Mitochondrion 9 (2009) 36–40

Contents lists available at ScienceDirect

Mitochondrion

journal homepage: www.elsevier.com/locate/mito

mtDNA GeneExtractor: A computer tool for mtDNA gene/region information extraction

Fernando Freitas [a], Sandra Oliveira [a], Ricardo Rocha [b], Luísa Pereira [a,c,*]

[a] Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal
[b] Departamento de Ciências de Computadores, Faculdade de Ciências da Universidade do Porto, Portugal
[c] Medical Faculty, University of Porto, Portugal

ARTICLE INFO

Article history:
Received 1 August 2008
Received in revised form 19 November 2008
Accepted 21 November 2008
Available online 3 December 2008

Keywords:
mtDNA
Computer tool
Extraction
Transition/transversion ratios

ABSTRACT

The analysis of considerable numbers of DNA sequences is largely dependent on the development of simple software tools for automatically process the genetic data deposited on public databases. However, there are some difficulties in the automation process due to diverse synonyms being used as qualifiers for genes and some inconsistencies in gene locations between related Primate species, this fact happening even in the carefully curated database RefSeq. Here, we present mtDNA GeneEXtractor, a Windows based computer tool developed for the extraction of information for particular gene/regions from mammal mitochondrial DNA sequences deposited under GenBank format. The tool was quite efficient in retrieving organized information for comparative mtDNA gene/region diversity analyses when tested for the evaluation of transition/transversion ratios in humans and between Primates. Taking phylogenetic information into account to avoid redundancy due to ancestry-sharing, the transition/transversion ratios in the 13 protein-coding genes had a mean value of 12.46 for Primates (from 6.46 in ND2 to 17.04 in COX1) and higher (34.74) but more heterogeneous (ranging from 17.30 in ND5 to 74.39 in ND4) in a worldwide human database. The similar patterns of transition/transversion ratios in all positions and in only four fold degenerate positions show no evidence for selection in the 13 mtDNA protein-coding genes.

1. Introduction

When Zuckerkandl and Pauling (1965) suggested the hypothesis of the evolutionary molecular clock, which would allow the dating of evolutionary events, it became possible to infer the phylogenetic history of a set of species (reviewed in Ayala, 1986). Supporting the existence of a molecular clock, the neutral theory of molecular evolution, developed by Kimura in the transition from the 1960s to the 1970s, postulated that the molecular evolution rates are stochastically constant given that the vast majority of molecular differences between species are selectively neutral (Kimura, 1977). The diversity would be driven mainly by genetic drift, rather than by selective pressure.

Meanwhile, some tests showed that evolution rates are higher than expected by the regularity of the neutral theory. But this characteristic does not invalidate the molecular clock (Ayala, 1986): evolution is sufficiently regular so that a molecular clock can be applied to most situations. Attention should be devoted to some uncertainties which can modify the properties of the molecular clock, such us events which can lead to big fluctuations on mutation rates between different periods of time or between different lineages.

By the time these theoretical concepts were being developed, a few DNA or protein sequences, from diverse organisms, were available for validation of evolution rates. One of the first molecules to be analysed was the mitochondrial protein cytochrome *c*, which displayed a regularity supporting its application as a molecular clock (Ayala, 1986).

The mitochondrial proteins continue to be an informative tool for the inference of phylogeny. A low proportion of these proteins are coded by a peculiar genome, the mitochondrial DNA (mtDNA), which is located inside mitochondria. Almost all mtDNAs are circular and not protected by histones, displaying high mutation rates and absence of recombination due to the exclusive maternal inheritance (haploid genome). The high mutation rates allow a stronger power to distinguish individuals; the absence of recombination allows an easier inference of the phylogenetic evolution as no shifting of portions of the DNA molecules occurs between homolog copies. Recently, considerable technological development is allowing the publication of a huge number of complete mtDNA sequences of many individuals from the same or different species. It is now possible to use these big public mtDNA databases to evaluate the fitness of diverse theoretical models. The most used and largest molecular public database is GenBank (Wheeler et al.,

* Corresponding author. Address: Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal. Tel.: +351 22 5570700; fax: +351 22 5570799.
E-mail address: lpereira@ipatimup.pt (L. Pereira).

