Mitochondrion 11 (2011) 946–953

Contents lists available at SciVerse ScienceDirect

Mitochondrion

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

Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies

Amanda Ramos [a], Elena Barbena [a], Ligia Mateiu [b], María del Mar González [a], Quim Mairal [a], Manuela Lima [c,d], Rafael Montiel [e], Maria Pilar Aluja [a], Cristina Santos [a,*]

[a] Unitat d'Antropologia Biològica, Departament BABVE, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain
[b] Laboratory for Molecular Plant Physiology and Biotechnology CGB, University of Antwerpen Groenenborgerlaan, 171 B-2020 Antwerpen, Belgium
[c] IBMC — Institute for Molecular and Cellular Biology, University of Porto, Porto, Portugal
[d] Center of Research in Natural Resources (CIRN), University of the Azores, Ponta Delgada, Portugal
[e] Laboratorio Nacional de Genómica para la Biodiversidad, CINVESTAV-IPN, Irapuato, Mexico

ARTICLE INFO

Article history:
Received 10 June 2011
Received in revised form 10 August 2011
Accepted 26 August 2011
Available online 2 September 2011

Keywords:
NUMTs
Insertions
Heteroplasmy
Cancer

ABSTRACT

Nuclear insertions of mitochondrial origin (NUMTs) can be useful tools in evolution and population studies. However, due to their similarity to mitochondrial DNA (mtDNA), NUMTs may also be a source of contamination in mtDNA studies. The main goal of this work is to present a database of NUMTs, based on the latest version of the human genome—GRCh37 draft. A total of 755 insertions were identified. There are 33 paralogous sequences with over 80% sequence similarity and of a greater length than 500 bp. The non-identical positions between paralogous sequences are listed for the first time. As an application example, the described database is used to evaluate the impact of NUMT contamination in cancer studies. The evaluation reveals that 220 positions from 256 with zero hits in the current mtDNA phylogeny could in fact be traced to one or more nuclear insertions of mtDNA. This is due to they are located in non-identical positions between mtDNA and nuclear DNA (nDNA). After *in silico* primer validation of each revised cancer study, risk of co-amplification between mtDNA and nDNA was detected in some cases, whereas in others no risk of amplification was identified. This approach to cancer studies clearly proves the potential of our NUMT database as a valuable new tool to validate mtDNA mutations described in different contexts. Moreover, due to the amount of information provided for each nuclear insertion, this database should play an important role in designing evolutionary, phylogenetic and epidemiological studies.

1. Introduction

The evolution of eukaryotic cells is linked to the phenomena of endosymbiosis and it is widely accepted that DNA transfer occurs between the cellular organelles (mitochondria and chloroplasts) and the nucleus. The general evolutionary tendency is a reduction in the gene content of cellular organelles to avoid genetic redundancy (Kleine et al., 2009). Although the origin of this transfer is unknown (Blanchard and Lynch, 2000), it is accepted that it was an important evolutionary mechanism in the prokaryotic–eukaryotic transition and it appears that it was in the early endosymbiosis when this transfer was the most significant. DNA transfer from cellular organelle to the nucleus is a process that remains active (Hazkani-Covo et al., 2010; Hazkani-Covo and Graur, 2007; Ricchetti et al., 2004) and has resulted in the formation of pseudogenes, known as NUMTs (nuclear insertions of mitochondrial origin) or NUPTs (nuclear insertions of chloroplast origin) depending on whether they are of mitochondrial or chloroplast origin (Leister, 2005). These insertions are non-uniformly distributed throughout the genomes and the patterns that enable prediction of the insertion position are currently unknown (Bensasson et al., 2001).

The most parsimonious explanation for the origin of specific NUMT integration is the non-homologous end joining (NHEJ) repair mechanism. NHEJ is the major mechanism of double-strand break repair (DSBR) in mammalian cells and during this process, NUMTs appear to prevent chromosomal deletions primarily through blunt-end repair (Hazkani-Covo and Covo, 2008).

De novo NUMT insertions have been described as being associated to diseases and to the aging process (Caro et al., 2010; Goldin et al., 2004; Turner et al., 2003). However, most reported NUMTs are polymorphic or fixed into the species.

Since the mutation rate is higher in mitochondrial DNA (mtDNA) than in nuclear DNA (nDNA), once an insertion is fixed, it undergoes a different evolutionary process to that of the original mtDNA sequence.

* Corresponding author at: Unitat d'Antropologia Biològica, Departament Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain. Tel.: +34 935811503; fax: +34 935811321.
E-mail address: cristina.santos@uab.cat (C. Santos).

1567-7249/$ – see front matter © 2011 Elsevier B.V. and Mitochondria Research Society. All rights reserved.
doi:10.1016/j.mito.2011.08.009

Thus, NUMTs are considered molecular fossils that provide information about the mtDNA ancestral state, allowing to root intra-specific mtDNA phylogenetic trees without needing an outgroup species. Moreover, they can be used as tools in the study of the evolutionary history of populations (Ricchetti et al., 2004; Thalmann et al., 2005; Turner et al., 2003) and in the establishment of molecular phylogenies (Hazkani-Covo, 2009; Hazkani-Covo and Graur, 2007; Jensen-Seaman et al., 2009).

