CHD genes: a reliable marker for bird populations and phylogenetic analysis?
Case study of the superfamily Sylvioidea (Aves: Passeriformes)

Mitică CIORPĂC1,2, Radu Constantin DRUICĂ1, Gogu GHIORGHIȚĂ1, Dumitru COJOCAȚĂ1, Dragos Lucian GORGAN1,*
1Department of Biology, Faculty of Biology, “Alexandru Ioan Cuza” University of Iasi, Iasi, Romania
2Interdisciplinary Research Department - Field Science, “Alexandru Ioan Cuza” University of Iasi, Iasi, Romania

Received: 07.10.2015 ● Accepted/Published Online: 09.04.2016 ● Final Version: 24.10.2016

Abstract: The chromo-helicase-DNA binding protein (CHD) genes are widely used markers for sex determination in birds and provide a rapid and low-cost method with applicability to a large number of taxa. A good phylogenetic marker displays highly conserved domains and a slow evolution rate, properties that CHD genes seem to have. This gives rise to the following question: is the CHD gene a reliable marker for phylogenetic and bird population analysis? The aim of this study is to investigate whether the CHD gene is a reliable tool for molecular phylogeny and for population analysis within the superfamily Sylvioidea. Our results reveal that CHD genes are good markers for these two analyses, even better than myoglobin.

Key words: CHD genes, cytochrome b, myoglobin, Sylvioidea

Sex identification in birds based on their external morphology is difficult (Griffiths et al., 1998), as they are mostly sexually monomorphic (Ong and Vellayan, 2008). With the discovery of the chromo-helicase-DNA binding protein (CHD) gene (Griffiths and Tlwarl, 1995) in the avian sex chromosome (Ellegren, 1996), molecular DNA noninvasive sexing methods, such as the analysis of feathers, became possible and generated applicability to wildlife DNA forensics (An et al., 2007). The avian CHD1 genes belong to a family composed of a chromatin organization modifier (An et al., 2007) domain, an SNF2-related helicase/ATPase domain, and a DNA binding domain; thus, the acronym CHD stands for these (Fridolfsson and Ellegren, 1999). CHD genes have an important role in the avian genome, as in other organisms, due to their involvement in chromatin remodeling in the control of transcription elongation (Simic et al., 2003). This gene has two introns with different lengths for the Z and W chromosome, allowing the discrimination between the amplicons of the Z and W chromosomes by gel electrophoresis (Dubiec and Zagalska-Neubauer, 2006). Being a functional part of the DNA and with a very slow evolution rate, the CHD gene is highly conserved, even among distant species. The alignment between the CHD-W sequence in birds and the CHD sequence in mice does not include any gaps, except for two 130- and 175-bp deletions (Vucicevic et al., 2013). Its high degree of conservation across species has led to the design of universal primers for birds’ sex determination (Griffiths et al., 1998; Kahn et al., 1998; Fridolfsson and Ellegren, 1999).

Genetic distances are a frequently used tool to identify and assess the species status of closely related taxa (Wesson et al., 1993; Hung et al., 1999; Burbink et al., 2000; Bradley and Baker, 2001; Cagnon et al., 2004; Parkin et al., 2004; Olsson et al., 2005; Newman et al., 2012). This way, genetic distances are frequently compared between different studies, even if different genetic markers are involved (Helbig et al., 1995; Baker et al., 2003; Johnson and Cicero, 2004; Zhang et al., 2007). Fast-evolving parts, like mitochondrial sequences (Heidrich et al., 1998) or nuclear introns, are suitable for resolving young evolutionary relationships, for example those between species, whereas older relationships are better analyzed with more conservative genes like nuclear exons (Lin and Danforth, 2004). Mostly, a combination of the two is used to target different parts of a phylogenetic tree.