1567-7249/$ - see front matter © 2008 Elsevier B.V. and Mitochondria Research Society. All rights reserved.
doi:10.1016/j.mito.2008.11.003

2008), which accounts now more than 1400 different species and 5000 human complete mtDNA genomes.

Many works centred on mtDNA diversity within (e.g. Behar et al., 2008) and between species (e.g. Belle et al., 2005) are being conducted. As particularly demonstrated in humans, there seems to be a high heterogeneity in mutation rates between positions inside some regions of the mtDNA molecule, mainly inside two control region segments, designated accordingly as hypervariable segments I and II (Meyer et al., 1999). The control region or D-loop (~1200 bp in humans), as the name indicates, has controlling functions for the molecule replication and transcription, including hypervariable regions with portions mutating very fast because they do not seem to have any function. The rest of the molecule is the coding region, and in mammals, it is segmented into 13 genes coding for proteins (having some evolution constraints dictated by the genetic code – some alterations modify the proteins while others do not), 22 tRNAs (having evolution constraints due to its complex secondary structure) and two rRNAs (another secondary structure requiring conservation heterogeneity in different portions).

Again, a discussion ensued on the adequacy of using the very quickly mutating mtDNA regions as molecular clocks (see Macaulay et al., 1997 versus Howell et al., 1996). The comparative study of mtDNA coding portions, a slower evolving portion of the molecule, is supporting inferences attained by the fast molecular clocks (e.g. Pereira et al., 2005). Additional tests, comparing different portions of the mtDNA molecule bearing diverse molecular characteristics, should be performed in other species for which increasing numbers of haplotypes are being published (for example, the mouse, *Mus musculus domesticus*; Goios et al., 2007).

The studies focused on ascertaining and using the genetic heterogeneity between different mtDNA portions depend on the development of computer tools which allow the automatic and efficient extraction of information from public databases (Goios et al., 2006). Some tools allow comparative analyses, as it is the case of the MamMiBase, developed by Vasconcelos et al. (2005), which has the limitation of surveying a few and pre-defined species for only the 13 mtDNA protein-coding genes. Another database, Mamit-tRNA, displays the alignments for the 22 tRNAs in several species (Helm et al., 2000).

In this work, we introduce a new computer tool developed for an easy and efficient extraction of any of the 38 gene/regions that constitute mammal mtDNA, from whichever mammal species the user intends to study. The software was developed for Windows based platforms and can be freely download from http://www.ipatimup.pt/downloads/mtDNAGeneExtractor.zip. The input data should be in the format GenBank, which provides annotated information, namely for gene locations, translation for protein-coding genes and tRNA codons recognized. However, as the responsibility for annotating a sequence belongs to the submitting lab, many inconsistencies can be found in GenBank entries, dramatically limiting the efficiency of computer tools to perform searches across the database (Karp, 2001). In order to solve some of these problems, the National Center for Biotechnology Information (NCBI), responsible for GenBank, created RefSeq database. RefSeq is nonredundant (one sequence per gene/genome per species), curated by GenBank staff with defined qualifiers (Pruitt et al., 2007), which intends to function as a reference for genome databases. We used the revised information of RefSeq for implementing the tool presented here.

