Universitat Autònoma de Barcelona

Unitat d'Antropologia Biològica
Departament de Biologia Animal Biologia Vegetal i Ecologia

Freqüència i Patró de l'Heteroplàsmia Mitocondrial Humana

Amanda Ramos Reche

Tesi Doctoral

2012

Treball finançat pel Ministeri d'Educació i Ciència
(Projecte: CGL2006-07374, CGL2008-03408-E i
CGL2009-08205)

Freqüència i Patró de l'Heteroplàsmia Mitocondrial Humana

Memòria presentada per Amanda Ramos Reche per a optar al títol de Doctor en Biologia, programa de doctorat en Biodiversitat del Departament de Biologia Animal, Biologia Vegetal i Ecologia de la Universitat Autònoma de Barcelona, dirigida per les professores:

- Dra. Mª Pilar Aluja Paris, Professora Titular de la Unitat d'Antropologia Biològica del Departament de Biologia Animal, Biologia Vegetal i Ecologia de la Universitat Autònoma de Barcelona.

- Dra. Cristina Pereira Santos, Professora Lector de la Unitat d'Antropologia Biològica del Departament de Biologia Animal, Biologia Vegetal i Ecologia de la Universitat Autònoma de Barcelona.

Dra. Mª Pilar Aluja Paris   Dra. Cristina Pereira Santos

Amanda Ramos Reche

I wanted to change the world.
But I have found that the only thing one
can be sure of changing is oneself.

Aldous Huxley

"Te recuerdo Amanda,
la calle mojada,
corriendo a la fábrica
donde trabajaba Manuel.
La sonrisa ancha, la lluvia en el pelo,
no importaba nada, ibas a encontrarte con él,
con él, con él, con él, con él.

Son cinco minutos.
La vida es eterna en cinco minutos.
Suena la sirena de vuelta al trabajo,
y tú caminando, lo iluminas todo.
Los cinco minutos te hacen florecer

Que partió a la sierra.
Que nunca hizo daño. Que partió a la sierra,
y en cinco minutos quedó destrozado.
Suena la sirena, de vuelta al trabajo.
Muchos no volvieron, tampoco Manuel"

Víctor Jara.

A mi padre, por haber creído en mí, siempre,

A tu, Pere, pel teu suport i ajuda incondicionals

RESUM

La investigació desenvolupada en la present Tesi està enfocada a l'estudi de l'heteroplàsmia mitocondrial humana. Els treballs d'heteroplàsmia mitocondrial que s'han publicat fins al moment d'inici de la Tesi s'han centrat en la regió no codificant del DNA mitocondrial (DNAmt) i no hi ha cap estudi on s'utilitzin seqüències completes de DNAmt. Per detectar correctament l'heteroplàsmia mitocondrial, cal tenir en compte la presència dels anomenats NUMTs (insercions nuclears d'origen mitocondrial). A causa de l'elevada similitud entre NUMTs i DNAmt, una co-amplificació entre ells podria comprometre la veracitat dels resultats. Actualment, no existeix en la bibliografia cap treball on s'utilitzin *primers* que evitin la co-amplificació amb NUMTs. D'altra banda, les bases de dades de NUMTs descrites fins el moment presenten una gran discrepància en el nombre total de NUMTs i han estat realitzades amb una versió antiga de l'esborrany del genoma humà.

Els objectius de la Tesi són els següents: 1) establir una metodologia adequada per a la correcta amplificació i seqüenciació del genoma mitocondrial humà que eviti la co-amplificació amb

NUMTs; 2) realitzar una exhaustiva compilació de les insercions nuclears d'origen mitocondrial basada en la darrera versió de l'esborrany del genoma humà (GRCh37); 3) avaluar l'impacte de la contaminació per NUMTs en estudis de càncer com un exemple d'aplicació de la base de dades descrita; 4) determinar la freqüència i el patró de l'heteroplàsmia en la totalitat del genoma mitocondrial; i 5) comparar el patró mutacional de l'heteroplàsmia mitocondrial amb l'observat a nivell poblacional.

La metodologia així com la presentació dels resultats i la discussió s'han organitzat en 3 capítols dirigits a respondre els diversos objectius. En el primer capítol s'ha dut a terme el disseny i la validació d'un conjunt de *primers* que permet l'amplificació i seqüenciació completa de genoma mitocondrial. A més, aquests *primers* estan dissenyats de manera que eviten la co-amplificació amb DNA nuclear (DNAn). Finalment, i com a conseqüència de l'actualització de l'esborrany del genoma humà, s'han re-avaluat aquests *primers*, havent-se de redissenyar una de les parelles. En el segon capítol es presenta la compilació de les 755 insercions nuclears d'origen mitocondrial realitzada tenint en compte el darrer esborrany del genoma humà GRCh37. En aquesta base de dades, es presenten per primera vegada les posicions no idèntiques entre el DNAmt i el DNAn. Aquestes posicions han estat la base per l'aplicació de l'exemple que s'ha realitzat en estudis de càncer, avaluant així l'impacte de la contaminació per NUMTs. En el tercer i darrer capítol es presenta l'estudi exhaustiu sobre la freqüència i patró de l'heteroplàsmia mitocondrial humana mitjançant la seqüenciació completa del genoma mitocondrial de 101 individus.

Alhora aquesta informació s'ha utilitzat per realitzar una comparativa amb l'espectre mutacional a nivell poblacional.

Aquest és el primer treball on es descriuen per primera vegada un conjunt d'eines essencials pel correcte estudi de l'heteroplàsmia mitocondrial humana: es presenta un conjunt de *primers* que permeten amplificar i seqüenciar selectivament la totalitat del genoma mitocondrial i s'ha generat una base de dades d'insercions nuclears d'origen mitocondrial basada en el nou esborrany del genoma humà GRCh37, reportant alhora les posicions no-idèntiques entre el DNAmt i els NUMTs.

És la primera vegada on s'analitza la freqüència i el patró de l'heteroplàsmia mitocondrial humana en individus sans. El 62.37% de la població presenta heteroplàsmia mitocondrial (de longitud i/o puntual) i un 24.75% d'individus presenta heteroplàsmia puntual, posant de manifest l'elevada freqüència d'heteroplàsmia mitocondrial a nivell poblacional. La distribució de l'heteroplàsmia mitocondrial comprèn tot el genoma mitocondrial, localitzant-se tant en posicions definides com a *hotspot* com en posicions altament estables i conservades. Comparant el patró mutacional de l'heteroplàsmia amb el descrit a nivell poblacional, sembla haver-hi un excés d'heteroplàsmia puntual localitzada en posicions que presenten zero *hits* en la filogènia. L'excés de transicions detectat en l'espectre mutacional de l'heteroplàsmia, en comparació amb el descrit a nivell poblacional, implica que determinades forces evolutives com la deriva o la selecció han hagut d'actuar perquè finalment siguin eliminades. En aquest context, les heteroplàsmies puntuals localitzades en posicions estables podrien tenir un gran

impacte en la supervivència de la pròpia mitocòndria, suggerint que la selecció purificadora hauria d'estar actuant per prevenir la seva fixació en els individus.

Aquest és el primer treball on s'analitza la freqüència i el patró de l'heteroplàsmia mitocondrial humana en individus sans. Tenint en compte l'elevada quantitat de dades que es presenta unit a la poca informació prèvia disponible, aquest treball proporciona una nova perspectiva per a l'estudi de les malalties associades al DNAmt així com en treballs a nivell poblacional, evolutiu i forense.

ABSTRACT

The investigation developed in this thesis is focused on the study of human mitochondrial heteroplasmy. Until the beginning of the present thesis, the heteroplasmy works published have been focused on the analysis of non-coding region, and no studies on determining the levels of heteroplasmy using complete mitochondrial DNA (mtDNA) sequences has been published. To properly detect the mitochondrial heteroplasmy, the presence of NUMTs (nuclear insertions of mitochondrial origin) has to be taken into account. Due to similarity between NUMTs and mtDNA, a co-amplification between them could compromise the veracity of the sequence results. Currently, the published primers to amplify the whole mtDNA do not prevent the co-amplification with NUMTs. Moreover, the NUMT databases described until now show discrepancies in the total number of NUMTs and all of them have been carried out with an older version of the Human Genome Reference Sequence.

The main goals of the thesis are: 1) to establish an appropriate methodology for the correct amplification and sequencing of the human mitochondrial genome that avoid the co-amplification with

NUMTs; 2) to perform an exhaustive compilation of nuclear insertions of mitochondrial origin based on latest version of the Human Genome Reference Sequence (GRCh37); 3) to evaluate the impact of NUMTs contamination on cancer studies as an application example of the described database; 4) to determine the frequency and pattern of heteroplasmy in the whole mitochondrial genome; and 5) to compare the heteroplasmic mutational spectrum with the observed at population level in the whole mtDNA.

The methodology used, the presentation of results and the discussion are organized into 3 chapters directed to answer the objectives. In the first chapter, the design and validation of the set of primers that allows a selective amplification and sequencing of complete mitochondrial genome has been performed. Moreover, these primers have been designed to avoid co-amplification with nuclear DNA (DNAn). Finally, and as a result of the updated Human Genome Reference Sequence, primer pairs have been re-evaluated and one of them had to be redesigned. The second chapter presents the compilation of the 755 nuclear insertions of mitochondrial origin performed with the last Human Genome Reference Sequence GRCh37. In this database, we present for the first time the non-identical positions between DNAmt and DNAn. These positions have been the basis for the evaluation of the impact of NUMTs contamination on cancer studies in the application example. In the third and final chapter is presented an exhaustive study of the frequency and pattern of the human mitochondrial heteroplasmy using the complete sequencing of the mitochondrial genome of 101

individuals. Moreover, this information has been used to perform a comparison with the mutational spectrum at the population level.