For the superfamily Sylvioidea, in addition to mitochondrial genes such as cytochrome b, control region, or ND2 (NADH-ubiquinone oxidoreductase chain 2), a number of nuclear genes are used for molecular phylogeny, genes like fibrinogen beta chain intron 5 (FGB), glyceraldehyde-3-phosphate dehydrogenase (GAPDH), lactate-dehydrogenase B intron 3 (LDHB), ornithine-decarboxylase exon 6–8, intron 7 (ODC1), recombination...
activation gene 1 (RAG1), and the widely used myoglobin intron 2 (myo) (Fregin et al., 2009). Considering this, the aim of this study is to investigate whether the CHD gene is a reliable tool for molecular phylogeny and population analysis of the superfamily Sylvioidea.

Blood samples from three species of the genus *Acrocephalus*, *A. schoenobaenus*, *A. arundinaceus*, and *A. scirpaceus*, were collected from five different bird ringing camps in Larga Jijia (Iasi County, Romania) during 2010 and stored in Queen's Lysis Buffer (Seutin et al., 1991). The genomic DNA was isolated and purified using the DNA IQ System (Promega, Madison, WI, USA), and quantified through spectrometry and electrophoresis. The CHD genes were amplified using PCR methods with one pair of specific primers: P2 5'-TCT GCA TCG CTA AAT CCT TT-3' (Griffiths and Tiwarl, 1995) and P8 5'-CTC CCA AGG ATG AGR AAY TG-3' (Griffiths et al., 1998). The PCR reaction was carried out in 25 µL of total volume containing GoTaq Green Master Mix (Promega), primers (0.2 µM final concentration), DNA template (~12.5 µg), and nuclease-free water up to the final volume. The amplification was conducted in a SensoQuest Labcycler (SensoQuest GmbH, Göttingen, Germany) under the following cycling conditions: 1.5 min of initial denaturation at 95 °C, followed by 30 cycles of denaturation at 95 °C for 30 s, primer alignment at 48 °C for 45 s, and elongation at 72 °C for 45 s, followed by a final elongation step at 72 °C for 5 min, in accordance with Griffiths et al. (1998). A 3% agarose gel was run for the PCR products and visualized in UV light. The distinction between males and females was made by the presence of one band for males and two bands for females.

A total of 21 sequences belonging to 7 taxa, with 6 from the superfamily Sylvioidea (Table 1), were used to assess the reliability of the CHD-Z gene in phylogenetic analysis. The cytochrome b (cytb [1012 bp; 337 variable sites, of which 176 were informative]), myoglobin intron 2 (myo [656 bp; 207 variable sites without gaps, of which 176 were informative]), and chromo-helicase-DNA-binding protein gene (CHD-Z exon and intron [326 bp; 62 variable sites without gaps, of which just 13 were informative]) sequences were aligned and concatenated using MEGA 6 (Tamura et al., 2013), defining 3 data sets for the same taxa: 1) cytb sequences; 2) cytb and CHD-Z sequences concatenated; 3) cytb and myo sequences concatenated.

The corrected HKY (Hasegawa et al., 1985) distance between taxa for each gene was obtained using PAUP v4.0b10 (http://paup.sc.fsu.edu/). The optimal substitution model was selected according to the Akaike information criterion (Akaikes, 1974) using three substitution schemes (+F; +I; +G, 4 nCat) and 8 different topologies in jModelTest v2.1 (Guindon and Gascuel, 2003; Darriba et al., 2012). The base composition heterogeneity and chi-square ($\chi^2$) test of deviation from base composition homogeneity of variable sites were calculated in PAUP v4.0b10 (http://paup.sc.fsu.edu/). Furthermore, the relative composition variability (RCV) was computed as a summary statistic for each marker, following the method of Phillips and Penny (2003).

The maximum likelihood (Baker et al., 2003) tree was designed in SeaView (Gouy et al., 2010) using PhyML (Guindon et al., 2010) and was visualized in DensiTree v2.1.11 (Bouckaert, 2010). Bayesian inference (BI) analysis was conducted in BEAST v1.8.0 [Bayesian Evolutionary