2. Design, implementation and functionality

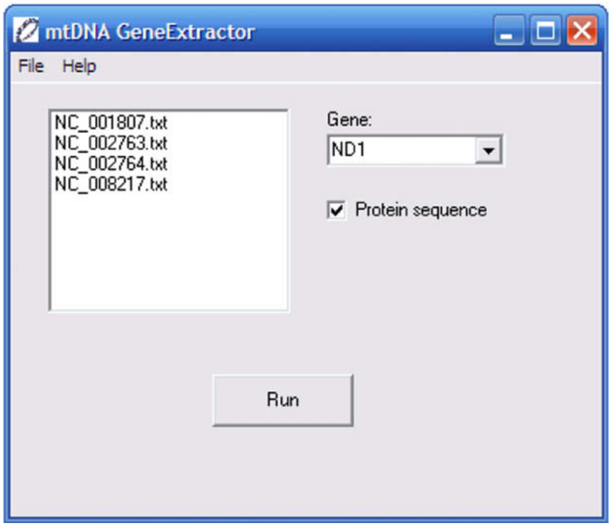
The mtDNA GeneExtractor tool was developed for Windows based platforms and implemented using the C++ language. Its graphical environment is very simple and intuitive (Fig. 1). The left window shows the input sequences (it allows the simultaneous input of several files in GenBank format saved as text file) and the right window is used for selection of the gene/region to extract during the parsing process. In order to find/extract the location of the selected gene/region inside the input sequences, the location is marked by a string used in the RefSeq database (a list can be found in Table 1). But even in the case of a curated database, some genes, such as rRNA's, have more than one way of identification, making it necessary to include several options as qualifiers for a single gene. As aminoacids leucine and serine can be coded by different codons recognized by two tRNAs each, additional information about codons recognized must be added (Table 1).

As the beginning of the circular sequence is arbitrary when the first individual of that species is sequenced, the D-loop (more than 1000 bp) is sometimes split into two fragments. When this occurs, the word "join" is shown with the location of this region. The program will bring together both halves, in the correct direction.

While testing the tool for GenBank files deposited in RefSeq, for Primate mtDNAs, we have found some other inconsistencies, which rendered difficult the automation process. For instance, the location can be preceded by the signal '>'or '<', meaning that the author was not sure about the beginning of the gene. In such cases, the signal is ignored and the location is interpreted as beginning in the position indicated. In other situations, some gene/regions locations are not indicated or are incomplete in the file, this occurring mainly for the D-loop region (totally missing in NC_008220 and only hypervariable regions were indicated in NC_008066). In such cases, a manual correction must be performed.

Users must be aware that, if using this tool to extract genes coded in the complementary strand, the sequence obtained is the one provided in the GenBank file (5′–3′). So, they should complement the segment obtained for further comparison and translation (as for ND6 gene).

To extract the protein sequence, the checkbox button must be selected. The same string is used to identify the protein in the file, but the amino acid sequence extracted is present some lines below and marked in the file with the word "translation". When the pro-



NC_001807.txt
NC_002763.txt
NC_002764.txt
NC_008217.txt

Gene: ND1

☑ Protein sequence

Fig. 1. Graphic interface of mtDNA GeneExtractor. Left window for input of sequences and right window for selection of gene to extract. If the option "Protein sequence" is selected, the information extracted will be the protein sequence.

Table 1
List of qualifiers used to identify the gene/regions and its location in the genome.