This is the first study that describes essential tools for the correct study of the human mitochondrial heteroplasmy: a set of primes that can selectively amplify and sequence the entire mitochondrial genome and a database of nuclear insertions of mitochondrial origin based on the last Human Genome Reference Sequence GRCh37, reporting the non-identical positions between mtDNA and NUMTs.

For the first time the analysis of the frequency and pattern of the human mitochondrial heteroplasmy in healthy individuals is reported. The results show that the 62.37% of the individuals are heteroplasmic (length and/or point) and 24.75% of individuals presented point heteroplasmy, highlighting the high frequency of mitochondrial heteroplasmy at population level. The distribution of mitochondrial heteroplasmy covers the whole mitochondrial genome, finding them in hotpots as well as in highly stable and conserved positions. Comparing the presented results with the mutational spectrum at population level, it seems that there is an excess of point heteroplasmy located in positions that have zero hits in the mitochondrial phylogeny. The excess of transitions detected in the heteroplasmic mutational spectrum, in comparison with the described at the population level, implies that some evolutionary forces may be acting to finally remove them. In fact, highly stable heteroplasmic positions could have a greater impact in the viability of mitochondrial survival, suggesting that purifying selection must be operating to prevent their fixation within individuals.

This is the first study that analyses the frequency of heteroplasmy in healthy population, carrying out an evolutionary evaluation of the detected changes. This work provides a new perspective which would be important for medical, evolutionary, and forensic proposes.

ÍNDEX

	Pàgina
PRÒLEG	**1**
1. INTRODUCCIÓ	**5**
1.1 CARACTERÍSTIQUES DEL DNA MITOCONDRIAL HUMÀ	**7**
1.1.1 Elevat nombre de còpies	13
1.1.2 Herència materna	13
1.1.3 Pèrdua de recombinació	15
1.1.4 Taxa de mutació	17
1.2 ORIGEN EVOLUTIU DE LA MITOCÒNDRIA	**21**
1.2.1 Insercions nuclears d'origen mitocondrial (NUMTs)	24
1.3 HETEROPLÀSMIA MITOCONDRIAL	**29**
1.3.1 Metodologies aplicades a l'estudi de l'heteroplàsmia mitocondrial	35
2. JUSTIFICACIÓ DEL TREBALLL I OBJECTIUS	**41**
2.1 JUSTIFICACIÓ DEL TREBALL	**43**
2.2 OBJECTIUS	**47**
3. METODOLOGIA, RESULTATS I DISCUSSIÓ	**49**
3.1 CAPÍTOL 1: Primer design and validation for mtDNA amplification and sequencing	**51**

3.1.1 Materials and methods 55

Primer design and validation *55*

Total DNA extraction, amplification and sequencing of the entire mtDNA *56*

Cloning *57*

3.1.2 Results 58

Primer design and validation *58*

PCR and sequencing primers to mtDNA analysis *61*

3.1.3 Discussion 65

3.1.3.1 Human mtDNA complete amplification and sequencing: a new validated primer set that prevents NUMTs co-amplification 65

3.1.3.2 Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: a revision based on the new human genome reference sequence (GRCh37) 67

3.2 CAPÍTOL 2: Updating of NUMTs database and usefulness in cancer studies **71**

3.2.1 Materials and methods 75

Updating of insertions of mitochondrial origin database *75*

Applications to Cancer *76*

3.2.2 Results 78

Database of nDNA sequences of mitochondrial origin based on the GRCh37 Human Genome Reference Sequence *78*

Applications to cancer *87*

3.2.3 Discussion 90

3.3 CAPÍTOL 3: Frequency and pattern of human mitochondrial heteroplasmy **95**
3.3.1 Materials and methods 97
Samples and DNA Extraction *97*
MtDNA Analysis *98*
Detection and Authentication of Heteroplasmy *98*
Data analysis *100*
3.3.2 Results 104
Frequency and type of heteroplasmy *104*
Pattern and distribution of point heteroplasmy along the mtDNA genome *107*
Stability of point heteroplasmic positions *111*
3.3.3 Discussion 119
Frequency and type of heteroplasmy *119*
Pattern and stability of point heteroplasmic positions along the mtDNA genome *123*
4. DISCUSSIÓ GENERAL **129**
Disseny i validació de primers per a l'amplificació i seqüenciació del DNAmt ***131***
Actualització de la base de dades de NUMTs i aplicació en estudis de càncer ***134***
Freqüència i patró de l'heteroplàsmia mitocondrial humana ***138***
5. CONCLUSIONS **147**
6. BIBLIOGRAFIA **153**
7. ANNEX **183**
7.1 Annex I **185**
7.2 Annex II **271**

PRÒLEG

El present estudi pretén ampliar significativament la informació que es té actualment de l'heteroplàsmia mitocondrial humana, aportant informació desconeguda fins el moment, que podrà aplicar-se en camps de la investigació tan diversos com l'evolutiu, el poblacional, així com els estudis forenses i genètics.

Actualment, hi ha una gran varietat de treballs sobre l'estudi del DNA mitocondrial (DNAmt), tant per la seva relació amb malalties com en la caracterització d'individus sans. Pel que fa als estudis poblacionals, la majoria dels treballs publicats fins el moment s'han centrat en la regió no codificant del genoma mitocondrial, que comprèn només el 7% del genoma mitocondrial. Val a dir que, el ràpid desenvolupament de les tècniques de seqüenciació automàtica ha permès ampliar aquests estudis a la resta del genoma mitocondrial. No obstant, pel que fa a l'estudi de l'heteroplàsmia mitocondrial, en el moment d'inici de la present Tesi Doctoral, no hi havia cap treball publicat centrat en l'anàlisi de l'heteroplàsmia que utilitzés seqüències completes de DNAmt.

Per identificar correctament l'heteroplàsmia mitocondrial, cal tenir en compte la presència dels anomenats NUMTs (insercions nuclears d'origen mitocondrial), ja que l'elevada similitud entre NUMTs i DNAmt, pot induir una co-amplificació entre ells i comprometre la correcta identificació de l'heteroplàsmia mitocondrial. Fins a la realització d'aquesta Tesi, no s'ha publicat cap treball on tots els *primers* utilitzats evitessin la co-amplificació entre el DNAmt i els NUMTs. Tanmateix, actualment existeixen diverses bases de dades de NUMTs, però totes elles s'han realitzat amb una versió antiga de l'esborrany del genoma humà.

La present Tesi Doctoral aborda aquestes mancances i s'inicia amb una introducció on es destaquen les principals característiques del DNAmt; es detalla l'origen evolutiu de la mitocòndria i es descriuen les característiques de les insercions nuclears d'origen mitocondrial; i finalment es fa una exhaustiva descripció de l'heteroplàsmia mitocondrial, recopilant els tipus de treballs duts a terme fins ara, així com les metodologies que s'han d'aplicar per la correcta detecció i interpretació d'aquesta. A continuació, s'exposa la justificació del treball i s'estableixen els objectius principals de la Tesi Doctoral. La metodologia, els resultats i la discussió s'estructuren en tres capítols adreçats a respondre als diferents objectius plantejats. Aquest capítols estan escrits en anglès per facilitar la difusió de la Tesi. En el primer capítol es descriu el disseny i validació d'un nou set de *primers* per a l'amplificació i seqüenciació completa del DNAmt humà, evitant alhora la co-amplificació amb NUMTs; així com la revisió d'aquest set de *primers* degut a l'actualització del nou esborrany del genoma humà

(GRCh37). El segon capítol presenta una nova base de dades d'insercions nuclears d'origen mitocondrial basada en darrer esborrany del genoma humà (GRCh37) i la seva aplicació en estudis de càncer. En el darrer capítol es descriu l'estudi de la freqüència i el patró de l'heteroplàsmia mitocondrial humana. Seguidament s'exposa la discussió general, organitzada en funció dels objectius plantejats i, finalment, es presenten les conclusions de la Tesi Doctoral.

1. INTRODUCCIÓ

" All these mitochondrial DNAs stem from one
woman who is postulated to have lived about 200,000 years ago,
probably in Africa."
Rebecca Cann, Mark Stoneking & Allan Wilson

1.1 CARACTERÍSTIQUES DEL DNA MITOCONDRIAL HUMÀ

El DNA mitocondrial (DNAmt) humà és una molècula circular de 16569 parells de bases (pb) de longitud. És un genoma de doble cadena, amb una cadena pesada H (*Heavy Strand*), rica en guanina i una cadena lleugera L (*Light Strand*) rica en citosina.

Aquest genoma es troba a les mitocòndries, els orgànuls responsables de la respiració cel·lular, que generen la major part dels subministraments d'adenosina trifosfat (ATP) que necessita la cèl·lula.

Una revisió de Cowdry (1918), citat per Lehninger (1965), conté més d'un dotzena de termes per referir-se a les estructures que ara es coneixen com a mitocòndries: lefaroplast, condricont,

condriòmits, condrioplasts, condriosomes, condriosferes, fila, grànuls fucsinofílics, Korner, Fadenkörper, mitogel, cossos parabasals, vermícules, sarcosomes, cossos intersticials, plasmosomes, plastocondris i bioblasts, entre d'altres. Més tard, l'any 1948, amb el estudis de Hogeboon, Schneider i Palade, i gràcies a la utilització del microscopi electrònic, es confirmà que les mitocòndries presenten dues membranes: una membrana externa llisa i una membrana interna plegada formant les crestes mitocondrials (Hogeboon et al. 1948). Així s'establí definitivament la mitocòndria com el lloc on es duu a terme la respiració cel·lular.

