

DNA barcoding of primates and the selection of molecular markers using African Great Apes as a model

Amy S. Jackson¹ & Vincent Nijman^{1,2}

1) Oxford Brookes University, Department of Social Sciences, Gipsy Lane, Oxford, OX3 0BP UK

e-mail: amysheere89@live.co.uk

2) Centre for Functional Genomics, Oxford Brookes University, Oxford, OX3 0BP UK

e-mail: vnijman@brookes.ac.uk

Summary - Ambiguities within species description and identification may compromise research validity. Species identification has typically been based upon morphological characteristics, yet recent technological advances have led to identifications achieved via DNA approaches, including DNA barcoding. DNA barcoding studies typically use cytochrome c oxidase subunit I (COI) as the proposed universal molecular marker for animals. Here, we test 12 mitochondrial protein coding genes for the presence of a clear barcoding gap allowing us to unequivocally define species. Using the African Great Apes as our model group, we assess this at the species (*Pan troglodytes*), genus (*Pan*) and family (Hominidae) level. Based on 279 complete mitochondrial genomes, sequences were partitioned by gene for analysis and pairwise distances were calculated. No barcoding gap was observed at the within species level, i.e., the four recognised chimpanzee taxa were not distinguishable through DNA barcoding. However, NADH dehydrogenase subunit 5 (ND5) and cytochrome c oxidase subunit II (COII) produce the largest barcoding gaps at the genus (ND5 2%, COII 0.5%) and family (ND5 1.5%, COII 0.5%) level. Rather than focusing on COI, our analysis suggests that these two genes may be more, or at least as, appropriate markers in primate species delineation, with uses in the identification of extinct and extant species. Further use may be beneficial to taxonomists, providing additional evidence and new insights for these morphologically similar species.

Keywords - COII, DNA barcoding, Great Apes, Molecular marker, ND5, Primate.

Introduction

The requirement for correct species description and identification is of high importance as ambiguities may compromise validity in many areas of research from Anthropology to Zoology. This may then result in data collection and evaluation issues, as well as classification errors (Groves, 2004, 2012, 2014; Isaac *et al.*, 2004; Meier *et al.*, 2006). Species identification has typically been based upon morphological characteristics, yet advances in technology have led to identifications achieved via DNA approaches using phylogenetic reconstructions and sequence similarity database searches (Tautz *et al.*, 2003; Blaxter, 2004; Baker & Bradley, 2006; Dawney *et al.*, 2007; Hajibabaei *et al.*, 2007) such as BOLD (The Barcode of Life Data System)

(Ratnasingham & Hebert, 2007) and GenBank (Benson *et al.*, 1999). One such technique, DNA barcoding (Hebert *et al.*, 2003), has become a well-established practice used by many as it is a relatively cheap, reliable and rapid method for species identification (Hebert & Gregory, 2005; Luo *et al.*, 2011; Čandek & Kuntner, 2015).

The DNA barcoding method employs the use of short standardised regions of the genome to act as taxon 'barcodes' (Hebert *et al.*, 2003). These regions are maintained by selection so as to be almost identical between a taxon but which vary between taxa, meaning genetic variation within species is substantially less than that between species (Hebert *et al.*, 2003; Blaxter, 2004; Hajibabaei *et al.*, 2007). This molecular identification system allows for species

Tab. 1 - The 12 protein coding genes (ATP6, ATP8, cytb, COI-III, ND1-ND5) of the mitochondrial genome used for this study. For each, their approximate length (bp) for DNA barcoding applications is listed along with basic functions and the complex they are a part of (Adapted from Luo *et al.*, 2011).

GENE	COMPLEX	BASE PAIRS	FUNCTION
ATP6	Complex V	681	ATP synthesis, oxidative phosphorylation
ATP8	Complex V	206	ATP synthesis, oxidative phosphorylation
COI	Complex IV	648	Catalytic properties, oxidative phosphorylation
COII	Complex IV	684	Catalytic properties, oxidative phosphorylation
COIII	Complex IV	783	Catalytic properties, oxidative phosphorylation
cytb	Complex III	1140	Electron transfer, oxidative phosphorylation
ND1	Complex I	955	Electron transfer, oxidative phosphorylation
ND2	Complex I	1032	Electron transfer, oxidative phosphorylation
ND3	Complex I	345	Electron transfer, oxidative phosphorylation
ND4	Complex I	1378	Electron transfer, oxidative phosphorylation
ND4L	Complex I	294	Electron transfer, oxidative phosphorylation
ND5	Complex I	1812	Electron transfer, oxidative phosphorylation

identification and discrimination and resolving taxonomic uncertainties in both extinct and extant species (Hebert *et al.*, 2003, 2004; Casiraghi *et al.*, 2010; Galimberti *et al.*, 2015) by comparing the unknown DNA sequences with known sequences deposited within a reference library (Hajibabaei *et al.*, 2007). Gene selection for evaluating species is therefore important as a genetic marker is required to have high reliability for species identification, thus must be orthologous between species and have an appropriate mutation rate for when distinguishing in phylogenetic analyses (Doyle & Gaut, 2000; Blaxter, 2004; Dawnay *et al.*, 2007; Alacs *et al.*, 2010; Dinh *et al.*, 2019).

The mitochondrial gene cytochrome *c* oxidase subunit I (COI) has been proposed as the standard DNA barcoding marker for animals, where a fragment length of 648 base pairs (bp) at the 5' end is sufficient for species identification (Hebert *et al.*, 2003; Hebert & Gregory, 2005; Lorenz *et al.*, 2005; Ivanova *et al.*, 2012). Although mitochondrial DNA (mtDNA) represents only a small fraction of an organisms total cellular DNA (1-2%), these genes are frequently used for DNA barcoding (Clayton, 1982; Galtier *et al.*, 2009; Yang *et al.*, 2014). DNA barcoding therefore goes beyond the use of just COI (Moritz & Cicero, 2004; DeSalle *et al.*, 2005; Casiraghi *et al.*, 2010). This gene alone may not provide sufficient amounts of information relating to the research being conducted, as other genes may be more advantageous for different studies (Meyer, 1994; DeSalle *et al.*, 2005). Choice of an appropriate molecular marker is therefore crucial for the project being undertaken.