Gene	Qualifiers	Necessary additional information
ATP6	/gene = “ATP6”	–
ATP8	/gene = “ATP8”	–
COX1	/gene = “COX1”	–
COX2	/gene = “COX2”	–
COX3	/gene = “COX3”	–
CYTB	/gene = “CYTB”	–
ND1	/gene = “ND1”	–
ND2	/gene = “ND2”	–
ND3	/gene = “ND3”	–
ND4	/gene = “ND4”	–
ND4L	/gene = “ND4L”	–
ND5	/gene = “ND5”	–
ND6	/gene = “ND6”	–
12S rRNA	/product = “12S ribosomal RNA” /product = “s-rRNA”	–
16S rRNA	/product = “16S ribosomal RNA” /product = “l–rRNA”	–
tRNA Alanine	/product = “tRNA–Ala”	–
tRNA Arginine	/product = “tRNA–Arg”	–
tRNA Asparagine	/product = “tRNA–Asn”	–
tRNA Aspartate	/product = “tRNA–Asp”	–
tRNA Cysteine	/product = “tRNA–Cys”	–
tRNA Glutamate	/product = “tRNA–Glu”	–
tRNA Glutamine	/product = “tRNA–Gln”	–
tRNA Glycine	/product = “tRNA–Gly”	–
tRNA Histidine	/product = “tRNA–His”	–
tRNA Isoleucine	/product = “tRNA–Ile”	–
tRNA Leucine UUR	/product = “tRNA–Leu”	/note=”codons recognized: UUR” /codon_recognized=”UUR”
tRNA Leucine CUN	/product = “tRNA–Leu”	/note=”codons recognized: CUN” /codon_recognized=”CUN”
tRNA Lysine	/product = “tRNA–Lys”	–
tRNA Methionine	/product = “tRNA–Met”	–
tRNA Phenylalanine	/product = “tRNA–Phe”	–
tRNA Proline	/product = “tRNA–Pro”	–
tRNA Serine UCN	/product = “tRNA–Ser”	/note=”codons recognized: UCN” /codon_recognized=”UCN”
tRNA Serine AGY	/product = “tRNA–Ser”	/note=”codons recognized: AGY” /codon_recognized=”AGY”
tRNA Threonine	/product = “tRNA–Thr”	–
tRNA Tryptophane	/product = “tRNA–Trp”	–
tRNA Tyrosine	/product = “tRNA–Tyr”	–
tRNA Valine	/product = “tRNA–Val”	–
D-loop	D-loop/note = “D-loop”/note = “control region”	–

tein sequence checkbox button is selected together with a nonprotein-coding gene, an error message is displayed.

The program reads each GenBank input file and writes the required information into an output file, with FASTA format and text extension. For this reason, the program does not need a great amount of memory. In the output file, each sequence is identified with the Accession Number and the chosen gene, being the order of the sequences the same as listed in the interface window. The FASTA format was chosen because it is an input file accepted by most used programs dealing with phylogenetic and diversity analyses.

When the program is not able to obtain the information of the location of the gene due to the absence of the standard identification string, the program will skip these files, continuing to the following ones, and a message indicating that the information was not possible to obtain from all files will appear.

3. Application to the estimation of transition/transversion ratios per protein-coding genes in humans and other Primates

To test the efficiency and automation of mtDNA GeneExtractor, the 13 protein-coding genes were extracted from two mtDNA datasets: (1) 53 humans representing a worldwide sampling (Ingman et al., 2000); and (2) 23 Primates (information displayed in Table S1). These data were then used to estimate the values of transition/transversion (ts/tv) ratios.

When using BioEdit (Hall, 1999) to perform the alignment of the 13 protein-coding genes extracted from the Primate dataset, we have noticed that many genes were beginning or ending in incorrect locations (this conclusion being reached from confirmation of match of the segments missing in some of the Primates species). The same occurred for tRNAs when comparing with information reported by Helm et al. (2000). The location inconsistencies detected in Primate mtDNAs are provided in Table S2.

For the human dataset, as they are not deposited in RefSeq, annotations were different in protein-coding genes, being necessary to add manually to all GenBank files the qualifiers for them, as for instance /gene = “ATP6”. This absence is however unusual in GenBank files.

The ts/tv ratios were evaluated in two ways: (1) by direct counting, dividing positions showing transitions by positions showing transversions; and (2) by taking in consideration the phylogenetic information to avoid redundancy due to ancestry-sharing of substitutions (substitution which occurred only once along time but being present in all related individuals). We applied both strategies to both datasets by using, respectively, the programs Arlequin (Excoffier et al., 2005) and PAML (standing for phylogenetic analysis by maximum likelihood; Yang, 2007). In the case of PAML, the molecular model used was HKY85 and it was necessary to input the phylogenetic trees for each dataset: for the human dataset, the reconstruction of Bandelt et al. (2006) of the dataset published by Ingman et al. (2000) was used; for Primates, the tree was reconstructed from information published in Purvis (1995) and is displayed in Fig. S1. The results for ts/tv ratios are presented in Fig. 2.