Due to the similarity with mtDNA, NUMTs are a potential source of contamination in mtDNA studies based on PCR amplification (Parr et al., 2006b; Yao et al., 2008). Although this problem has previously been considered to be muted in consequence of the high copy number of mtDNA compared with the corresponding nuclear loci, caution is mandatory (Parfait et al., 1998) since amplification of NUMTs is a real problem for accurate sequence interpretation, and particularly for the interpretation of mtDNA heteroplasmy (Parr et al., 2006a, 2006b). Parr et al. (2006b), in a study performed with rho0 cells, suggested that one of the most important factors that determine whether a NUMT will or will not co-amplify with mtDNA are the region of the mtDNA targeted by the PCR and the number of NUMT copies. Moreover, samples of ancient DNA or any tissue with a low mtDNA copy number, in both normal and pathological states (such as, for example, tumor samples), also seem to be important factors that could determine the co-amplification of NUMTs (Goios et al., 2008; Parr et al., 2006b). Goios et al. (2008) stated that, in standard sequencing of samples used for population characterization, the amplification of NUMTs does not constitute a real problem, since the mtDNA content in samples is much higher than the nDNA content and the detection of mtDNA heteroplasmy is not a priority. Yao et al. (2008) demonstrated, nevertheless, that co-amplification of NUMTs can occur even when samples of a standard quality and amount are used.

From the analysis of several studies concerning the prevalence of NUMTs, a very different number of nuclear insertions of mitochondrial origin are reported. This discordance can be attributed to different factors: 1) after insertion of a NUMT, it becomes susceptible to suffering post-insertional processes, such as duplications, translocations, deletions, etc., that alter the initially inserted sequence (Hazkani-Covo et al., 2003, 2010); these post-insertional processes have not been equally considered in different studies; and 2) there are no unified criteria for the *in silico* detection of NUMTs. This lack of standardization affects both the selection of bioinformatic tools as well as the criteria for acceptance or not of the NUMT paralogous sequences (Lascaro et al., 2008). On the other hand, the existing NUMT compilations were based on a previous version of the Human Genome Reference Sequence, instead of the current version (GRCh37) and an update is required. This update would be useful for population and phylogenetic studies and the prevention of mtDNA and nDNA co-amplification.

The main goal of this study is to present a comprehensive database of NUMTs, based on the latest version of the human genome — GRCh37 draft. Moreover, as an application example, the described database is used to assess the impact of NUMT contamination on cancer studies.

2. Materials and methods

2.1. Updating of insertions of mitochondrial origin database

NUMT detection was performed *in silico*; the new version of the human genome draft (GRCh37) and the criteria for NUMTs detection proposed by Hazkani-Covo and Graur (2007) were used. The Basic Local Alignment Search Tool (BLAST), available from the NCBI (National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/BLAST/) (Altschul et al., 1990), was used to identify regions of similarity between mitochondrial and nuclear genomes. The Human mtDNA Reference Sequence (NC_012920) was compared against the human RefSeq Genomic database at NCBI. The similar nucleotide sequences were found using the RemoteBlast package in Bioperl bundle (Stajich et al., 2002) with the parameters set to restrict the search to human organism and the E-value of 10^{-3} (Hazkani-Covo and Graur, 2007). Moreover, regions with less than 20 bp were excluded. The BLAST report contains the basic information for each hit and the lists of the identical positions in the high scoring alignment pairs, both in the query and in the hit sequence. An *in-house* Perl script was written to further process this information. For each alignment pair, the array of sites listed in the query range was compared against the identical sites found. The same was carried out for positions in the hit range, and finally the sites of interest, namely the non-identical nucleotides, were extracted.

BLAST results were manually inspected to select only the hits obtained for the GRCh37 primary reference assembly and no post-insertional processes were taken into account.

2.2. Applications to cancer

MtDNA mutations previously described in cancer samples, and classified as having zero hits in the mtDNA phylogeny (that are therefore not polymorphic in human populations) by Santos et al. (2008) were used to evaluate the impact of NUMT contamination in cancer studies. First, mutations compiled by Santos et al. (2008) were re-evaluated using the updated mtDNA phylogeny – mit. Tree build 8 – (Van Oven and Kayser, 2009) and mutations that had one or more hits in the new phylogeny were not considered for subsequent analysis.

Mutations detected in cancer, with no hits in the mtDNA phylogeny, were then searched in the database of non-identical positions between the paralogous sequences. Furthermore, for those studies in which mutations coincided with non-identical positions between the paralogous sequences, an *in silico* validation of primers used in each study was performed. In brief, the PCR primers were submitted to BLAST. The specific Basic BLAST tool, optimized for highly similar sequences (Megablast), was performed using the Reference Genomic Sequence database for *Homo sapiens* (refseq_genomic: Genomic sequences from National Center for Biotechnology Information Reference Sequence Project). If both primers showed homology (expected number of chance matches in a random model – E-value – lower than 1) inside the same chromosome region they were selected as being susceptible to co-amplify mtDNA and nDNA.

3. Results

3.1. Database of nDNA sequences of mitochondrial origin based on the GRCh37 Human Genome Reference Sequence

The NUMT database based on the GRCh37 Human Genome Reference Sequence is reported in reported in Table I (Online Supplementary Material I). Additionally, a BLAST-searchable database of all NUMTs is also reported in Online Supplementary Material II.

Seven hundred and fifty-five insertions were found to be spread throughout the entire genome. For each insertion the following information is reported: location at the chromosome; length of the chromosome; access number; score and the E-value of the match; fraction and the percentage identity; number of gaps; total length of match (alignment mtDNA/nDNA); the effective length in nDNA and in the mtDNA; the number of identical positions; the region in the mtDNA and nDNA; and finally, matrix of the identical and non-identical positions in mtDNA and nDNA (Fig. 1).

The 755 insertions are distributed throughout the genome with variable frequencies and sizes for each chromosome (Table 1 and Fig. 2). A large number of insertions are observed in most chromosomes. There are only six chromosomes (14, 15, 16, 18, 20 and Y) that accumulate 15 or fewer insertions, representing less than 2% of the total insertions present in each.

There is a positive correlation between the number of insertions and chromosome size (Spearman correlation: 0.822 $p<0.001$) (Fig. 3a). Chromosome 2, the second largest chromosome in relation to the entire