The choice of a suitable gene that performs well for various levels of taxa whilst ensuring consistently high phylogenetic resolutions are achieved is important for species identification and recovery (Lorenz *et al.*, 2005; Nijman & Aliabadian, 2010; Luo *et al.*, 2011). Comparisons between intra- and interspecific genetic distances have been used to test gene efficiency in this way, highlighting the existence or lack of a barcoding gap; Interspecific average genetic distances should be at least ten times greater than intraspecific average genetic distances (Meyer & Paulay, 2005).

Many primate taxonomies are still debated and remain largely unresolved, with various studies attempting to clarify phylogenetic relationships (Pastorini *et al.*, 2002; Singer *et al.*, 2003; Whittaker *et al.*, 2007; Zinner *et al.*, 2013), whereby new molecular evidence is being used to revise taxonomies. Despite a proliferation of DNA barcoding studies, thus far only four have been published that specifically deal with primates (Lorenz *et al.*, 2005; Hajibabaei *et al.*, 2006; Nijman & Aliabadian, 2010; Luo *et al.*, 2011); only Nijman & Aliabadian (2010) and Luo *et al.* (2011) compare the performance of more than one mitochondrial gene.

Presented here is an evaluation of the efficacy of 12 mitochondrial protein coding genes (cytochrome *c* oxidase subunit I, II and III, COI, COII and COIII; cytochrome *b*, *cytb*; NADH dehydrogenase subunits 1, 2, 3, 4, 4L and 5, ND1, ND2, ND3, ND4, ND4L and ND5; and ATPase subunits 6 and 8, ATP6 and ATP8; Tab. 1) within the Great Apes, tested at the species (*Pan troglodytes*), genus (*Pan*) and family (*Hominidae*) level; ND6 was excluded due to it being the only protein coding gene on the L-strand and containing many indels (Clayton, 2000; Luo *et al.*, 2011).

Chimpanzees (*Pan troglodytes*) and bonobos (*P. paniscus*) have allopatric distributions (Groves, 2001). The bonobo is monotypic and for chimpanzees, four subspecies are recognised, all separated by geographic barriers: western (*P. troglodytes verus*), central (*P. t. troglodytes*), and eastern (*P. t. schweinfurthii*), and the Nigeria-Cameroon chimpanzee (*P. t. ellioti*) (Oates *et al.*, 2009). Morphological differences between the latter taxa are slight, but molecular studies have found support for the recognition of these taxa (Gonder *et al.*, 2006; Becquet *et al.*, 2007; Hallast *et al.*, 2016). DNA barcoding may provide novel insights into the genetic differentiation between these taxa, and allows for a systematic comparison with other related taxa.

As our null hypothesis, we postulate that there will be no discernible differences between any of the 12 mitochondrial genes for the species, genus or family level comparisons. However, we do expect to find a more conserved genetic pattern at the species level than at the genus and family level.

Methods

We downloaded complete mitochondrial genome sequences from GenBank on 10 January 2020 using the nucleotide database for Hominidae, with parameter settings to include only submissions with a sequence length of 15,000–17,000 bp, producing 51,421 results. Only extant species are included and many sequences available are clones or shotgun

sequences and therefore excluded. Thus, a total of 279 high quality sequences were selected, including 174 sequences for chimpanzee species and subspecies, 43 for bonobos, 30 for gorilla species and its four subspecies, 30 for humans and 2 for orangutan species (Supplementary Material 1). Humans return the largest amount of results in GenBank, however, only a small representative sample have been selected along with orangutans to allow for family level analysis. As this research is completely desk-based, it wholly complies with all ethical and legal requirements of the UK, where we conducted the research.

Complete genomes have been used for consistency throughout datasets and subsequent analyses. Sequences were exported in FASTA format and aligned in MEGA 7 (Kumar *et al.*, 2016) using MUSCLE, a multiple alignment software, due to its high accuracy (Edgar, 2004). Phylogenetic trees were constructed (Gregory, 2008) in MEGA 7 using the neighbour-joining (NJ) method (Saitou & Nei, 1987) with Kimura's 2-parameter (K2P) correction model (Kimura, 1980) and bootstrap of 500 replicates (Felsenstein, 1985), as indeed deployed in previous DNA barcoding studies (Hebert *et al.*, 2003; Casiraghi *et al.*, 2010). Probable erroneous sequence names were corrected: sequences JF727166, JF727173, F727176 and JF727179 were relabelled as *P. t. schweinfurthii*, X93335 was relabelled as *P. t. verus*, and KF914213 was relabelled as an eastern gorilla (*Gorilla beringei*).

We determined each of the 12 protein coding gene positions in order to partition the genomes for analysis. This was achieved via accessing GenBank and locating the average start and end positions for each gene (Supplementary Material 2). We tested each gene region three times; (i) for intra- and interspecific distances within *Pan troglodytes* and its four subspecies, (ii) for intra- and interspecific distances among the genus *Pan* (thus including bonobos), and (iii) for intra- and interspecific distances among the Great Apes (Supplementary Material 3). Due to their extremely close genetic relationship, for the purpose of intra- and interspecific comparisons between species at the genus and family level,