### Table 1. Phylogenetic analysis data set.

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Family</th>
<th>GenBank Acc. No.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>cytb</td>
<td>CHD-Z</td>
</tr>
<tr>
<td><em>Acrocephalus_sechellensis</em></td>
<td>Acrocephalidae</td>
<td>AJ004284 1</td>
</tr>
<tr>
<td><em>Phylloscopus_collybita</em></td>
<td>Phylloscopidae</td>
<td>HQ608821 4</td>
</tr>
<tr>
<td><em>Phylloscopus_trochilus</em></td>
<td>Phylloscopidae</td>
<td>Z73492 7</td>
</tr>
<tr>
<td><em>Pycnonotus_sinensis</em></td>
<td>Pycnonotidae</td>
<td>FJ487714 17</td>
</tr>
<tr>
<td><em>Pycnonotus_taivanus</em></td>
<td>Pycnonotidae</td>
<td>NC013483 9</td>
</tr>
<tr>
<td><em>Parus_atricapillus</em></td>
<td>Paridae</td>
<td>AF347937 10</td>
</tr>
<tr>
<td><em>Gallus_gallus</em></td>
<td>out-group</td>
<td>AP003319 13</td>
</tr>
</tbody>
</table>

---

1Heidrich et al. (1998); 2Dawson et al. (2005); 3Fregin et al. (2009); 4Lei et al. (2010); 5Bensch et al. (2006); 6Fuchs et al. (2006); 7Helbig et al. (1995); 8Zuccon and Ericson (2010); 9Chang et al. (2010); 10Gill et al. (2005); 11Harvey et al. (2006); 12Johansson et al. (2013); 13Nishibori et al. (2005); 14Griffiths et al. (1998); 15Boardman et al. (2002); 16Alstrom et al. (2006); 17Lohman et al. (2010); 18Chang et al. (2008).
Analysis Sampling Trees (Drummond et al., 2012) using the substitution model selected by the AIC for each gene and a Relaxed Clock: Uncorrelated Log-normal (Drummond et al., 2006) molecular clock model with a tree prior setup to the Speciation: Yule Process (Yule, 1925; Gernhard, 2008) model, applied for a UPGMA starting tree. One independent Markov chain Monte Carlo simulation of 10 million iterations for each data set was run through CIPRES Science Gateway v.3.3 (Miller et al., 2010). Convergence for the posterior distributions of the parameter estimates was checked using Tracer v1.5 (http://tree.bio.ed.ac.uk/software/tracer/), and then the summary maximum clade credibility (MCC) tree was computed using TreeAnnotator. The MCC trees were visualized and graphically edited in FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/). Finally, the posterior marginal likelihoods, treeness index (Phillips and Penny, 2003), Colless tree imbalance (Colless, 1982), and mean posterior probability for each tree were used for model comparisons and identification of the best tree model fit.

The CHD gene is an easy and reliable tool in bird sex determination, defining the differences even from the size of the PCR products. This gene presents two alleles of different sizes; the allele from the W chromosome is bigger than that of the Z chromosome, allowing for discrimination between genera. Sex determination using the CHD gene was conducted for three species of Acrocephalus: A. schoenobaenus, A. arundinaceus, and A. scirpaceus. For A. schoenobaenus, a 166.66% sex ratio from a sample size of 32 was identified, which suggests sex disequilibrium within the population, as the male percentage is 62.50%. This sex ratio trend was constant during the sampling dates (Figure 1). A 42.85% sex ratio was observed for A. arundinaceus in a sample size of 20 individuals, with 30% males and a constant trend during the sampling campaign. A similar situation was observed for A. scirpaceus, with a higher sex disequilibrium and a sex ratio of 311.11% in a sample size of 37 individuals with 75.68% males.

Taking into account the possibility of hybridization between A. arundinaceus and A. scirpaceus, identified across Europe, the correlation degree between these two higher sex disequilibria was tested based on a correlation test that showed a strong positive level of correlation (0.81) for the males and females of both species and a negative correlation between the males and females of different species.

In the world of birds, quite often both sexes have similar phenotypic traits, and even an experienced ornithologist may have difficulty in identifying an ambiguous individual (Dubiec and Zagalska-Neubauer, 2006). Other characteristics, such as sexually monomorphic colors for around 60% of all songbird species (Price and Birch, 1996) and the morphological uniformity of nestlings, increase the difficulty of taxonomical classification. Therefore, the utility of the CHD-based technique in molecular sexing of bird species is unquestionable, having been proven by numerous studies. Furthermore, it is a fast and accurate tool for sex determination.
method, but at the same time easy to perform and relatively cheap. Furthermore, numerous bird species or populations are protected by different conservation programs and the sampling process must be as noninvasive as possible. Previously, Ong and Vellayan (2008) showed the applicability of CHD-based molecular sexing as a completely noninvasive method, using feathers as a source of DNA.

