 to 95,000 trees (depending on the alignment set) from each chain were collected to compute

posterior consensus trees after reaching convergence, when the largest discrepancy observed across all bipartitions (maxdiff) was much below the recommended threshold 0.1. The gamma-distributed rate variation across the sites was approximated by five discrete rate categories in Bayesian approaches.

The data about the presence and absence of duplication in the parrot mitogenomes were collected based on PCR screening obtained by Schirtzinger et al. (2012) and ourselves as well as mitochondrial genomes deposited in GenBank and obtained in this study (Table S7). The data was mapped on the IQ-TREE maximum likelihood tree obtained for *ND2* gene and mtDNA markers using Mesquite (Maddison and Maddison 2017). The analyses were carried out for two *ND2* data sets. One consisted of 238 parrot species with both known and unknown information about the mitogenome duplications. The lack of data about the duplication was coded as missing data. The second set included 141 taxa with the confirmed presence or absence of duplication. We applied maximum parsimony and maximum likelihood reconstruction methods. In the latter case, we used Mk1 model (Markov k-state 1 parameter model) because it fit the data better according to AIC criterion than the alternative AsymmMk model (asymmetrical Markov k-state 2 parameter model).

References

Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst Biol 65:997-1008.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772.

Eberhard JR, Wright TF. 2016. Rearrangement and evolution of mitochondrial genomes in parrots. Mol Phylogenet Evol 94:34-46.

Huelsenbeck JP, Larget B, Alfaro ME. 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol Biol Evol 21:1123-1133.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780.

Lanfear R, Calcott B, Ho SY, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695-1701.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095-1109.

Maddison WP, Maddison DR. 2017. Mesquite: a modular system for evolutionary analysis. Version 3.31.

Miller AD, Good RT, Coleman RA, Lancaster ML, Weeks AR. 2013. Microsatellite loci and the complete mitochondrial DNA sequence characterized through next generation sequencing and de novo genome assembly for the critically endangered orange-bellied parrot, Neophema chrysogaster. Mol Biol Rep 40:35-42.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268-274.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539-542.

Schirtzinger EE, Tavares ES, Gonzales LA, Eberhard JR, Miyaki CY, Sanchez JJ, Hernandez A, Mueller H, Graves GR, Fleischer RC, et al. 2012. Multiple independent origins of mitochondrial control region duplications in the order Psittaciformes. Mol Phylogenet Evol 64:342-356.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564-577.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189-1191.