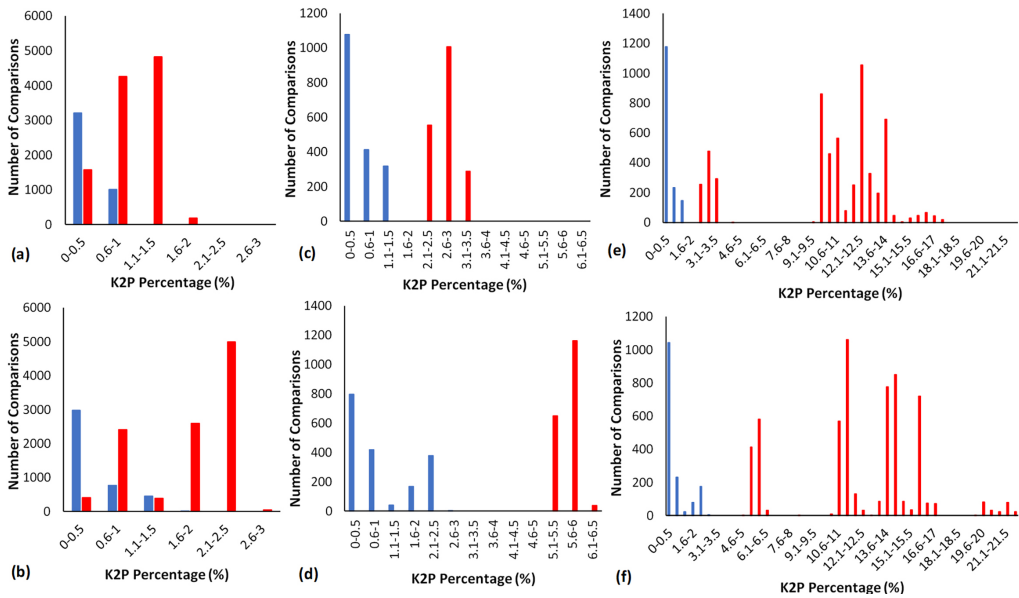


Fig. 1 - Histograms representing the number of sequence comparisons plotted against K2P (Kimura 2-parameter) percentage distances (%) of chimpanzee subspecies using (a) ND5 and (b) COII mitochondrial genes show no clear barcoding gap. However, these mitochondrial genes do appear to perform well due to the more pronounced dip and less overlap in intraspecific (blue) and interspecific (white) variation of K2P distances (%) of species of the genus using (c) ND5 and (d) COII mitochondrial genes show clear barcoding gaps (ND5, 2%; and COII, 0.5%). Intraspecific (blue) and interspecific (white) variation of K2P distances (%) of species of the genus using (e) ND5 and (f) COII mitochondrial genes show clear barcoding gaps (ND5, 1.5%; and COII, 0.5%). The colour version of this figure is available at the JASs website.

all chimpanzee subspecies are classed as one collective species, so as to emphasise the effect of the barcoding gap (Becquet *et al.*, 2007; Prado-Martinez *et al.*, 2013). For each, we computed pairwise distances in MEGA 7 using default parameter settings with the K2P method, bootstrap of 500 replicates and substitutions including both transitions and transversions. We then exported pairwise distances and checked them using ExcaliBAR, a software program used to facilitate the determination of the barcoding gap in pairwise genetic distances (Aliabadian *et al.*, 2014). ExcaliBAR also allows for easy and efficient extraction and separation of intra- and interspecific distances. We created histograms for each gene across each of the three taxonomic levels, where the number of sequence comparisons

were plotted against K2P percentage distances. We also plotted mean K2P percentage distances for each gene at each taxonomic level.

In the Statistical Package for the Social Sciences (SPSS), we used the Kolmogorov-Smirnov test of normality due to the large datasets, followed by a one-way ANOVA to determine any significant variance between the means for each gene across the different taxonomic levels (Dytham, 2003; Davis, 2013). We log transformed ($\log_{10}$) data when deemed necessary so as to approach a normal distribution more closely. Results were assumed to be significant when $P < 0.05$. We calculated the classical barcoding gap for each gene across each taxonomic level, by dividing the average interspecific distance by the average intraspecific distance to determine if a

tenfold difference was apparent between genetic distances (Meyer & Paulay, 2005).

Results

The three condensed NJ trees for all three taxonomic levels using complete mitochondrial genome sequences before being partitioned are represented in Supplementary Material 4. As expected, there are clear groupings of all recognised species across all three taxonomic levels which follow traditional taxonomic classifications. The close grouping of *P. t. troglodytes* and *P. t. schweinfurthii* subspecies is apparent across all three phylogenetic trees, as is the close genetic relationship between gorilla species (western gorilla, *Gorilla gorilla*; eastern gorilla, *G. beringei*) and subspecies (western lowland gorilla, *G. gorilla gorilla*; eastern lowland gorilla, *G. beringei graueri*) within the Hominidae NJ tree. Small discrepancies are apparent between gorilla species and subspecies, with some sequences grouping together at the same branch, which could be due to their high similarity.

Results of the Kolmogorov-Smirnov test found non-normally distributed K2P intra- and interspecific distances for all genes across all taxonomic levels. Although no clear barcoding gap is evident for any of the genes tested at the species level, it would appear that COII, cytb, ND2, ND4 and ND5 perform slightly better than the remaining seven genes due to their more pronounced dip and less overlap between intra- and interspecific genetic distances (Fig. 1). At the genus level, the most pronounced gap is visible for ND5 (2%), closely followed by COII (0.5%) and ND2 (0.5%). Although no clear gap is present, COI also appears to perform well due to no observable overlap. The most pronounced gap for comparisons within Hominidae are visible for ND5 (1.5%), closely followed by COII (0.5%) and ND2 (0.5%). Each gene was also plotted against the mean K2P percentage distances at the species, genus and family level (Fig. 2), with the classical barcoding gap being tested for each gene across each taxonomic level. All ANOVA results

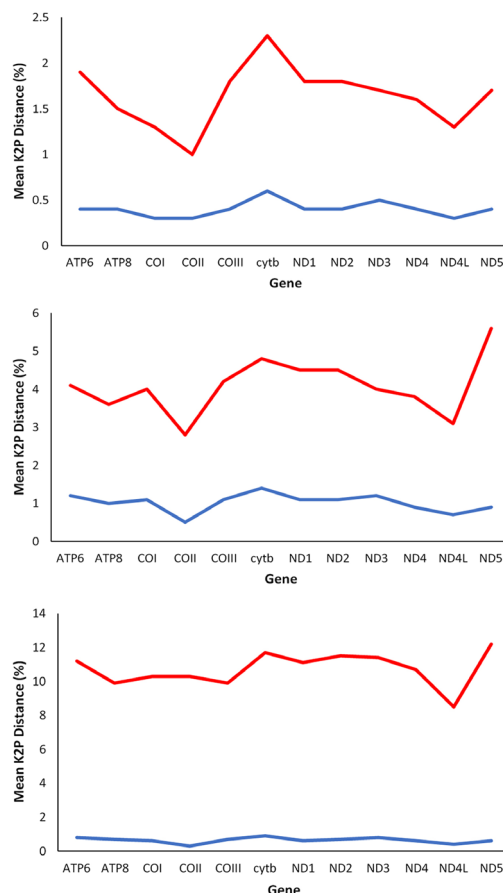


Fig. 2—Mean intra- (blue) and inter-specific (red) distance for each of the 12 mitochondrial genes at the (a) species, (b) genus and (c) family level. The colour version of this figure is available at the JASS website.

are significant at each taxonomic level for all 12 genes (Tab. 2), with results as follows.