The direct counting leads to lower estimations than PAML, because it is centred in variable positions, not evaluating the pairwise differences between all individuals. Nevertheless, the direct counting allowed us to confirm that we were obtaining the same results as Belle et al. (2005) who checked this value in the same human

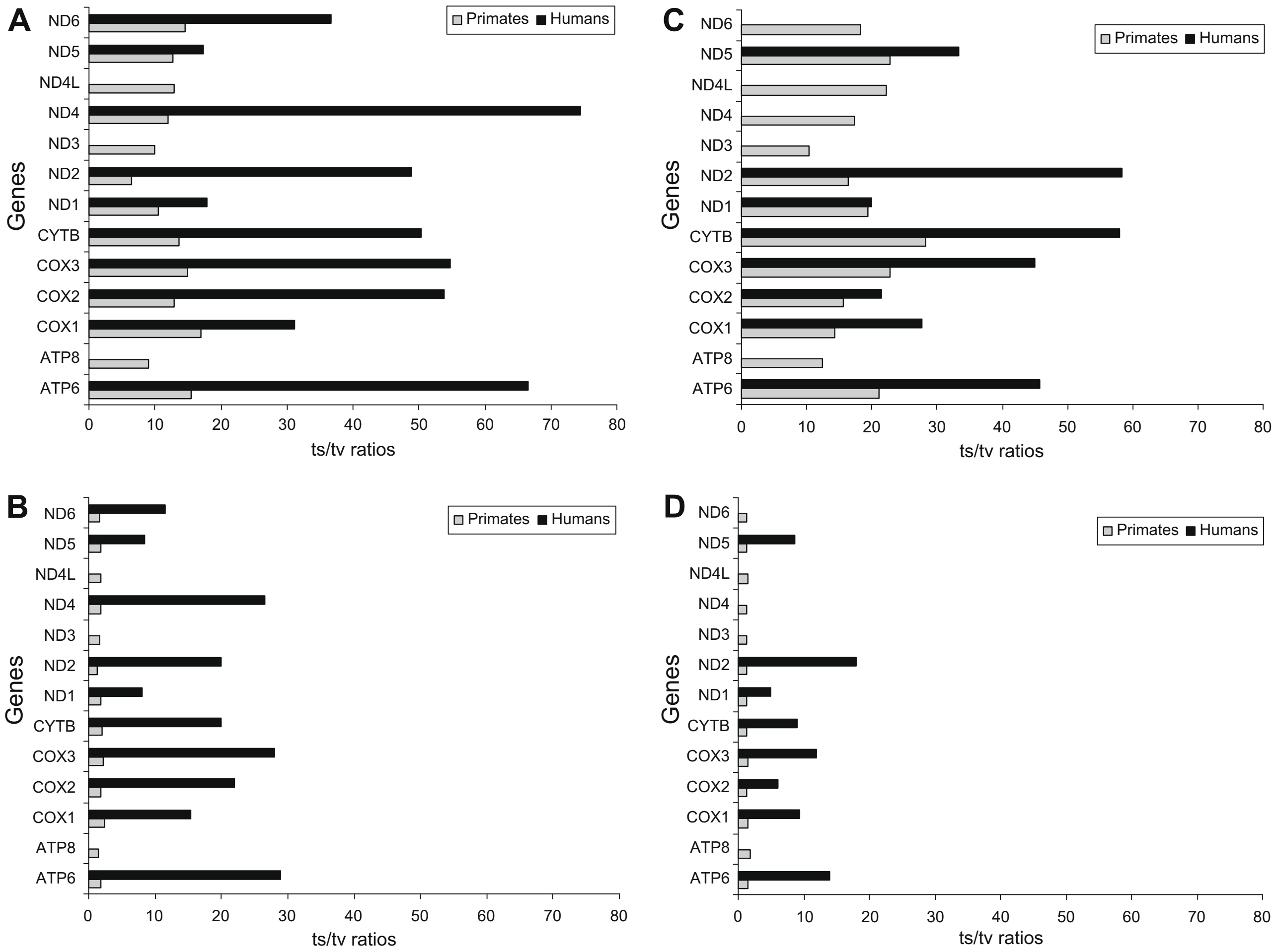


Fig. 2. Transition/transversion ratios in the 13 mtDNA protein-coding genes in humans and Primates, determined for all positions in PAML (A) and by direct counting (B) and for 4-fold positions in PAML (C) and by direct counting (D).

database, but for the 13 protein-coding genes together. Both methods calculated a higher bias towards transitions inside humans than between Primates, in accordance with previous findings that rates of mtDNA variation are higher inside than between species (Nachman et al., 1996).

So, for instance, taking phylogenetic information into account, the transition/transversion ratios in the 13 protein-coding genes had a mean value of 12.46 for Primates (from 6.46 in ND2 to 17.04 in COX1) and higher (34.74) but more heterogeneous (ranging from 17.30 in ND5 to 74.39 in ND4) in a worldwide human database. All genes had transversions when comparing Primates, in contrast with the human dataset for which no transversions were observed in ATP8, ND3 and ND4L.

We also tested these estimations in only four fold degenerate positions, which are the third codon positions that can vary for any of the four bases, meaning the same amino acid; these positions are never under the effects of selection. The distributions were very similar between all positions and four fold degenerate positions. This result seems to point to the absence of clear selection effects in any of the 13 protein-coding genes. Previous claims of selection driving evolution of human mtDNA protein-coding genes inside some haplogroups have been published (Mishmar et al., 2003), but latter evidence favoured a slight purifying selection affecting the younger branches of the human phylogeny (Kivisild et al., 2006).