El nombre de mitocòndries per cèl·lula és un paràmetre que sembla variar significativament d'un tipus de cèl·lula a un altre, sobretot per la importància de la biogènesi mitocondrial en el context de la diferenciació i especialització cel·lulars. Estimacions a partir de seccions seriades de cèl·lules donen valors d'entre uns pocs centenars fins a milers per cèl·lula, per exemple, unes 800 mitocòndries en el cas dels hepatòcits. S'estima que l'oòcit humà conté més de 100.000 mitocòndries, mentre que els espermatozoides en tenen relativament poques (tot i que aquestes poden ser excepcionalment grans) (Scheffler 2008).

Cada mitocòndria pot presentar múltiples còpies de DNAmt, de manera que cada cèl·lula pot contenir-ne milers. El primer informe definitiu que afirmava la presència de DNA en aquest orgànul va aparèixer l'any 1960, però es considerava tan especialitzat i excepcional que la troballa no va ser generalitzada a totes les mitocòndries (Steinert and Steinert 1960).

Donada la reduïda mida del genoma mitocondrial, és evident que només un nombre limitat de gens hi podrien ser presents. Durant les tres darreres dècades, centenars de genomes mitocondrials d'animals, plantes i fongs s'han caracteritzat i fins i tot seqüenciat genèticament. La primera seqüenciació completa d'aquest genoma va ser la dels éssers humans (Anderson et al. 1981), i poc després va completar-se la seqüència bovina (Anderson et al. 1982).

L'estimació que es va realitzar sobre el nombre de polipèptids diferents en cada mitocòndria va situar-se per sobre dels mil (Schatz 1995), però les anàlisis proteòmiques més recents han identificat entre 500 i 800 proteïnes (Da Cruz et al. 2005; Forner et al. 2006; Gabaldon and Huynen 2004; Johnson et al. 2007; Prokisch et al. 2006; Vo and Palsson 2007). Això vol dir que, uns 1000 gens nuclears contribueixen a la biogènesi i funció de les mitocòndries. Molts d'ells són coneguts, però encara n'hi ha que no s'han identificat i caracteritzat.

El DNAmt dels metazous codifica per dos RNAs ribosomals (rRNA) (12S i 16S), 22 RNAs de transferència (tRNAs) (amb excepcions en algunes espècies), tres gens de les subunitats del citocrom oxidasa (COXI-COXIII), el gen del citocrom b (Cyt B), set gens de les subunitats del NAD deshidrogenasa (N1-N6, N4L), dos gens de les subunitats del ATPasa (6 i 8) i el gen del citocrom b (Scheffler 2008) (Figura 1).

El DNAmt també inclou una regió no codificant o D-loop que conté l'origen de replicació de la cadena pesada i les regions promotores de la transcripció en direccions oposades. Per aquestes

raons, es coneix com a "regió control". La seva longitud és de 1122 pb en humans i, degut a la seva identitat funcional, aquesta regió no està ben conservada entre espècies, sent polimòrfica fins i tot dins d'una sola espècie com en els éssers humans. Tot i així, en la regió control hi ha petites zones funcionalment significatives i segments conservats. Podríem establir doncs, que la regió del D-loop és funcionalment conservada en tots els vertebrats, però les seqüències de nucleòtids que la composen no ho són (Scheffler 2008).

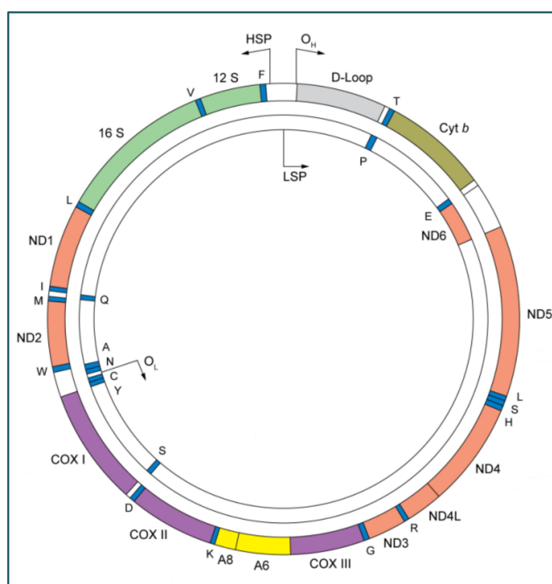


Figura 1. Representació esquemàtica del genoma mitocondrial humà amb les seves regions funcionals: els dos rRNAs (12S i 16S), tres gens de les subunitats del citocrom oxidasa (COXI-COXIII), set gens de les subunitats del NAD deshidrogenasa (N1-N6, N4L), dos gens de les subunitats del ATPasa (6 i 8), el gen del citocrom b, 22 tRNAs (amb el codi estàndard de lletra única), dos orígens de replicació de les cadenes lleugera i pesada (O_H i O_L), els promotors de la transcripció tant de la cadena pesada com de la lleugera (HSP i LSP) i la regió control. Figura adaptada de DiMauro i Schon (2008).

L'any 1987, Cann, Stoneking i Wilson (Cann et al. 1987) van publicar un treball sobre l'anàlisi del DNAmt mitjançant fragments de restricció de 147 individus provinents de cinc poblacions geogràficament espaiades (34 asiàtics, 21 aborígens australians, 26 aborígens de Nova Guinea, 46 caucàsics i 20 africans). L'impacte científic d'aquest treball va ser enorme. L'aplicació de mètodes de genètica molecular en l'estudi de l'evolució de primats i humans, amb especial atenció al DNAmt, ja havia estat suggerida per Brown (1981), i diverses publicacions de Brown, Wilson i els seus col·legues l'havien precedit a la publicació de referència del 1987 (Brown et al. 1982; Cann et al. 1982; Cann et al. 1984). No obstant això, el gran nombre de persones representades en aquest estudi va permetre obtenir proves convincents de la força potencial d'aquest enfocament en el tractament de preguntes sobre l'evolució humana, sobre els orígens dels éssers humans moderns i sobre els patrons migratoris dels nostres avantpassats (Cann et al. 1987).

En les últimes dues dècades s'han dut a terme anàlisis per enzims de restricció abastament refinats, s'ha aplicat la reacció en cadena de la polimerasa (PCR), s'han seqüenciat regions concretes i, fins i tot s'han analitzat seqüencies de DNAmt humà de les zones més remotes del globus. Darrerament, degut a la introducció de les tècniques de seqüenciació d'última generació (plataformes 454 de Roche o Illumina de Solexa, entre d'altres) s'estan realitzant seqüenciacions massives en una gran diversitat de treballs que ampliaran enormement el volum de seqüencies completes de DNAmt per a la seva utilització en futurs estudis [per exemple Ansorge (2009), Briggs et al. (2009) i Metzker et al. (2010)].

El DNAmt ha estat la regió més estudiada dels genomes eucariotes i ha tingut un paper fonamental en el desenvolupament de l'estudi de les poblacions i la genètica evolutiva (Avise 1991; Avise et al. 1987; Brown 1985; Harrison 1989; Moritz et al. 1987; Rand 1994). Durant la primera dècada dels estudis, en el camp de la biologia evolutiva, va assumir-se que el DNAmt actuava de manera neutra com a marcador genètic. Més recentment, ha anat creixent l'interès de com la selecció podria estar actuant en el DNAmt, i dos enfocaments diferents s'han postulat per abordar aquesta qüestió. Anàlisis directes, mitjançant l'estudi de diferents poblacions, han intentat identificar els efectes de la *fitness* o supervivència en els diferents haplotips (Clark and Lyckegaard 1988; Fos et al. 1990; Garcia-Martinez et al. 1998; Hutter and Rand 1995; Kambhampati et al. 1992; Kilpatrick and Rand 1995; MacRae and Anderson 1988; Nigro and Prout 1990). D'altra banda, s'han buscat evidències de selecció en el DNAmt realitzant anàlisis estadístiques de models neutres d'evolució molecular basant-se en seqüències de DNA de poblacions naturals (Ballard 2000; Ballard and Kreitman 1994; Blouin 2000; Excoffier 1990; Nachman 1998; Nachman et al. 1994; Nachman et al. 1996; Quesada et al. 1999; Rand et al. 1994; Rand and Kann 1996; Rand and Kann 1998; Rand et al. 2000; Templeton 1996; Weinreich and Rand 2000; Whittam et al. 1986; Wise et al. 1998). Tots dos enfocaments han identificat una evolució del DNAmt no neutre, implicant l'actuació de la selecció a nivell individual. No obstant això, un augment ràpid de la literatura sobre biologia cel·lular i genètica molecular ha demostrat la importància de la selecció i la deriva genètica dins i entre citoplasmes com a mecanisme primari, afectant a l'expressió fenotípica de les mutacions

en el DNAmt (Pereira et al. 2009; Ruiz-Pesini et al. 2004; Ruiz-Pesini and Wallace 2006).

L'ús generalitzat d'aquest genoma es deu a les característiques úniques del DNAmt, que el fan especialment susceptible als estudis evolutius. Aquestes característiques inclouen un alt nombre de còpies, l'herència materna, la manca de recombinació i una taxa de mutació més alta que la que es troba en el DNA nuclear (DNAn).

1.1.1 Elevat nombre de còpies

El DNAmt és present en un gran nombre de còpies en les cèl·lules humanes. La mitjana de cèl·lules somàtiques només té dues còpies d'un determinat gen o segment de DNAn, però centenars de milers de còpies de DNAmt (Robin and Wong 1988). Aquesta propietat, juntament amb el fet de localitzar-se fora del nucli, fa que sigui més fàcil d'obtenir, sent la molècula escollida per a l'anàlisi en estudis de DNA antic i per certes aplicacions forenses. Però a la vegada, aquesta propietat també complica la genètica de poblacions del DNAmt perquè hi ha diferents nivells en els quals les poblacions de molècules de DNAmt poden ser definides -dins del mitocondri, dins d'una cèl·lula, dins d'un determinat teixit, dins d'un individu, i dins d'un grup de persones (la tradicional definició d'una població).