The analysis of the sequence divergences for 7 taxa representing 4 families of the superfamily Sylvioidea, plus an outgroup, *Gallus gallus*, indicates that the HKY distance for CHD-Z was between 0.000 and 0.290 (data not shown). Divergences between cytochrome b (cytb) and myoglobin, intron 2 (myo), were higher than those of the CHD-Z gene. A graphical representation of the pairwise corrected sequence divergences (HKY distance) for CHD-Z, cytb, and myo is shown in Figure 2 and indicates a linear relation among these three genes. The cytb sequence generally diverges much faster than CHD-Z (the ratio of cytb/CHD-Z is 1.41); the myoglobin gene also diverges faster than CHD-Z (the ratio of myo/CHD-Z is 1.15) for the same set of taxa. Considering just the ingroup taxa (the taxa from superfamily Sylvioidea), the myoglobin and CHD-Z genes have the same percentage of divergence (0.086%), much lower than cytb, which presents a value twice as large (0.178).

Comparing the sequence divergence and the current classification within the superfamily Sylvioidea, the values are closely related for each taxonomic position. The species from the same genus revealed a divergence between 0.000 and 0.004 (genera *Phylloscopus* and *Pycnonotus*) for the CHD-Z gene compared to 0.003–0.102 for the cytb gene and 0.007–0.028 for the myo gene. CHD-Z interfamiliar distances showed a minimum value for Acrocephalidae/Phyllophaginae (0.066) and a maximum value for Acrocephalidae/Paridae (0.086). Myoglobin showed closer values: a maximum of 0.086 for the same families, but a minimum between Pycnonotidae/Phyllophaginae (0.051). This last value is rather unexpected given that these two families are taxonomically located in different clades. Divergence between taxa based on cytb gene analysis shows high values that are uniformly distributed across taxa and comparable to those of the CHD-Z and myo genes.

The evolution time of the CHD-Z gene indicates a higher substitution rate for transitions (2.536) and a lower rate for transversions (0.511–1.000) (Table 2). The optimal substitution model for CHD-Z gene evolution, selected by the AIC algorithm in jModelTest (Guindon and Gascuel, 2003; Darriba et al., 2012), was TPM1uf + G (Kimura, 1981), with 1.304 shape for 4 gamma categories.

The chi-square test of homogeneity of base frequencies across taxa was nearly significant ($\chi^2 = 23.829, P = 0.068$) for CHD-Z marker across 39 variable sites within the ingroup. Meanwhile, the chi-square test was clearly statistically insignificant for the myo and cytb markers ($P = 0.384$ for myo and $P = 0.929$ for cytb), suggesting a more stable base composition among taxa. The base composition heterogeneity (RCV) confirms the chi-square test of homogeneity, showing low levels of variability for the widely used markers cytb (RCV = 0.018) and myo (RCV = 0.025) and a higher value for the new proposed marker CHD-Z (RCV = 0.046). Phillips and Penny (2003) showed that nucleotide homogeneity within the ingroup and countable outgroup heterogeneity can equally affect the root position and the tree topology. The high nucleotide homogeneity and the
higher distance within individuals from four different families, observed for the cytb and myo markers, could explain the presence of uncertain relationships within the families of Sylvioidea. In contrast, the CHD-Z marker presents higher nucleotide heterogeneity, determined by the intron region, and a much lower distance within individuals, mainly determined by the exon region.

Knowing the implications of molecular marker selection for phylogenetic analysis and the risk of obtaining different topologies using an improper marker, the topology given by the CHD-Z gene was tested along with those of the cytb and myo genes, the most common markers used in bird phylogeny. The ML tree of a single nuclear gene reveals a wired topology for both analyses, CHD-Z and myo (Figure 3), probably caused by an improper tree algorithm (ML usually gives good topology when the analyzed sequence has a high number of nucleotides) or by a short distance between all studied taxa. Therefore, the phylogeny of a single nuclear gene, in our case CHD-Z or myo, is insufficient and unreliable.