Barcoding gap among chimpanzee species and subspecies

For chimpanzee species and subspecies, none of the 12 genes tested have a tenfold difference in K2P mean genetic distances, with the highest reaching only fivefold for four genes; ATP6 intraspecific mean 0.4 ± 0.4 % vs interspecific

Tab. 2 - Summary of descriptive statistics for intra- and interspecific K2P genetic distances (%) for the 12 mitochondrial genes tested across each of the taxonomic levels. For each, the average (%) and standard deviation ($\pm$ SD, %) are recorded, along with minimum and maximum genetic distances observed (range), all to one decimal place. Statistical results of the ANOVA are also reported (P value and F statistic). COMP= Comparison; N comp= Number of Comparisons.

GENE	COMP	INTRASPECIFIC		INTERSPECIFIC		P VALUE	F STATISTIC
		N COMP	MEAN $\pm$ SD (RANGE)	N COMP	MEAN $\pm$ SD (RANGE)		
ATP6	Species	4218	0.4 $\pm$ 0.4 (0-1.5)	10833	1.9 $\pm$ 1.0 (0.2-3.7)	0.001	3801.6
	Genus	1806	1.2 $\pm$ 1 (0-3.3)	1849	4.1 $\pm$ 0.5 (3.1-5.4)	0.001	5429.9
	Family	1555	0.8 $\pm$ 0.9 (0-3.4)	5766	11.2 $\pm$ 4.8 (2.9-25.8)	0.001	20884.4
ATP8	Species	4218	0.4 $\pm$ 0.4 (0-1.5)	10833	1.5 $\pm$ 0.8 (0-3.0)	0.001	5223.4
	Genus	1806	1.0 $\pm$ 0.8 (0-3.0)	1849	3.6 $\pm$ 0.9 (2.0-6.2)	0.001	5771.2
	Family	1555	0.7 $\pm$ 0.7 (0-3.0)	5766	9.9 $\pm$ 4.8 (2.0-31.5)	0.001	14731.9
COI	Species	4218	0.3 $\pm$ 0.3 (0-1.4)	10833	1.3 $\pm$ 0.6 (0.3-2.5)	0.001	5590.9
	Genus	1806	1.1 $\pm$ 0.7 (0-2.5)	1849	4.0 $\pm$ 0.4 (2.0-7-5.0)	0.001	6510.8
	Family	1555	0.6 $\pm$ 0.7 (0-2.5)	5766	10.3 $\pm$ 3.2 (2.9-19.5)	0.001	19984.7
COII	Species	4218	0.3 $\pm$ 0.3 (0-1.0)	10833	1.0 $\pm$ 0.4 (0.1-1.8)	0.001	10121.9
	Genus	1806	0.5 $\pm$ 0.4 (0-1.5)	1849	2.8 $\pm$ 0.3 (2.2-3.5)	0.001	8622.7
	Family	1555	0.3 $\pm$ 0.4 (0-1.5)	5766	10.3 $\pm$ 3.8 (2.2-17.3)	0.001	23434.7
COIII	Species	4218	0.4 $\pm$ 0.3 (0-1.4)	10833	1.8 $\pm$ 0.9 (0.1-3.6)	0.001	11797.7
	Genus	1806	1.1 $\pm$ 1 (0-3.5)	1849	4.2 $\pm$ 0.4 (3.3-5.0)	0.001	4875.6
	Family	1555	0.7 $\pm$ 0.9 (0-3.5)	5766	9.9 $\pm$ 3 (2.5-18.6)	0.001	24078.9
cytb	Species	4218	0.6 $\pm$ 0.4 (0-1.9)	10833	2.3 $\pm$ 0.9 (0.3-4.0)	0.001	18586.1
	Genus	1806	1.4 $\pm$ 1.1 (0-3.7)	1849	4.8 $\pm$ 0.4 (3.9-5.7)	0.001	4269.6
	Family	1555	0.9 $\pm$ 1 (0-3.6)	5766	11.7 $\pm$ 5.2 (4.0-18.0)	0.001	25028.7
ND1	Species	4218	0.4 $\pm$ 0.4 (0-1.6)	10833	1.8 $\pm$ 0.9 (0.2-3.2)	0.001	8440.6
	Genus	1806	1.1 $\pm$ 0.9 (0-3.1)	1849	4.5 $\pm$ 0.4 (3.0-5.3)	0.001	5622.1
	Family	1555	0.6 $\pm$ 0.8 (0-3.1)	5766	11.1 $\pm$ 3.7 (3.8-19.4)	0.001	26920.7
ND2	Species	4218	0.4 $\pm$ 0.3 (0-1.5)	10833	1.8 $\pm$ 0.9 (0-3.1)	0.001	10996.6
	Genus	1806	1.1 $\pm$ 0.9 (0-3.0)	1849	4.5 $\pm$ 0.3 (3.8-5.2)	0.001	5509.9
	Family	1555	0.7 $\pm$ 0.8 (0-3.0)	5766	11.5 $\pm$ 4 (3.6-22.0)	0.001	25137.8
ND3	Species	4218	0.5 $\pm$ 0.5 (0-3.4)	10833	1.7 $\pm$ 0.9 (0-4.3)	0.001	5504.7
	Genus	1806	1.2 $\pm$ 1 (0-3.9)	1849	4.0 $\pm$ 0.7 (2.7-5.5)	0.001	4324.9
	Family	1555	0.8 $\pm$ 0.9 (0-3.9)	5766	11.4 $\pm$ 4.1 (2.7-24.7)	0.001	18629.5
ND4	Species	4218	0.4 $\pm$ 0.4 (0-1.5)	10833	1.6 $\pm$ 0.7 (0.3-2.9)	0.001	14202.4
	Genus	1806	0.9 $\pm$ 0.7 (0-2.8)	1849	3.8 $\pm$ 0.2 (3.3-4.2)	0.001	5881.0
	Family	1555	0.6 $\pm$ 0.6 (0-2.7)	5766	10.7 $\pm$ 3.8 (3.3-20)	0.001	26245.1
ND4L	Species	4218	0.3 $\pm$ 0.4 (0-2.1)	10833	1.3 $\pm$ 0.8 (0-3.5)	0.001	3145.0
	Genus	1806	0.7 $\pm$ 0.7 (0-3.1)	1849	3.1 $\pm$ 0.5 (2.1-4.6)	0.001	6295.9
	Family	1555	0.4 $\pm$ 0.6 (0-3.1)	5766	8.5 $\pm$ 2.9 (2.1-19)	0.001	16230.1
ND5	Species	4218	0.4 $\pm$ 0.4 (0-1.6)	10833	1.7 $\pm$ 0.7 (0.4-2.7)	0.001	17569.5
	Genus	1806	0.9 $\pm$ 0.8 (0-2.7)	1849	5.6 $\pm$ 0.2 (5.2-6.3)	0.001	8288.8
	Family	1555	0.6 $\pm$ 0.7 (0-2.7)	5766	12.2 $\pm$ 3.8 (5.0-22.0)	0.001	35998.7