1.1.2 Herència materna

Encara que l'herència paterna del DNAmt es produeix en musclos (Zouros et al. 1992), i s'ha observat tant de manera inter

com intraespecífica en Drosophila, ratolins, i híbrids d'aus (Gyllensten et al. 1991; Kaneda et al. 1995; Kondo et al. 1990; Kvist et al. 2003), des de fa anys, l'herència estrictament materna del DNAmt humà ha estat considerada com un dogma indestructible (Stoneking 1993; Stoneking and Soodyall 1996; Wallace et al. 1999). Aquesta herència uniparental és un dels grans avantatges del DNAmt, ja que permet als investigadors traçar llinatges relacionats a través del temps, destacant l'ascendència materna d'una població, sense els efectes de confusió de l'herència biparental i la recombinació inherent al DNAn. No obstant però, l'any 2002 va fer-se públic l'informe d'un home amb intolerància severa a l'exercici, on el DNAmt del teixit muscular era predominantment d'origen patern, fet que va provocar certs dubtes sobre la validesa de la hipòtesi de l'estricta herència materna del DNAmt (Schwartz and Vissing 2002). Això provocà que s'alcessin veus advertint sobre la prudència a l'hora de fer inferències sobre les poblacions humanes i de la història basada en el supòsit ara qüestionable de l'herència estrictament materna (Bromham et al. 2003). Investigacions posteriors amb un major nombre de pacients amb miopaties mitocondrials no han reportat més casos d'herència paterna (Filosto et al. 2003; Schwartz and Vissing 2004; Taylor et al. 2003). Es coneix des de fa anys, que les mitocòndries de l'esperma són selectivament destruïdes a l'oòcit (Manfredi et al. 1997; Shitara et al. 1998), i s'ha demostrat que el DNAmt patern és destruït en els oòcits per ubiquitinització (Sutovsky et al. 1999; Sutovsky et al. 2000). Per tant, sembla molt probable que el cas d'herència paterna del DNAmt fos conseqüència d'una ruptura en el reconeixement normal i eliminació de les molècules d'aquest DNAmt patern. D'altra banda,

milers de comparacions mare-fill no han demostrat cap indici de contribució paterna (Giles et al. 1980; Howell et al. 2003; Jazin et al. 1998; Santos et al. 2008b; Santos et al. 2005b). Per tant, a l'actualitat, l'herència materna del DNAmt en els éssers humans encara es pot considerar com la norma general (Schwartz and Vissing 2003). Aquesta característica d'herència uniparental fa que tots els individus pertanyents a un mateix llinatge comparteixin una mateixa seqüència, el que confereix la possibilitat d'estudiar llinatges al llarg del temps, esclarint els ancestres materns de les poblacions (Pakendorf and Stoneking 2005).

1.1.3 Pèrdua de recombinació

Un altre principi de l'Antropologia Molecular, que es va veure sacsejat fa uns anys, és que el DNAmt no recombina. Això va ser considerat com un fet establert (Stoneking 1993; Stoneking and Soodyall 1996; Wallace et al. 1999) fins els anys 1999-2000, quan quatre publicacions reportaren evidències de recombinació en el DNAmt humà (Awadalla et al. 1999; Awadalla et al. 2000; Eyre-Walker et al. 1999; Hagelberg et al. 1999). Tres d'aquests estudis es van basar en anàlisis filogenètiques i estadístiques mitjançant seqüències de DNAmt. Els autors van demostrar l'evidència de recombinació amb l'argument d’un excés de llocs homoplàsics observats en els arbres filogenètics (Eyre-Walker et al. 1999), i la correlació del desequilibri de lligament amb la distància a través del genoma mitocondrial (Awadalla et al. 1999; Awadalla et al. 2000). El quart estudi va afirmar tenir proves directes de recombinació en el DNAmt d’individus de la Melanèsia (Hagelberg et al. 1999). No obstant això, posteriorment es va demostrar que els estudis

filogenètics i/o estadístics havien utilitzat dades errònies i/o mètodes estadístics qüestionables (Arctander 1999; Jorde and Bamshad 2000; Kumar et al. 2000; Macaulay et al. 1999; Merriweather and Kaestle 1999; Parsons and Irwin 2000), i el treball de Hagelberg et al. (1999) va haver-se de retractar (Hagelberg et al. 2000). Tres estudis sobre la correlació del desequilibri de lligament i la distància en grans conjunts de dades de seqüències completes de DNAmt (Elson et al. 2001; Ingman et al. 2000; Piganeau and Eyre-Walker 2004) no van trobar evidències de recombinació, encara que, de nou, va detectar-se un excés de posicions homoplàsiques (Piganeau and Eyre-Walker 2004) - això però, s'atribueix generalment a l’heterogènia taxa de mutació del DNAmt humà (Excoffier and Yang 1999; Howell and Smejkal 2000; Stoneking 2000). Fins el moment però, només ha estat publicat un cas de recombinació en el DNAmt humà en la única persona coneguda que conté tant DNAmt d’origen matern com patern (Kraytsberg et al. 2004; Schwartz and Vissing 2002). En aquest cas, la recombinació entre el DNAmt matern i patern es va produir en aproximadament el 0,7% del total de DNAmt en el teixit muscular del pacient. Aquesta troballa posa de manifest que la recombinació és possible, ja que les mitocòndries posseeixen una recombinasa funcional (Thyagarajan et al. 1996).

No obstant això, recentment s’ha demostrat que la fusió entre mitocòndries d’una mateixa cèl·lula proporcionaria protecció enfront a mutacions en el DNAmt mitjançant l’intercanvi genètic, fet que permetria als genomes mitocondrials amb diferents mutacions complementar-se entre ells, compensant així la càrrega mutacional (Nakada et al. 2001; Nakada et al. 2009). Això podria donar-se fins i tot dins d’una sola mitocòndria, sempre i quan fos entre nucleoides i

no entre molècules de DNAmt en el sí del mateix nucleoide, fet que limitaria la recombinació (Chen et al. 2010; Gilkerson et al. 2008). A més, les fuites de DNAmt patern són un fenomen molt rar, per tant la recombinació no sembla ser un procés fonamental (Hagelberg 2003; Slate and Gemmell 2004) - en absència de molècules de DNA en heteroplàsmia, els esdeveniments de recombinació es traduirien en DNAs mitocondrials que no diferirien dels originals.

1.1.4 Taxa de mutació

Per lògica es pot pensar que com més importància té el contingut genètic més lentament hauria d'actuar un canvi evolutiu, però tot i que la majoria del DNAmt conté gens indispensables per a la vida cel·lular, la taxa de mutació d'aquest genoma se situa per sobre de la dels gens nuclears en diversos ordres de magnitud (Nachman et al. 1996; Schriner et al. 2000).

La proximitat als efectes mutagènics dels radicals oxigènics generats per la fosforilació oxidativa i el fet de que aquest genoma no està associat a histones que podrien protegir-lo, són explicacions per aquest augment de la pressió mutagènica en el DNAmt (Gray et al. 1999; Shigenaga et al. 1994). De totes maneres, aquesta elevada taxa de mutació proporciona una considerable variació a nivell poblacional, ja que podria permetre a alguns individus sobreviure a sobtats canvis ambientals. Per tant, l'elevada taxa de mutació que es dóna en la línia germinal del DNAmt dels mamífers podria representar un compromís entre la generació de suficient variació adaptativa per a poder augmentar la probabilitat de que les espècies sobrevisquin a un canvi ambiental catastròfic, *versus* la supressió de la taxa de mutació del DNAmt suficientment com per a garantir que

l'acumulació de mutacions perjudicials al llarg dels llinatges materns no condueixi a l'extinció com a resultat del trinquet de Muller (Wallace 2007).

Tot i existir una gran controvèrsia pel que fa a l'acceptació d'una taxa de mutació que pugui ser utilitzada en diferents camps d'investigació, una estimació d'aquesta seria $0{,}017 \times 10^{-6}$ substitucions/posició/any pel que fa a tot el genoma, exceptuant la regió de control (Ingman et al. 2000). En les dues regions hipervariables (HVRI i HVRII) de la regió no codificant, la taxa és encara major, tot i que saber quant major és, és una qüestió que encara genera discrepàncies.

Aquesta gran controvèrsia pel que fa a la taxa de mutació del mtDNA, és deguda en gran mesura, al procediment que es realitza per calcular-la. Existeixen dues vessants clarament diferenciades: els mètodes basats en comparacions filogenètiques, ja estiguin fonamentades en comparacions interespecífiques o intraespecífiques, que donen estimacions de 0.0250-0.1770 substitucions/posició/milió d'anys pel que fa a la regió control, i els mètodes empírics d'observació directa de les mutacions del DNAmt en famílies o llargs llinatges, on es consideren totes les substitucions, que donen estimacions de 0.3791-1.7957 substitucions/posició/milió d'anys [per revisió veure Santos et al. (2005a; 2008b)]. Aquesta última sembla ser una taxa significativament més elevada que les obtingudes filogenèticament. A l'actualitat existeix un debat pel que fa a la incògnita de com la taxa reflecteix el "veritable" estat de coses, i com s'utilitza en estudis poblacionals (Heyer et al. 2001; Howell and Mackey 1997; Howell et al. 2003; Jazin et al. 1998; Macaulay et al. 1997; Paabo 1996; Parsons et al. 1997;

Sigurgardottir et al. 2000). Una bona reflexió al respecte és la suggerida per Santos et al (2008b; 2005b) que proposa una sèrie de factors a tenir en compte i que eliminen les discrepàncies, si s'apliquen les correccions necessàries, com per exemple la taxa d'heteroplàsmia (0.4332 substitucions/posició/milió d'anys).