The mitochondrial sequence is a reliable tool for phylogenetic relationships between supraspecific taxa, but usually it is not a good indicator of intra- or interspecific relationships. Over time, combined data sets and nuclear and mitochondrial markers have proven their usefulness in phylogeny inference by increasing the resolution. For example, Pons et al. (2011) drew the green woodpecker's phylogeography using mitochondrial marker cytb and another four nuclear markers (one of them being Z-linked), showing a clear distinction between the Iberian and European subspecies of green woodpecker, soon confirmed by another study by Perktas et al. (2011) that used a different mitochondrial marker over an increased number of individuals. Therefore, combined mitochondrial and nuclear data sets are mandatory in order to eliminate any taxonomic uncertainties. Combining the cytochrome b sequence with the CHD-Z

<table>
<thead>
<tr>
<th>Gene</th>
<th>AIC model selected</th>
<th>R(AC)</th>
<th>R(AG)</th>
<th>R(AT)</th>
<th>R(CG)</th>
<th>R(CT)</th>
<th>R(GT)</th>
<th>ti/tv</th>
<th>gamma</th>
<th>p-inv</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD-Z</td>
<td>TPM1uf+G</td>
<td>1.000</td>
<td>2.536</td>
<td>0.511</td>
<td>0.511</td>
<td>2.536</td>
<td>1.000</td>
<td>-</td>
<td>1.304</td>
<td>-</td>
</tr>
<tr>
<td>cytb</td>
<td>TVM+G</td>
<td>3.887</td>
<td>10.131</td>
<td>1.761</td>
<td>0.459</td>
<td>10.131</td>
<td>1.000</td>
<td>-</td>
<td>0.244</td>
<td>-</td>
</tr>
<tr>
<td>myo</td>
<td>HKY</td>
<td>1.914</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.914</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 3. The ML tree for myoglobin gene (in blue) and for CHD-Z gene (in red); both trees are scaled in substitutions and the length of each tree and different topologies indicate that both of them are improper for phylogeny using a single gene.
gene and myoglobin, two data sets were obtained for a complete BI phylogenetic analysis to test the reliability of the CHD-Z gene in phylogeny. A topology comparison of the MCC trees with cytochrome b highlights that all the MCC trees have a similar topology, with differences given by the total length of the branches (Figure 4), determined by the mutation rates specific for each gene. All these three MCC trees are scaled in substitutions and vary between 0.322 for cytb and 0.175 for myo and cytb.

The BI analysis showed a clock rate for CHD-Z-cytb genes (1.457), intermediary between myoglobin - cytb (0.974) and the cytb gene (1.772) independently analyzed, meaning that the myo gene is changing faster through time and can negatively affect the time-calibrated trees, giving wrong divergence species time. A marginal likelihood estimation was performed for the tree models comparison by the AIC through Markov chain Monte Carlo (AICM) algorithm, inferred using the method-of-moments estimator (Baele et al., 2012) as a measure of accuracy of the molecular clock models. The minimum value of the AICM, corresponding to the combined data set cytb + CHD-Z, indicates the better model fit. The analyzed models’ accuracy is expressed as the difference between AICM values, which means that the positive values (+2483.463) indicate a better model relative fit of the cytb + CHD-Z tree compared to the cytb + myo tree (Table 3). Furthermore, the AICM values are confirmed by the tree mean posterior probability (cytb + CHD-Z 0.901, S.E. = 0.078; cytb 0.882, S.E. = 0.0084; and cytb + myo 0.835, S.E. = 0.119), suggesting a lower level of phylogenetic noise. Furthermore, the tree topology asymmetry was evaluated in order to assess the combined data set constraining the species’ different potential for speciation (Blum and Francois, 2005). The Colless index of tree imbalance (Colless, 1982) seems to be unaffected by data sets, with a value of 0.466. Li et al. (2007) described a good phylogenetic marker as a gene with an intermediate substitution rate and high treeness values, highlighting the importance of the optimal rate and base composition stationarity in marker suitability. Even if the CHD-Z base composition is less stationary than myo or cytb, the low number of variable sites with high base composition heterogeneity seems to increase the fitness of