mean $1.9 \pm 1\%$, $F_{1,13715}=3801.58$, $P<0.001$; COIII intraspecific mean $0.4 \pm 0.3\%$ vs interspecific mean $1.8 \pm 0.9\%$, $F_{1,14416}=11797.72$, $P<0.001$; ND1 intraspecific mean $0.4 \pm 0.4\%$ vs interspecific mean $1.8 \pm 0.9\%$, $F_{1,14244}=8440.61$, $P<0.001$; and ND2 intraspecific mean $0.4 \pm 0.3\%$ vs interspecific mean $1.8 \pm 0.9\%$, $F_{1,14358}=10996.60$, $P<0.001$. Six genes have a fourfold in mean differences (COI, cytb, ATP8, ND4, ND4L and ND5), while the remaining two have only a threefold difference in means (COII and ND3).

Barcoding gap within the genus Pan

For genus, none of the 12 genes tested have a tenfold difference, with the highest reaching sixfold for COII and ND5; COII is almost six times greater for intra- ($0.5 \pm 0.4\%$) and interspecific ($2.8 \pm 0.3\%$) K2P mean distances, $F_{1,3336}=8622.69$, $P<0.001$, and ND5 reaches just over sixfold for intra- ($0.9 \pm 0.8\%$) and interspecific ($5.6 \pm 0.2\%$) mean distances, $F_{1,3554}=8288.77$, $P<0.001$. Seven genes have a fourfold in mean differences (COI, COIII, ATP8, ND1, ND2, ND4 and ND4L). The remaining three genes have only a threefold difference in means (cytb, ND3 and ATP6).

Barcoding gap within Hominidae

At the taxonomic level of family, all genes have more than a tenfold difference. The top performing gene is COII with an interspecific ($10.3 \pm 3.8\%$) distance of 34 times greater than the intraspecific ($0.3 \pm 0.4\%$) K2P mean distance, $F_{1,6772}=23434.69$, $P<0.001$. The next is ND4L followed closely by ND5, with 21 and 20 times greater interspecific distances, respectively (ND4L intraspecific mean $0.4 \pm 0.6\%$ vs interspecific mean $8.5 \pm 2.9\%$, $F_{1,6523}=16230.09$, $P<0.001$; and ND5 intraspecific mean $0.6 \pm 0.7\%$ vs interspecific mean $12.2 \pm 3.8\%$, $F_{1,7067}=35998.65$, $P<0.001$). Four genes range between 16 and 19 times greater (ND1=19, ND4=18, COI=17, ND2=16), four genes have a difference of 14 times greater (ATP6, ATP8, COIII and ND3), and cytb has a difference of 13 times greater.

Discussion

Here we find support for one of the two hypotheses stated. We reject that there will be no difference in performance of the 12 mitochondrial protein coding genes by demonstrating their variability. We found support for their more conserved nature at the species level than at the genus and family level due to no observable barcoding gap within species.

As previously stated, the condensed NJ trees for all taxonomic levels are consistent with traditional taxonomic classifications, with the close genetic relationship of *P. t. troglodytes* and *P. t. schweinfurthii* evident throughout due to their tight clustering. This is supported by previous research where they are known collectively as equatorial Africa chimpanzees (Becquet *et al.*, 2007; Hey, 2010; Gonder *et al.*, 2011). Their divergence is predicted to have occurred only 0.08–0.12 million years ago (MYA), which is much more recent than the divergence between any other subspecies (Gonder *et al.*, 2011), thus offering an explanation for their close grouping. The same is apparent between gorilla species and subspecies, although there are some ambiguities within their alignment. It seems plausible that these discrepancies are a result of the close genetic relationship between species and subspecies (Prado-Martinez *et al.*, 2013; Marques-Bonet & Hvilson, 2018). However, it could also be due to unreliable sequence downloads from GenBank (Bridge *et al.*, 2003; Harris, 2003; Vilgalys, 2003; Galimberti *et al.*, 2015). NUMTs (nuclear mitochondrial DNA sequences) and other pseudogenes are known to produce biases and overestimations of results (Thalmann *et al.*, 2004; Song *et al.*, 2008; Galtier *et al.*, 2009; Casiraghi *et al.*, 2010) and should therefore ideally be identified and excluded. Previous research has documented the presence of a variety of NUMTs among gorilla sequences, with the remaining Great Apes being less problematic (Thalmann *et al.*, 2004). This could therefore explain the discrepancies observed within the Hominidae family NJ tree.