*"The first cell fusion, precursor of fertilization,
may have been the result of cannibalism: a microorganism ate
another without digesting."*

Lynn Margulis

1.2 ORIGEN EVOLUTIU DE LA MITOCÒNDRIA

Poc després d'extreure, per primera vegada, mitocòndries de cèl·lules d'organismes superiors, fa uns 120 anys, ja es va expressar d'una forma més o menys explícita la idea de que estaven relacionades d'alguna manera amb els bacteris (Margulis 1981). Llavors, els pensaments sobre l'origen de les mitocòndries i els cloroplasts varen culminar en una declaració extremista de la Teoria Endosimbiòtica que deia: "la 'cèl·lula' eucariota són múltiples 'cèl·lules' procariotes" (Taylor 1974), i Lynn Margulis es va convertir en una divulgadora particularment eloqüent i contundent sobre les idees de la simbiosi en l'evolució cel·lular eucariota.

Formalment, és necessari almenys tenir en compte dues hipòtesis alternatives sobre l'origen evolutiu de les mitocòndries. Una d'elles proposa que les mitocòndries provenen de l'interior d'una

cèl·lula mitjançant un procés de compartimentació intracel·lular i especialització funcional. Quan es va descobrir el DNA en aquests orgànuls, calia afegir que la informació genètica d'un genoma nuclear s’havia distribuït entre aquests dos compartiments. La segona hipòtesi suggereix que les mitocòndries es van originar a partir d'una relació simbiòtica entre una proto-cèl·lula eucariota i un organisme procariota primitiu capaç de dur a terme la fosforilació oxidativa. Només la segona hipòtesi s'està considerant seriosament a l'actualitat (Scheffler 2008).

Una revisió crítica dels arguments a favor i en contra sobre la hipòtesi endosimbiont publicada per Gray i Doolittle l’any 1982 (Gray and Doolittle 1982) va fer un intent per definir el tipus de dades que provaven aquesta hipòtesi, i va concloure que "un origen endosimbiont dels mitocondris és més complex i molt menys creïble" que el cas d’un origen dels cloroplasts (plastidis) a partir d’eubacteris fotosintètics (cianobacteris). Durant la dècada de 1980 van començar a compilar-se una gran quantitat de dades addicionals que eren revisades periòdicament per Gray i altres, culminant l’any 1992 en una revisió exhaustiva i avaluada (Gray 1992). L'autor va arribar a la conclusió que re-avaluar hipòtesis alternatives no tenia sentit en vista de l'enorme quantitat de proves a favor de la hipòtesi endosimbiòtica.

El DNAmt codifica per proteïnes molt específiques i úniques, tot i que és evident que és un genoma massa petit per codificar la informació suficient perquè un organisme visqui per si sol. La pèrdua d'informació genètica de les mitocòndries primordials es va veure compensat per les funcions codificades pel nucli, però en aquest

moment s'hauria de distingir entre dues possibilitats. Una d'òbvia és que els gens proto-mitocondrials van ser transferits al nucli, havent-hi una evidència convincent de que aquests esdeveniments s'havien produït. També hauria de ser obvi però, que la simbiosi implica que els bacteris destinats a convertir-se en mitocòndria van haver d'envair una cèl·lula hoste, un proto-eucariota, que abans d'aquest esdeveniment evolutiu era igualment capaç d'una existència independent. La distinció entre les dues cèl·lules des de la visió més simple, és que el proto-eucariota ja tenia el seu DNA compartimentat en un nucli amb una membrana nuclear, mentre que el proto-mitocondri s'assemblava a un bacteri sense nucli. Tots dos havien de ser capaços de replicar i transcriure el seu DNA, així com de sintetitzar proteïnes i diversos lípids per a formar la membrana. El cicle de Krebs també l'haurien de poder fer totes dues. Per tant, l'associació inicial entre aquestes dues cèl·lules havia de conduir a una considerable i redundant informació genètica. Els gens redundants en la mitocòndria podrien simplement haver-se perdut, d'altres van ser transferits al nucli i només un nombre molt petit va romandre a la mitocòndria (Kleine et al. 2009).

Hi ha però, un límit establert en la reducció d'aquest genoma mitocondrial (Palmer 1997)? Està força acceptat que els pèptids que resten codificats en el genoma mitocondrial són extremadament hidrofòbics i no poden ser importats mitjançant la maquinària de translocació mitocondrial. D'altra banda, dos pèptids d'ancoratge hidrofòbic del complex II són codificats per gens nuclears en la majoria dels organismes examinats fins al moment. Els arguments teòrics s'han remès a la predicció que el genoma mitocondrial, en

última instància, necessita retenir només dos gens que codifiquen pèptids, a més dels RNAr i els RNAt (Claros 1995; Claros et al. 1995; Popot and de Vitry 1990). Els dos gens assenyalats són el Cyt B i el COXI, que codifiquen per pèptids amb més de tres o quatre segments transmembrana hidrofòbics. Alguns descobriments proporcionen evidències convincents de que la pèrdua completa dels gens funcionals del genoma mitocondrial va ocorre, de fet, abans de l'emergència dels animals, fa aproximadament 1000 milions d'anys (Palmer 1997). Una de les idees emergents és que la mitocòndria endosimbiont original va evolucionar de forma independent en les diverses espècies que habiten en aquests entorns, del que en resultà una col·lecció de variants d'orgànuls morfològica, genètica i funcionalment heterogènia, incloent mitocòndries aeròbiques, anaeròbiques, hidrogenosomes i mitosomes (Van der Giezen and Tovar 2005).

1.2.1 Insercions nuclears d'origen mitocondrial (NUMTs)

Tot i que hi ha evidències de que la transferència de material gènic funcional de les mitocòndries al nucli ha finalitzat, la transferència de DNA organular no funcional al nucli és un procés que continua actiu (Hazkani-Covo and Graur 2007; Hazkani-Covo et al. 2010; Ricchetti et al. 2004) i que té com a resultat la formació de pseudogens coneguts com a NUMTs o NUPTs, depenent de si tenen un origen mitocondrial o cloroplastídic (Leister 2005) (Figura 2).

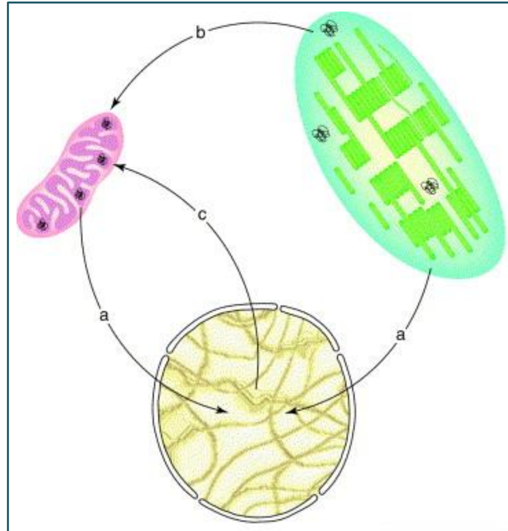


Figura 2. Visió esquemàtica dels diferents tipus de transferència de DNA intercompartimental. (a) Orgànul-nucli; (b) cloroplast→mitocondri; (c) nucli→mitocondri. Figura adaptada de Leister et al (2005).

Fins al moment, tots els NUMTs estudiats en mamífers han perdut la seva funció. Probablement a causa de les diferències entre els codis genètics nuclear i mitocondrial, aquestes insercions van passar immediatament a ser no funcionals quan van arribar al nucli (Hazkani-Covo et al. 2003).

Diversos estudis han demostrat que els NUMTs són similars però no idèntics a les seves contraparts en el DNAmt; varien en grandària i es poden repetir en tàndem (Tourmen et al. 2002). A més, els NUMTs es caracteritzen per la seva heterogeneïtat estructural i la desigual distribució en els cromosomes (Bensasson et al. 2001;

Tourmen et al. 2002). Alguns d'aquests NUMTs s'han integrat recentment al genoma nuclear mantenint la seva estructura. D'altres però, es van inserir fa molt temps i, posteriorment a la inserció, han patit processos com ara duplicacions, delecions, inversions o desplaçaments dins de la seqüència nuclear (Tourmen et al. 2002).

L'explicació més parsimoniosa que explica l'origen dels NUMTs és la reparació dels trencaments de doble cadena (DSBR) pel mecanisme d'unió d'extrems no homòlegs (NHEJ) (Hazkani-Covo and Covo 2008). Durant aquest procés de reparació, una seqüència de DNAmt serveix de pegat i preserva el DNAn de la pèrdua de nucleòtids. D'aquesta manera, el NUMT previndria la formació de delecions cromosòmiques (Hazkani-Covo and Covo 2008).

S'han descrit NUMTs en més de 80 espècies i el nombre d'aquestes seqüències és diferent entre elles (Bensasson et al. 2001; Hazkani-Covo and Graur 2007). Richly i Leister (2004) suggereixen que l'abundància variable dels NUMTs entre les espècies pot ser la causa de les diferències en la freqüència de la transferència de DNAmt al nucli o les distintes taxes de pèrdua de NUMTs. No obstant això, l'estimació del contingut i abundància de NUMTs encara continua incompleta (Bensasson et al. 2001).

L'anàlisi de NUMTs ha permès estudiar diversos mecanismes d'integració del DNAmt al nucli, així com conèixer les forces evolutives que actuen sobre les seqüències no codificants (Kleine et al. 2009). Aquest procés evolutiu ha estat àmpliament estudiat

mitjançant comparacions paràlogues dins dels genomes (Hazkani-Covo et al. 2003; Mourier et al. 2001; Woischnik and Moraes 2002).