Figure 4. Maximum clade credibility trees for three data sets obtained by BI analysis in BEAST v1.8.0. MCC genealogy based on a Yule speciation process, summarized from the output of a Markov chain Monte Carlo chain run for 10 million iterations and sampled every 1000 iterations.
this phylogenetic tool. Overall, our new proposed marker CHD-Z has the greatest phylogenetic signal-to-noise index (treeness index: 0.371), and hence a lower potential for nonphylogenetic signals to influence phylogeny reconstruction compared to myo (treeness 0.357), which is diverging faster than CHD-Z (the ratio of myo/CHD-Z is 1.15). Due to the coding and noncoding regions, the CHD-Z marker is able to discriminate both close and distant sylvioid taxa, respectively.

In conclusion, the implications of the CHD gene for population analysis are substantial, offering a good image of population dynamics, sex distribution, and population trends. Additionally, the use of the CHD-Z gene in phylogenetic reconstruction gives the most effective results, having a moderate clock rate and being capable of revealing the relationship between closer and distant taxa. The CHD genes are reliable tools not just for sex determination in birds, but also for population and phylogenetic analysis, providing better results compared to the myoglobin gene, a frequently used nuclear marker in bird analysis.

Acknowledgment
This work was supported by the project “Doctoral and Post-doctoral programs of excellence for highly qualified human resources training for research in the field of Life sciences, Environment and Earth Science”, Grant No. POSDRU/159/1.5/S/133391, cofinanced by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007–2013, University of Bucharest, Romania.

References


Olsson U, Alström P, Ericson PG, Sundberg P (2005). Non-mono-
phyletic taxa and cryptic species - Evidence from a molecular
phylogeny of leaf-warblers (Phylloscopus, Aves). Mol Phylo-
genet Evol 36: 261-276.

sets for sex typing of birds from feathers. Zoo Biology 27: 62-
69.

Parkin DT, Collinson JM, Helbig AJ, Knox AG, Sangster G, Svensson
L (2004). Species limits in Acrocephalus and Hippolais warblers
from the Western Palearctic. Brit Birds 97: 276-299.

Perktas U, Barrowclough GF, Groth JG (2011). Phylogeography and
species limits in the green woodpecker complex (Aves: Pici-
dae): multiple Pleistocene refugia and range expansion across

Phillips MJ, Penny D (2003). The root of the mammalian tree in-
ferred from whole mitochondrial genomes. Mol Phylogenet

the Eurasian green woodpecker (Picus viridis). J Biogeogr 38:
311-325.

Price T, Birch GL (1996). Repeated evolution of sexual color dimor-
phism in passerine birds. The Auk 842-848.


Simic R, Lindstrom DL, Tran HG, Roinick KL, Costa PJ, Johnson
protein CHD1 interacts with transcription elongation factors
and localizes to transcribed genes. EMBO J 22: 1846-1856.

MEGA6: Molecular Evolutionary Genetics Analysis version

Vucicevic M, Stevanov-Pavlovic M, Stevanovic J, Bosnjak J, Gajic B,
Aleksic N, Stanimirovic Z (2013). Sex determination in 58 bird
species and evaluation of CHD gene as a universal molecular

Investigation of the validity of species status of Ixodes-Dam-
mini (Acari, Ixodidae) using rDNA. P Natl Acad Sci USA 90:
10221-10225.

Yule GU (1925). A mathematical theory of evolution, based on the
conclusions of Dr. JC Willis, FRS. Philos T R Soc Lon B 213:
21-87.

Zhang S, Yang L, Yang XJ, Yang JX (2007). Molecular phylogeny of
the yuhinas (Sylviidae: Yuhina): a paraphyletic group of bab-
bblers including Zosterops and Philippine Stachyris. J Ornithol

Zuccon D, Ericson PGP (2010). The phylogenetic position of the