For intra- and interspecific differences at the species level, there is no clear barcoding gap for any

of the genes. Despite the visible overlap observed, this does not necessarily result in failure (Collins & Cruickshank, 2013), as some do appear to perform slightly more well than others; COII, ND4 and ND5 in particular, result in a smaller overlap and a more pronounced dip between intra- and intersubspecific distances. Statistically, *cytb* is the top performing gene at the species level ($F_{1, 14809} = 18586.12$), closely followed by ND5 ($F_{1, 14811} = 17569.5$). These results combined may suggest ND5 to be the most suitable gene to use for species comparisons. Within the genus *Pan*, COII and ND5 both produce the highest classical barcoding gap and F-statistic (sixfold, $F_{1, 3336} = 8622.69$ and $F_{1, 3554} = 8288.77$, respectively), with the intra- and interspecific histograms also confirming this as they visually demonstrate the largest barcoding gaps. Although ND5 does not result in the highest classical barcoding gap at the family level, it does produce the largest F-statistic ($F_{1, 7067} = 35998.65$), suggesting this gene may be the optimal molecular marker. The results of the histogram mapping intra- and interspecific K2P genetic distances also confirms the ND5 gene to be optimal, as it demonstrates the largest barcoding gap. COII does however produce the largest classical barcoding gap result and is also the second top performing gene at the family level when comparing the results of the histograms. However, the large barcoding gaps observed may also be exaggerated as chimpanzee subspecies were classed as one collective species to emphasise the results of the barcoding gap.

The use of average K2P genetic distances to test for the classical barcoding gap may also exaggerate the results (Meier *et al.*, 2008; Kochzius *et al.*, 2010). Here, particularly for intra- and interspecific distances at the family level, the use of means for barcoding calculations appears to exaggerate the results, however, results of the histograms show ND5 to produce the largest barcoding gap. Similarly, at the species level, although no gene resulted in the recommended tenfold difference, statistical analysis of the means would suggest that there should be a significant difference and thus a probable visible barcoding gap. A more precise method may therefore be to

report medians or the overlap between the smallest interspecific and highest intraspecific distances rather than the means (Meier *et al.*, 2008; Čadež & Kuntner, 2015).

The most commonly used molecular marker for primates in previous studies are COI and *cytb*, where they are often used for species identification (Branicki *et al.*, 2003; Verma & Singh, 2003; Lorenz *et al.*, 2005; Nijman & Aliabadian, 2010). This current research statistically supports *cytb* as the top performing gene for within species. Nijman & Aliabadian (2010) found *cytb* to perform better than COI for intra- and interspecific genetic distance analysis among primates due to a much more pronounced barcoding gap. The use of *cytb* for species recovery and identification among primates, specifically chimpanzees and humans, has performed well (Branicki *et al.*, 2003; Verma & Singh, 2003). A range of studies on non-primate taxa using *cytb* have also demonstrated its use for species identification within birds (Dinh *et al.*, 2019), fish (Kochzius *et al.*, 2010), rodents and bats (Bradley & Baker, 2001).

The NADH dehydrogenase genes, excluding ND1, have been found to show higher rates of sequence diversity in primates than the cytochrome *c* oxidase genes (Liang *et al.*, 2012), which may apply to this research as ND5 in particular, as well as other NADH dehydrogenase genes, feature consistently throughout comparisons between all taxonomic levels. ND4 and ND5 have also been used to infer phylogenies and divergence estimates among primate families, including Hominidae, with our results supporting previous research (Hayasaka *et al.*, 1988). Similar to our results, a study on the discrimination of common carp subspecies using ND5 and ND6 has demonstrated their variance due to the large genetic distances observed (Zhou *et al.*, 2003).

Although DNA barcoding has many benefits, concerns regarding the reliability of these studies in general still exist (Moritz & Cicero, 2004; Kress *et al.*, 2015). In particular, the use of COI as the universal molecular marker often receives speculation (Collins & Cruickshank, 2013), again calling attention to the importance of gene choice. A single gene with a short barcode sequence is argued

to not contain sufficient information to infer species relationships that are representative of their full phylogenetic histories (DeSalle *et al.*, 2005; Hajibabaei *et al.*, 2006; Rubinoff, 2006; Collins & Cruickshank, 2013). However, its combined use with classical morphological, physiological and behavioural taxonomy can provide knowledge and new insights, particularly for extant species (Ebach & Holdrege, 2005). Archaeological anthropology studies using DNA barcoding techniques may also benefit by selecting ND5 and COII as molecular markers. As many primate relationships remain somewhat unresolved, DNA barcoding and archaeology may thus provide further understandings and offer resolutions to taxonomic uncertainties. Prior to any genetic research, careful consideration should be taken as to the appropriate molecular marker selected in order to ensure the suitability of a gene for a particular taxon, with the possibility of a small pilot study beforehand (DeSalle *et al.*, 2005; Jafari *et al.*, 2018).

DNA barcoding therefore goes beyond just the use of COI. Based on our findings, it can be concluded that ND5 and COII should be recommended for future use in DNA barcoding research within the African Great Apes. Being morphologically similar, their continued use may provide taxonomists with additional genetic evidence and may also offer new insights into their demographic and geographic histories. However, other species may display different relationships between intra- and interspecific distances depending on gene choice, thus future research should consider and explore the use of alternative genes as not all can be expected to perform optimally across all taxa. Rather than focusing on COI, our analysis suggests that these two genes may be more, or at least as, appropriate markers in primate species delineation.

Data Sharing

The data generated for this research is available from RADAR the institutional repository of Oxford Brookes University: <https://radar.brookes.ac.uk/radar/home.do>

Acknowledgments

We thank Tim Robbins for his assistance with MEGA software.

Author contributions

ASJ organised the structure of the paper and wrote the general parts; VN revised and edited extensively the whole manuscript. Both authors read, provided feedback and approved the manuscript.