Els NUMTs es classifiquen com ortòlegs, quan deriven d'un procés d'especiació i estan presents en el mateix locus al genoma de diferents espècies. Així doncs, podem considerar els NUMTs com fòssils moleculars que ens aporten informació sobre l'estat ancestral del DNAmt, alhora que ens permet arrelar els arbres filogenètics recents dins d'una mateixa espècie sense necessitat d'utilitzar una espècie *outgroup*. Si profunditzem en aquest ús dels NUMTs, podem comprendre fenòmens poblacionals i establir històries evolutives de les poblacions d'una mateixa espècie (Ricchetti et al. 2004; Thalmann et al. 2005). La taxa en que ocorren les insercions de NUMTs, la seva ortologia i l'estudi dels processos post-insercionals que pateixen, es perfilen com una bona eina per l'estudi filogenètic de les espècies (Hazkani-Covo and Graur 2007; Jensen-Seaman et al. 2009) i de fet, s'han construït filogènies de primats a partir de l'estudi dels NUMTs amb eines bioinformàtiques, concordants amb les consensuades fins al moment (Hazkani-Covo 2009). D'altra banda, els NUMTs poden classificar-se com a no-ortolòlegs quan deriven d'un procés de duplicació i estan presents en un sol genoma. En aquest cas, es classifiquen en insercions, delecions parcials o totals i en duplicacions en tàndem (Hazkani-Covo and Graur 2007). En aquest context, es coneix que la duplicació de NUMTs ja establerts en el genoma nuclear, és un procés que contribueix a un augment substancial en el nombre d'insercions (Richly and Leister 2004).

Diversos estudis publicats fins ara, mostren discrepàncies significatives en el nombre total de NUMTs presents en el genoma humà. Aquest nombre oscil·la entre 190 (Lascaro et al. 2008) i 1105 (Tourmen et al. 2002), havent treballs que reporten 206 insercions (Richly and Leister 2004), 211 (Ricchetti et al. 2004), 247 (Mishmar et al. 2004), 296 (Mourier et al. 2001), 452 (Hazkani-Covo and Graur 2007), 585 (Simone et al. 2011), 612 (Woischnik and Moraes 2002) i 871 (Hazkani-Covo et al. 2010). D'acord amb Lascaro et al. (2008), les principals causes per explicar aquest desacord són el criteri propi de cada investigador en la utilització dels mètodes bioinformàtics i l'aplicació a genomes nuclears que encara no estan acoblats correctament. D'altra banda, hi ha molts estudis que no tenen en compte els possibles efectes dels processos post-insercionals de les duplicacions en el genoma nuclear (Hazkani-Covo et al. 2003). A més, les recopilacions i bases de dades existents no estan actualitzades, ja que s’han dut a terme amb una versió antiga del genoma humà enlloc de l’actual versió (GRCh37).

"It is quite likely that all of us harbour more than one mtDNA type among the trillions and trillions of mtDNA genomes in our bodies"

Brigitte Pakendorf & Mark Stoneking.

1.3 HETEROPLÀSMIA MITOCONDRIAL

Les múltiples còpies del genoma mitocondrial dins d'un individu no han de ser necessàriament idèntiques; l'existència de diferents tipus de DNAmt dins d'una mateixa mitocòndria, cèl·lula o individu es coneix com a heteroplàsmia (Figura 3).

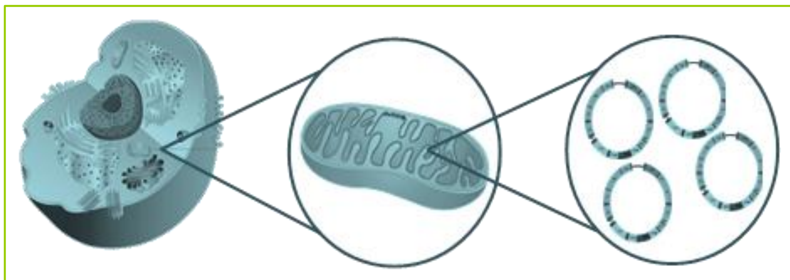


Figura 3. Representació esquemàtica dels diferents nivells on pot observar-se l'heteroplàsmia mitocondrial

Existeixen dos tipus d'heteroplàsmia mitocondrial:

- Heteroplàsmia puntual: s'estableix quan unes còpies de DNAmt pateixen una substitució (Figura 4a).

- Heteroplàsmia de longitud, s'estableix quan unes còpies de DNAmt pateixen una inserció o deleció (Figura 4b). Aquesta acostuma a produir-se en zones repetitives com les regions de poliC del D-loop: la regió compresa entre les posicions 16184-16193 de la HVRI i 303-315 de la HVRII.

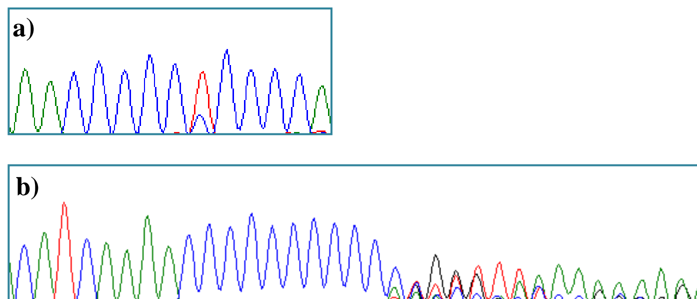


Figura 4. Electroferograma corresponent a la seqüència d'una regió del DNAmt. **a)** Heteroplàsmia puntual donada per una substitució C/T. **b)** Heteroplàsmia de longitud en una poliC de la HVRII situada entre les posicions 303-309.

L'estudi de l'heteroplàsmia mitocondrial humana es remunta als anys 80 i va ser identificada per primera vegada en estudis genètics sobre malalties mitocondrials, on es van observar nivells variables de delecions (Holt et al. 1988) o mutacions (Wallace et al.

1988) en els pacients afectats. No obstant això, l'heteroplàsmia no és necessàriament un estat relacionat amb patologies. Bendall et al. (1996) va suggerir que la freqüència de l'heteroplàsmia mitocondrial és superior a la detectada fins el moment i que no s'havien pogut detectar abans degut, probablement, a dificultats metodològiques. Per tant, la freqüència observada dependrà de com es mesura.

Les estimacions inicials sobre la freqüència d'heteroplàsmia puntual detectaven nivells mínims o nuls d'heteroplàsmia, degut a que els mètodes emprats eren poc sensibles. Una de les estimacions més actuals considera que aproximadament un 24% de la població té més d'un tipus de DNAmt present, amb una freqüència d'almenys un 0.08% (Li et al. 2010), i de fet és molt probable que tothom albergui més d'un tipus de DNAmt entre els trilions i trilions de genomes mitocondrials que tenim al nostre cos (Pakendorf and Stoneking 2005). No obstant això, l'homogeneïtat total del DNAmt entre individus indica que una restricció en el nombre de molècules de DNAmt, o "coll d'ampolla", està darrere d'aquest procés: un nombre molt petit de molècules de DNAmt, o fins i tot una sola molècula, podria servir de model en la replicació per a formar part dels oòcits dels qual es transmet el DNAmt a la descendència (Clayton 1996; Chinnery 2002; Poulton 1995). Diferents autors estan d'acord en que, ja sigui per deriva genètica o per un complex procés de tria, és possible explicar fàcilment la ràpida segregació dels genomes mitocondrials; el veritable enigma resideix en l'estabilitat que tenen certes heteroplàsmies observades en alguns arbres genealògics (Lightowlers et al. 1997; Santos et al. 2005b). Una dona amb una fracció relativament petita de DNAmt mutat pot tenir descendència amb una proporció significativament més alta d'aquest DNAmt

mutat, si aquest és aleatòriament amplificat. No obstant això, l'heteroplàsmia hauria de ser reduïda en la majoria dels seus descendents.

Actualment però, no està del tot clar en quin moment del desenvolupament es dóna aquesta reducció dràstica de material genètic mitocondrial. Estudis recents han postulat que el coll d'ampolla es produeix durant el desenvolupament embrionari com a conseqüència d'una dràstica reducció del contingut de mtDNAs a les cèl·lules germinals (Jansen and de Boer 1998; Krakauer and Mira 1999), mentre que altres estudis proposen que el coll d'ampolla és el resultat de la replicació d'una subpoblació de mtDNAs durant la maduració de l'oòcit a nivell postnatal i que, per tant, el coll d'ampolla es dóna sense una reducció del contingut de molècules de mtDNA a les cèl·lules germinals (Cao et al. 2010; Wai et al. 2008). S'han dut a terme estimacions de la mida del coll d'ampolla (Bendall et al. 1996; Howell et al. 1992; Marchington et al. 1998) utilitzant diferents models de selecció. Aplicant el model de selecció única descrit per Bendall et al. (1996), en parelles de bessons, han estimat una mida del coll d'ampolla de 3 a 30 unitats de segregació. Marchington et al. (1998), utilitzant oòcits han obtingut una estimació similar de 1 a 31 unitats de segregació. D'altra banda, estudis de segregació de l'heteroplàsmia mitocondrial quantifiquen, a través de les generacions, de l'ordre de només 10 unitats de segregació per individu (Bendall et al. 1996; Brown et al. 2001; Poulton and Marchington 2002).

Actualment existeixen un nombre important de treballs centrats en l'observació de les variacions heteroplàsmiques en

poblacions sense malaltia mitocondrial evident (Calloway et al. 2000; He et al. 2010; Irwin et al. 2009; Kirches et al. 2001; Santos et al. 2008c). En general, aquests estudis s'han centrat en l'anàlisi de la regió de control, zona que només cobreix el 7% del DNAmt total. Aquests treballs detecten elevades freqüències pel que fa a heteroplàsmia de longitud (aproximadament un 70%) i al voltant d'un 4% en quant a la freqüència d'heteroplàsmia puntual. Els escassos estudis en els que s'han analitzat regions de la porció codificant del DNAmt en individus sans mostra que l'heteroplàsmia mitocondrial és més freqüent a la regió de control (Jazin et al. 1996; Li et al. 2010; Santos et al. 2008b).