Further use of our tool in similar analyses being conducted in larger datasets of species from different taxa can help to clarify the disagreement between Yang and Yoder (1999) to Belle et al. (2005): Yang and Yoder claim to detect heterogeneity in rates between different Primates branches, while Belle et al. found no variation not only between closely related species but also within orders.

4. Final remarks

mtDNA GeneExtractor can be applied to the analyses of large datasets in GenBank format. As there is still some inconsistency in public databases relatively to the annotation of deposited genomes, manual survey of extracted files is advised. Diverse evaluations of comparative genetic variability between mtDNA gene/regions can be easily performed, as demonstrated here for the evaluation of transition/transversion ratios in humans and between Primates. Results obtained showed that the bias towards transitions is more pronounced inside humans than between Primates. The method attending to the phylogeny led to higher estimations of transition/transversion ratios than the direct counting of polymorphic positions. And the most significant result was the no evidence for selection in the 13 mtDNA protein-coding genes as no differences were observed for patterns of transition/transversion ratios between all positions and only four fold degenerate positions.

Acknowledgements

This work was partially supported by Fundação para a Ciência e a Tecnologia, through Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mito.2008.11.003.

References

Ayala, F.J., 1986. On the virtues and pitfalls of the molecular evolutionary clock. J. Heredity 77, 226–235.

Bandelt, H.-J., Kong, Q.-P., Yao, Y.-G., Richards, M., Macaulay, V., 2006. Estimation of mutation rates and coalescence times: some caveats. In: Bandelt, H.-J., Macaulay, V., Richards, M. (Eds.), Mitochondrial DNA and the Evolution of Homo sapiens. Springer-Verlag, Berlin, Germany.

Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., Comas, D., Bertranpetit, J., Quintana-Murci, L., Tyler-Smith, C., Wells, R.S., Rosset, S.Genographic Consortium, 2008. The dawn of human matrilineal diversity. Am. J. Hum. Genet. 82, 1130–1140.

Belle, E.M., Piganeau, G., Gardner, M., Eyre-Walker, A., 2005. An investigation of the variation in the transition bias among various animal mitochondrial DNA. Gene 355, 58–66.