References

- Alacs E.A., Georges A., FitzSimmons N.N. *et al.* 2010. DNA detective: A review of molecular approaches to wildlife forensics. *Forensic Sci. Med. Pathol.*, 6: 180-194.
- Aliabadian M., Nijman V., Mahmoudi A. *et al.* 2014. ExcaliBAR: A simple and fast software utility to calculate intra-and interspecific distances from DNA barcodes. *Cont. to Zool.*, 83: 79-84.
- Baker R.J. & Bradley R.D. 2006. Speciation in mammals and the genetic species concept. *J. Mammal.*, 87: 643-662.
- Becquet C., Patterson N., Stone A.C. *et al.* 2007. Genetic structure of chimpanzee populations. *PLoS Genet.*, 3: e66. DOI: <https://doi.org/10.1371/journal.pgen.0030066>
- Benson D.A., Boguski M.S., Lipman D.J. *et al.* 1999. GenBank. *Nucleic Acids Res.*, 27: 38-43.
- Blaxter M.L. 2004. The promise of a DNA taxonomy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 359: 669-679.
- Bradley R.D. & Baker R.J. 2001. A test of the genetic species concept: Cytochrome-*b* sequences and mammals. *J. Mammal.*, 82: 960-973.
- Branicki W., Kupiec T. & Pawlowski R. 2003. Validation of cytochrome *b* sequence analysis as a method of species identification. *J. Forensic Sci.*, 48: 1-5.
- Bridge P.D., Roberts P.J., Spooner B.M. *et al.* 2003. On the unreliability of published DNA sequences. *New Phytol.*, 160: 43-48.

- Čandek K. & Kuntner M. 2015. DNA barcoding gap: Reliable species identification over morphological and geographical scales. *Mol. Ecol. Resour.*, 15: 268-277.
- Casiraghi M., Labra M., Ferri E. *et al.* 2010. DNA barcoding: A six-question tour to improve users' awareness about the method. *Brief. Bioinform.*, 11: 440-453.
- Clayton D.A. 1982. Replication of animal mitochondrial DNA. *Cell*, 28: 693-705.
- Clayton D.A. 2000. Transcription and replication of mitochondrial DNA. *Hum. Reprod.*, 15: 11-17.
- Collins R.A. & Cruickshank R.H. 2013. The seven deadly sins of DNA barcoding. *Mol. Ecol. Resour.*, 13: 969-975.
- Davis C. 2013. *SPSS step by step: Essentials for social and political science*. Policy Press, Bristol.
- Dawnay N., Ogden R., McEwing R., *et al.* 2007. Validation of the barcoding gene COI for use in forensic genetic species identification. *Forensic Sci. Int.*, 173: 1-6.
- DeSalle R., Egan M.G. & Siddall M. 2005. The unholy trinity: Taxonomy, species delimitation and DNA barcoding. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 360: 1905-1916.
- Dinh T.D., Ngatia J.N., Cui L.Y. *et al.* 2019. Influence of pairwise genetic distance computation and reference sample size on the reliability of species identification using Cyt *b* and COI gene fragments in a group of native passerines. *Forensic Sci. Int. Genet.*, 40: 85-95.
- Doyle J.J. & Gaut B.S. 2000. Evolution of genes and taxa: A primer. In J.J. Doyle & B.S. Gaut (eds): *Plant Molecular Evolution*, pp. 1-23. Kluwer Academic Publishers, Dordrecht.
- Dytham C. 2003. *Choosing and using statistics: A biologist's guide* (3rd ed.). Blackwell Science, Chichester.
- Ebach M.C. & Holdrege C. 2005. More taxonomy, not DNA barcoding. *BioScience*, 55: 822-824.
- Edgar R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32: 1792-1797.
- Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39: 783-791.
- Galimberti A., Sandionigi A., Bruno A. *et al.* 2015. DNA barcoding in mammals: What's new and where next? *Hystrix, It. J. Mamm.*, 26: 13-24.
- Galtier N., Nabholz B., Glémin S. *et al.* 2009. Mitochondrial DNA as a marker of molecular diversity: A reappraisal. *Mol. Ecol.*, 18: 4541-4550.
- Gonder M.K., Disotell T.R. & Oates J.F. 2006. New genetic evidence on the evolution of chimpanzee populations and implications for taxonomy. *Int. J. Primatol.*, 27: 1103-1127.
- Gonder M.K., Locatelli S., Ghobrial L. *et al.* 2011. Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *Proc. Natl. Acad. Sci. USA*, 108: 4766-4771.
- Gregory T.R. 2008. Understanding evolutionary trees. *Evo. Edu. Outreach*, 1: 121-137.
- Groves C. 2004. The what, why and how of primate taxonomy. *Int. J. Primatol.*, 25: 1105-1126.
- Groves C. 2012. Species concept in primates. *Am. J. Primatol.*, 74: 687-691.
- Groves C. 2014. Primate taxonomy: Inflation or real? *Annu. Rev. Anthropol.*, 43: 27-36.
- Hajibabaei M., Singer G.A. & Hickey D.A. 2006. Benchmarking DNA barcodes: An assessment using available primate sequences. *Genome*, 49: 851-854.
- Hajibabaei M., Singer G.A., Hebert P.D. *et al.* 2007. DNA barcoding: How it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.*, 23: 167-172.
- Hallast P., Delser P.M., Batini C. *et al.* 2016. Great ape Y Chromosome and mitochondrial DNA phylogenies reflect subspecies structure and patterns of mating and dispersal. *Genome Res.*, 26: 427-439.
- Harris D.J. 2003. Can you bank on GenBank? *Trends Ecol. Evol.*, 18: 317-319.
- Hayasaka K., Gojobori T. & Horai S. 1988. Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.*, 5: 626-644.
- Hebert P.D., Cywinska A., Ball S.L. *et al.* 2003. Biological identifications through DNA