S'ha observat que els nivells d'heteroplàsmia varien entre els teixits (Bendall et al. 1997; Calloway et al. 2000; Goto et al. 2011; Grzybowski et al. 2003; He et al. 2010). En aquests treballs s'ha posat de manifest que, teixits com el muscular o l'epiteli bucal presenten freqüències més elevades d'heteroplàsmia respecte altres teixits com el sanguini. També s'ha detectat que la freqüència observada a les arrels dels cabells és molt variable, podent presentar tant els nivells mínims com els més elevats d'heteroplàsmia que pot tenir un mateix individu. D'altra banda, la freqüència d'heteroplàsmia també pot variar entre poblacions (Irwin et al. 2009). Al treball de Irwin et al. (2009) es destaquen, d'entre les poblacions estudiades, les poblacions de Jordània (9.5% en sang) i Vietnam (15.5% en epiteli bucal) com les que presenten freqüències més elevades d'heteroplàsmia, mentre que les de Nairobi (1% en sang) i Kazakhstan (4.3%, en epiteli bucal) són les que tenen freqüències més baixes.

Si enfoquem l'heteroplàsmia mitocondrial com una etapa intermèdia entre la generació de mutacions i la fixació d'aquestes en un individu o una cèl·lula, aquesta representa un pas obligatori en l'evolució del DNAmt (Rand 2001). Per tant, l'heteroplàsmia mitocondrial és una eina potencial per a l'estudi del patró de mutació, el possible paper de la selecció i, fins i tot, de l'existència de la recombinació del mtDNA (Zsurka et al. 2005).

"There's about 1,000 times more mitochondrial DNA than nuclear DNA in our cells, so it's much easier to pick up"
Chris Stringer

1.3.1 Metodologies aplicades a l'estudi de l'heteroplàsmia mitocondrial

La detecció de l'heteroplàsmia mitocondrial ha evolucionat significativament des de la seva primera troballa, cap als volts dels anys 80. De fet, és la millora en les tècniques el que ha permès aprofundir més i millor en seu estudi. A banda d'intentar millorar el nivell de detecció, les tècniques per observar-la són molt variades i sovint depenen del tipus d'estudi que es pretengui realitzar.

S'ha dut a terme un recull de les publicacions relacionades amb la detecció de l'heteroplàsmia puntual des de l'any 2004 fins a l'actualitat. Aquesta revisió inclou 24 estudis poblacionals (Afonso Costa et al. 2010; Bai and Wong 2004; Ballana et al. 2008; Biggin et al. 2005; Cassandrini et al. 2006; Chong et al. 2005; De Camargo et al. 2011; Irwin et al. 2009; Jacobs et al. 2007; Lo et al. 2005; Lutz-Bonengel et al. 2008; Meierhofer et al. 2005; Nakahara et al. 2008;

Paneto et al. 2010; Picornell et al. 2006; Poe et al. 2007; Rose et al. 2010; Santos et al. 2008b; Singh et al. 2006; Sondheimer et al. 2011; Theves et al. 2006; Wang and Boles 2006; White et al. 2005; Zsurka et al. 2005) i 43 estudis realitzats amb mostres patològiques (Abnet et al. 2004; Alston et al. 2010; Ambrosini-Spaltro et al. 2010; Betts et al. 2006; Bi et al. 2010; Bidooki et al. 2004; Brinckmann et al. 2007; Canter et al. 2005; Cardaioli et al. 2008; Coskun et al. 2004; D'Aurelio et al. 2010; Debray et al. 2007; Dobrowolski et al. 2009; Fendt et al. 2011; Frederiksen et al. 2009; Gambello et al. 2006; Genasetti et al. 2007; Gigarel et al. 2005; He et al. 2010; Jacobs et al. 2005; Jakupciak et al. 2008; Kaare et al. 2009; Kaplanova et al. 2004; Karppa et al. 2004; Kloss-Brandstatter et al. 2010; Laloi-Michelin et al. 2009; Leshinsky-Silver et al. 2005; Li et al. 2005; Lindberg et al. 2008; Lu et al. 2009; Marchington et al. 2011; Meierhofer et al. 2006; Monnot et al. 2011; Phasukkijwatana et al. 2006; Sarzi et al. 2007; Suen et al. 2010; Tan et al. 2008; Uusimaa et al. 2007; Whittaker et al. 2007; Wong et al. 2004; Zaragoza et al. 2010; Zhadanov et al. 2007; Zsurka et al. 2011).

Aquest recull ha evidenciat que la tècnica més habitual per determinar els nivells d'heteroplàsmia és la seqüenciació automàtica basada en el mètode Sanger. D'aquesta manera, un 41,18% dels estudis poblacionals s'han realitzat mitjançant la seqüenciació automàtica (Afonso Costa et al. 2010; Biggin et al. 2005; Cassandrini et al. 2006; Chong et al. 2005; De Camargo et al. 2011; Irwin et al. 2009; Jacobs et al. 2007; Lo et al. 2005; Nakahara et al. 2008; Paneto et al. 2010; Picornell et al. 2006; Santos et al. 2008b; Theves et al. 2006; Wang and Boles 2006), havent-se observat una

freqüència similar (40,32%) en els estudis centrats en malalties (Abnet et al. 2004; Ambrosini-Spaltro et al. 2010; Bi et al. 2010; Bidooki et al. 2004; Brinckmann et al. 2007; Cardaioli et al. 2008; D'Aurelio et al. 2010; Fendt et al. 2011; Gigarel et al. 2005; He et al. 2010; Jacobs et al. 2005; Jakupciak et al. 2008; Kloss-Brandstatter et al. 2010; Laloi-Michelin et al. 2009; Leshinsky-Silver et al. 2005; Li et al. 2005; Lu et al. 2009; Marchington et al. 2011; Meierhofer et al. 2006; Monnot et al. 2011; Sarzi et al. 2007; Tan et al. 2008; Zaragoza et al. 2010; Zhadanov et al. 2007; Zsurka et al. 2011).

La segona tècnica més emprada en els estudis poblacionals (17,65%), és la cromatografia líquida d'alta resolució (DHPLC) (Ballana et al. 2008; Biggin et al. 2005; Jacobs et al. 2007; Lo et al. 2005; Meierhofer et al. 2005; Rose et al. 2010). Per contra, en els estudis de malalties (Brandstatter and Parson 2003; Kaare et al. 2009; Meierhofer et al. 2006), aquesta tècnica es fa servir només en un 3,23% dels casos. Alternativament, el segon mètode més utilitzat en els estudis de malalties és el de PCR-RFLP (Alston et al. 2010; Betts et al. 2006; Brinckmann et al. 2007; Canter et al. 2005; Cardaioli et al. 2008; Coskun et al. 2004; D'Aurelio et al. 2010; Debray et al. 2007; Genasetti et al. 2007; Kaplanova et al. 2004; Karppa et al. 2004; Lindberg et al. 2008; Phasukkijwatana et al. 2006; Sarzi et al. 2007; Suen et al. 2010; Uusimaa et al. 2007; Whittaker et al. 2007; Wong et al. 2004; Zhadanov et al. 2007; Zsurka et al. 2011), amb una freqüència d'un 32,26%, aplicant-se tan sols en un 14,71% dels estudis poblacionals (Bai and Wong 2004; Meierhofer et al. 2005; Singh et al. 2006; White et al. 2005; Zsurka et al. 2005).

Les tècniques de seqüenciació de nova generació, com les plataformes 454 (Roche), Illumina (Solexa) i Life (APG), s'utilitzen tant en estudis poblacionals (14,71%) (Ballana et al. 2008; Biggin et al. 2005; Lutz-Bonengel et al. 2008; Sondheimer et al. 2011; White et al. 2005) com en estudis de malalties (6,45%) (Alston et al. 2010; Brinckmann et al. 2007; He et al. 2010; Zaragoza et al. 2010).

Finalment, un 11,76% i un 17,74% respecte als estudis poblacionals (Bai and Wong 2004; Cassandrini et al. 2006; Poe et al. 2007; Theves et al. 2006) i de malalties respectivament (Bi et al. 2010; Dobrowolski et al. 2009; Frederiksen et al. 2009; Gambello et al. 2006; Genasetti et al. 2007; Jacobs et al. 2005; Jakupciak et al. 2008; Laloi-Michelin et al. 2009; Li et al. 2005; Tan et al. 2008; Wong et al. 2004), apliquen tecnologies diferents i més específiques. L'electroforesi temporal per gradient de temperatura (TTGE), el *melting* d'alta resolució (HRM) i el sistema d'amplificació refractari de mutació (ARMS), són alguns dels mètodes utilitzats als estudis de malalties. Pel que fa als estudis poblacionals, s'usen altres tècniques com la detecció a temps real de la fluorescència, la hibridació per Southern blot i LATE-PCR de molècula única.

Al voltant dels anys 80, alguns autors van concloure que la seqüenciació automàtica (Maxam and Gilbert 1977; Sanger and Coulson 1975; Smith et al. 1985; Smith et al. 1986) era capaç de detectar clarament el doble pic d'una heteroplàsmia puntual on la variant minoritària fos ≥20%, però donava senyals ambigües i/o indetectables a nivells inferiors al 5-10% (Bai and Wong 2004; Irwin et al. 2009; Tang and Huang 2010). Malgrat això, s'ha demostrat clarament que una heteroplàsmia on la variant minoritària sigui del

5% pot ser observada per seqüenciació automàtica si es segueixen estrictes protocols de validació i autentificació d'aquestes (Santos et al. 2008b). Aquests protocols estan bastats en un criteri molt exigent d'acceptació de seqüències vàlides (amb un nul soroll de fons) i d'un seriat de repeticions independents, utilitzant *primers* de seqüenciació en ambdós sentits (directe i revers), diferents productes de PCR i fins i tot diferents extractes de DNA.