Excoffier, L., Laval, G., Schneider, S., 2005. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. Evolution. Bioinformatics 1, 47–50. Online.

Goios, A., Meirinhos, J., Rocha, R., Lopes, R., Amorim, A., Pereira, L., 2006. RepeatAround: a software tool for finding and visualizing repeats in circular genomes and its application to a human mtDNA database. Mitochondrion 6, 218–224.

Goios, A., Pereira, L., Bogue, M., Macaulay, V., Amorim, A., 2007. mtDNA phylogeny and evolution of laboratory mouse strains. Genome Res. 17, 293–298.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41, 95–98.

Helm, M., Brule, H., Friede, D., Giege, R., Putz, D., Florentz, C., 2000. Search for characteristic structural features of mammalian mitochondrial tRNAs. RNA 6, 1356–1379.

Howell, N., Kubacka, I., Mackey, D.A., 1996. How rapidly does the human mitochondrial genome evolve? Am. J. Hum. Genet. 59, 501–509.

Ingman, M., Kaessmann, H., Paabo, S., Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408, 708–713.

Karp, P.D., 2001. Many genbank entries for complete microbial genomes violate the genbank standard. Comput. Funct. Genomics 2, 25–27.

Kimura, M., 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267, 275–276.

Kivisild, T., Shen, P., Wall, D.P., Do, B., Sung, R., Davis, K., Passarino, G., Underhill, P.A., Scharfe, C., Torroni, A., Scozzari, R., Modiano, D., Coppa, A., de Knijff, P., Feldman, M., Cavalli-Sforza, L.L., Oefner, P.J., 2006. The role of selection in the evolution of human mitochondrial genomes. Genetics 172, 373–387.

Macaulay, V.A., Richards, M.B., Forster, P., Bendall, K.E., Watson, E., Sykes, B., Bandelt, H.J., 1997. mtDNA mutation rates-no need to panic. Am. J. Hum. Genet. 61, 983–990.

Meyer, S., Weiss, G., von Haeseler, A., 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics 152, 1103–1110.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., Sukernik, R.I., Olckers, A., Wallace, D.C., 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA 100, 171–176.

Nachman, M.W., Brown, W.M., Stoneking, M., Aquadro, C.F., 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics 142, 953–963.

Pereira, L., Richards, M., Goios, A., Alonso, A., Albarran, C., Garcia, O., Behar, D.M., Golge, M., Hatina, J., Al-Gazali, L., Bradley, D.G., Macaulay, V., Amorim, A., 2005. High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res. 15, 19–24.

Pruitt, K.D., Tatusova, T., Maglott, D.R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35 (database issue), D61–65.

Purvis, A., 1995. A composite estimate of primate phylogeny. Philos. Trans. R. Soc. Lond. B Biol. Sci. 348, 405–421.

Vasconcelos, A.T., Guimaraes, A.C., Castelletti, C.H., Caruso, C.S., Ribeiro, C., Yokaichiya, F., Armoa, G.R., Pereira, G.daS., da Silva, I.T., Schrago, C.G., Fernandes, A.L., da Silveira, A.R., Carneiro, A.G., Carvalho, B.M., Viana, C.J., Gramkow, D., Lima, F.J., Correa, L.G., Mudado, M.deA., Nehab-Hess, P., Souza, R., Correa, R.L., Russo, C.A., 2005. MamMiBase: a mitochondrial genome database for mammalian phylogenetic studies. Bioinformatics 21, 2566–2567.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Ostell, J., Pruitt, K.D., Schuler, G.D., Shumway, M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L., Yaschenko, E., 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36 (database issue), D13–21.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Yang, Z., Yoder, A.D., 1999. Estimation of the transition/transversion rate bias and species sampling. J. Mol. Evol. 48, 274–283.

Zuckerkandl, E., Pauling, L., 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8, 357–366.