- barcodes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 270: 313-321.
- Hebert P.D., Penton E.H., Burns J.M. *et al.* 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl. Acad. Sci. USA*, 101: 14812-14817.
- Hebert P.D. & Gregory T.R. 2005. The promise of DNA barcoding for taxonomy. *Syst. Biol.*, 54: 852-859.
- Hey J. 2010. The divergence of chimpanzee species and subspecies as revealed in multipopulation isolation-with-migration analyses. *Mol. Biol. Evol.*, 27: 921-933.
- Isaac N.J., Mallet J. & Mace G.M. 2004. Taxonomic inflation: Its influence on macroecology and conservation. *Trends Ecol. Evol.*, 19: 464-469.
- Ivanova N.V., Clare E.L. & Borisenko A.V. 2012. DNA barcoding in mammals. In W.J. Kress & D.L. Erickson (eds): *DNA barcodes: Methods and protocols*, pp. 153-182. Humana Press, Totowa.
- Jafari M., Akram W., Pang Y. *et al.* 2018. Genetic diversity and biogeography of *T. officinale* inferred from multi locus sequence typing approach. *PLoS One*, 13: 1-18. DOI: <https://dx.doi.org/10.1371/journal.pone.0203275>
- Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16: 111-120.
- Kochzius M., Seidel C., Antoniou A. *et al.* 2010. Identifying fishes through DNA barcodes and microarrays. *PLoS One*, 5: e12620. DOI: <https://doi.org/10.1371/journal.pone.0012620>
- Kress W.J., García-Robledo C., Uriarte M. *et al.* 2015. DNA barcodes for ecology, evolution, and conservation. *Trends Ecol. Evol.*, 30: 25-35.
- Kumar S., Stecher G. & Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.*, 33: 1870-1874.
- Liang Z., Zhang X., Xingkui T.A.O. *et al.* 2012. Preliminary analysis of the mitochondrial genome evolutionary pattern in primates. *Zool. Res.*, 33: 47-56.
- Lorenz J.G., Jackson W.E., Beck J.C. *et al.* 2005. The problems and promise of DNA barcodes for species diagnosis of primate biomaterials. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 360: 1869-1877.
- Luo A., Zhang A., Ho S.Y. *et al.* 2011. Potential efficacy of mitochondrial genes for animal DNA barcoding: A case study using eutherian mammals. *BMC Genomics*, 12: 1-13. DOI: <https://doi.org/10.1186/1471-2164-12-84>
- Marques-Bonet T. & Hvilson C. 2018. Genomic variation of the great apes and the application to conservation. *International Zoo Yearbook*, 52: 25-33.
- Meier R., Shiyang K., Vaidya G. *et al.* 2006. DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Syst. Biol.*, 55: 715-728.
- Meier R., Zhang G. & Ali F. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. *Syst. Biol.*, 57: 809-813.
- Meyer A. 1994. Shortcomings of the cytochrome *b* gene as a molecular marker. *Trends Ecol. Evol.*, 9: 278-280.
- Meyer C.P. & Paulay G. 2005. DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biol.*, 3: e422. DOI: <https://doi.org/10.1371/journal.pbio.0030422>
- Moritz C. & Cicero C. 2004. DNA barcoding: Promise and pitfalls. *PLoS Biol.*, 2: e354. DOI: <https://doi.org/10.1371/journal.pbio.0020354>
- Nijman V. & Aliabadian M. 2010. Performance of distance-based DNA barcoding in the molecular identification of primates. *C. R. Biol.*, 333: 11-16.
- Oates J.F., Groves C.P. & Jenkins P.D. 2009. The type locality of *Pan troglodytes vellerosus* (Gray, 1862), and implications for the nomenclature of West African chimpanzees. *Primates*, 50: 78-80.
- Pastorini J., Forstner M.R. & Martin R.D. 2002. Phylogenetic relationships among Lemuridae (Primates): evidence from mtDNA. *J. Hum. Evol.*, 43: 463-478.
- Prado-Martinez J., Sudmant P.H., Kidd J.M. *et al.* 2013. Great ape genetic diversity and population history. *Nature*, 499: 471-475.

- Ratnasingham S. & Hebert P.D. 2007. BOLD: The Barcode of Life Data System (<http://www.barcodinglife.org>). *Mol. Ecol. Notes*, 7: 355-364.
- Rubinoff D. 2006. Utility of mitochondrial DNA barcodes in species conservation. *Cons. Biol.*, 20: 1026-1033.
- Saitou N. & Nei M. 1987. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4: 406-425.
- Singer S.S., Schmitz J., Schwegk C. *et al.* 2003. Molecular cladistic markers in New World monkey phylogeny (Platyrrhini, Primates). *Mol. Phylogenet. Evol.*, 26: 490-501.
- Song H., Buhay J.E., Whiting M.F. *et al.* 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proc. Natl. Acad. Sci. USA*, 105: 13486-13491.
- Tautz D., Arctander P., Minelli A. *et al.* 2003. A plea for DNA taxonomy. *Trends Ecol. Evol.* 18: 70-74.
- Thalmann O., Hebler J., Poinar H.N. *et al.* 2004. Unreliable mtDNA data due to nuclear insertions: A cautionary tale from analysis of humans and other great apes. *Mol. Ecol.*, 13: 321-335.
- Verma S.K. & Singh L. 2003. Novel universal primers establish identity of an enormous number of animal species for forensic application. *Mol. Ecol. Notes*, 3: 28-31.
- Vilgalys R. 2003. Taxonomic misidentification in public DNA databases. *New Phytol.*, 160: 4-5.
- Whittaker D.J., Morales J.C. & Melnick D.J. 2007. Resolution of the *Hylobates* phylogeny: congruence of mitochondrial D-loop sequences with molecular, behavioural, and morphological data sets. *Mol. Phylogenet. Evol.*, 45: 620-628.
- Yang L., Tan Z., Wang D. *et al.* 2014. Species identification through mitochondrial rRNA genetic analysis. *Sci. Rep.*, 4: 4089-4099.
- Zhou J., Wang Z., Ye Y. *et al.* 2003. PCR-RFLP analysis of mitochondrial DNA ND5/6 region among 3 subspecies of common carp (*Cyprinus carpio* L.) and its application to genetic discrimination of subspecies. *Chin. Sci. Bull.*, 48: 465-468.
- Zinner D., Wertheimer J., Liedigk R. *et al.* 2013. Baboon phylogeny as inferred from complete mitochondrial genomes. *Am. J. Phys. Anthropol.*, 150: 133-140.

Editor, Giovanni Destro Bisol



This work is distributed under the terms of a Creative Commons Attribution-NonCommercial 4.0 Unported License <http://creativecommons.org/licenses/by-nc/4.0/>