Segons Grzybowski et al. (2000), l'ús de diferents tècniques d'amplificació pot donar diferències en la quantificació dels nivells d'heteroplàsmia. No obstant això, aquests resultats no van poden ser reproduïts en un altre estudi (Brandstatter and Parson 2003). De la mateixa manera, en un estudi recent, l'ús de la seqüenciació de la cadena lleugera (L) dóna millors resultats en quant a qualitat de la seqüència que no pas si es seqüencia la cadena pesada (H) (Brandstatter and Parson 2003).

D'altra banda, la presencia de NUMTs porta a replantejar els estudis de DNAmt. Degut a l'elevada similitud amb el DNAmt, els NUMTs són una font de contaminació en els treballs de DNAmt basats en la seva amplificació per PCR (Parr et al. 2006b; Yao et al. 2008b). Tot i que aquest problema ha estat silenciat amb l'argument de l'elevat nombre de còpies del DNAmt, respecte el locus corresponent al DNAn, la precaució ha de ser-hi present (Parfait et al. 1998), ja que la co-amplificació amb un NUMT donarà problemes en la interpretació de les seqüències i, sobretot, en la interpretació de l'heteroplàsmia mitocondrial (Parr et al. 2006a; Parr et al. 2006b). En un estudi de Parr et al. (2006b), on utilitzà cèl·lules ρ0 (absència de mitocòndries), va suggerir que un dels factors més importants que

determina si un NUMT serà co-amplificat amb el DNAmt és la regió de DNAmt que es pretén amplificar així com el número de còpies del NUMT. A més, les mostres de DNA antic o qualsevol teixit amb un baix nombre de còpies de DNAmt, tant en teixit sa com en mostres patològiques (per exemple mostres tumorals) també semblen ser factors importants que determinarien la possibilitat de co-amplificació (Goios et al. 2008; Parr et al. 2006b). Goios et al. (2008) va postular que en seqüenciacions estàndards de mostres utilitzades per a la caracterització poblacional, en la que la detecció d'heteroplàsmia mitocondrial no constituïa una prioritat, la co-amplificació de NUMTs no constitueix un problema real, ja que el contingut de DNAmt en aquestes mostres és molt superior al contingut de DNAn. Però Yao et al. (2008b) i Parr et al. (2006b) van demostrar que la co-amplificació de NUMTs pot donar-se fins i tot quan s'utilitzen mostres amb una qualitat i quantitat de DNA estàndard. A l'estudi de Parr et al. (2006b) es demostra aquesta problemàtica amb la seqüenciació de 46 fragments de DNA paràlegs que representen tot el genoma mitocondrial i Hirano et al. (1997) assenyalaren que una heteroplasmia aparent en un malalt d'Alzheimer era en realitat el resultat d'una co-amplificació amb DNAn.

2. JUSTIFICACIÓ DEL TREBALL I OBJECTIUS

2.1 JUSTIFICACIÓ DEL TREBALL

El DNAmt humà segueix sent un dels DNAs més estudiats. Els atributs distintius, en contrast amb el genoma nuclear, com ara la seva abundància en les cèl·lules humanes, l’herència materna i l’elevada taxa de mutació, fan d'aquesta molècula una atractiva font d'informació per a estudis evolutius i de genètica de poblacions. Una altra característica del genoma mitocondrial és la presència de més d’un tipus de DNAmt en una mateixa mitocòndria, cèl·lula o individu, particularitat anomenada heteroplàsmia mitocondrial. Actualment, hi ha una gran varietat de treballs centrats en l'estudi de l’heteroplàsmia mitocondrial, tant per la seva relació amb malalties com estudis amb individus sans. Aquests últims s'han centrat en l'estudi de la regió no codificant del genoma mitocondrial, que representa només el 7% del seu genoma. Tot i així, pel que fa a l'anàlisi de l’heteroplàsmia en la totalitat del genoma mitocondrial, en el moment d’inici de la present Tesi Doctoral, no hi havia cap treball publicat enfocat en aquest tema on s’utilitzessin seqüències completes de DNAmt. Recentment però, s’ha publicat un treball on s’analitza l’heteroplàsmia mitocondrial mitjançant seqüències completes de DNAmt, tot i que el propòsit se centra en l’aplicació de

la plataforma de seqüenciació massiva Illumina (Solexa) (Li et al. 2010).

Un estudi exhaustiu de les posicions heteroplàsmiques ens ajudarà a entendre millor quina és la dinàmica i quines forces estan intervenint en l'evolució del genoma mitocondrial. La comprensió de les bases de l'heteroplàsmia mitocondrial, aportarà una valuosa informació en una gran diversitat de camps d'investigació com l'evolutiu, el poblacional, el forense i el de la genètica.

Per treballar amb el DNAmt i, més encara, si es pretén estudiar l'heteroplàsmia mitocondrial, cal tenir en compte la presència dels anomenats NUMTs. A causa de l'elevada similitud entre NUMTs i DNAmt, una co-amplificació entre ells podria comprometre la veracitat dels resultats. Actualment, no existeix en la bibliografia cap treball on tots els *primers* utilitzats evitin la co-amplificació amb NUMTs. És per això que, dur a terme una bona estratègia per evitar-ne aquesta co-amplificació, és de vital importància.

Degut a la possibilitat de co-amplificació amb el DNAmt, seria molt útil disposar d'una base de dades de totes les insercions d'origen mitocondrial. Actualment existeixen diverses bases de dades de NUMTs, però entre elles presenten una gran discrepància en el nombre total de NUMTs. A més, totes elles es troben desactualitzades ja que s'han realitzat amb una versió antiga de l'esborrany del genoma humà i no amb la darrera versió (GRCh37). Això comporta la necessitat de generar una nova base de dades basada en el nou esborrany del genoma humà. Aquesta pràctica eina permetrà poder validar mutacions del DNAmt, així com l'elaboració

d'unes estratègies adequades per a la realització d'estudis evolutius, filogeogràfics i, fins i tot, epidemiològics.

2.2 OBJECTIUS

El propòsit general d'aquest estudi és determinar la freqüència i el patró de l'heteroplàsmia en la totalitat del genoma mitocondrial humà. Per fer-ho, es plantegen els següents objectius:

1. **Establir una metodologia adequada per a l'amplificació i seqüenciació del genoma mitocondrial humà que permeti la correcta interpretació de l'heteroplàsmia tant en mostres d'individus sans com amb patologies.**
2. **Realitzar una exhaustiva compilació de les insercions nuclears d'origen mitocondrial basada en la darrera versió de l'esborrany del genoma humà (GRCh37).**
3. **Avaluar l'impacte de la contaminació per NUMTs en estudis de càncer com un exemple d'aplicació de la base de dades descrita.**
4. **Determinar la freqüència i el patró de l'heteroplàsmia en la totalitat del genoma mitocondrial a nivell poblacional.**
5. **Comparar el patró mutacional de l'heteroplàsmia mitocondrial amb l'observat a nivell poblacional.**

3. METODOLOGIA, RESULTATS I DISCUSSIÓ

3.1 Capítol 1:
Primer design and validation for mtDNA amplification and sequencing

The described methodology as well as the results of the primer design are detailed below and in the following publications (Annex II):

Ramos, A; Santos, C; Alvarez, L; Nogués, R; Aluja, MP. ***Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification****. Electrophoresis. 2009 May; 30(9):1587-1593.*

Ramos, A; Santos, C; Barbena, E; Mateiu, L; Alvarez, L; Nogues, R; Aluja MP. ***Validated primer set that prevents nuclear DNA sequences of mitochondrial origin coamplification: A revision based on the New Human Genome Reference (GRCh37).*** *Electrophoresis. 2011 Mar;32(6-7):782-3.*

3.1.1 Materials and methods

Primer design and validation

In order to analyze the whole mitochondrial genome, primers previously published (Torroni et al. 2001) were tested and new ones were designed. The primer design was performed using the Lasergene 7.2 software, from the DNASTAR package, following the general criteria of primers design (Roche 2006) . Particular attention was devoted to adjust, as much as possible, the melting temperatures (Tm) of all the primer pairs.

To discard co-amplifications of nDNA and mtDNA, PCR primers were submitted to the Basic Local Alignment Search Tool (BLAST) available in the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/ BLAST/) (Altschul et al. 1990). BLAST finds regions of local similarity between sequences comparing nucleotide sequences with sequence databases and calculating the statistical significance of matches. In this work, the specific tool Basic BLAST, optimized for highly similar sequences (Megablast), was performed using the Reference genomic sequence

data base for Homo sapiens (refseq_genomic: Genomic sequences from National Center for Biotechnology Information Reference Sequence Project). A primer pair was discarded if both primers showed similarity (expected number of chance matches in a random model – E-value – lower than 1) inside the same chromosome region.

After BLAST validation, primers were further tested in PCR experiments employing total DNA standard samples (see next section) and nDNA obtained from sperm cells. The nDNA isolation was performed by sperm differential lyses using the QIAmp DNA Investigator Kit (Qiagen) according to the manufacturer's specifications. To guarantee the complete mtDNA elimination, the initial washing step mentioned in QIAmp DNA Investigator Kit manual was repeated at least four times. Moreover, to ensure the integrity of nDNA, Y-chromosome STRs were amplified using AmpFlSTR® Y-filer™ PCR kit (Applied Biosystems), under conditions recommended by the manufacturer and all the markers were correctly amplified.

Total DNA extraction, amplification and sequencing of the entire mtDNA

Total DNA from blood samples was extracted using JETQUICK Blood DNA Spin Kit (Genomed) and isoamilic phenol-chloroform (Sambrook 2001).

Optimized primers and PCR conditions for the mtDNA amplification are reported in Table 1. The PCR mix for each sample consisted of 50 pmol of each primer, 200 mM of each dNTP, 2mM of $MgCl_2$, $10NH_4$-based reaction buffer, 1 U of *Taq* DNA polymerase, and 30 ng of DNA in a final volume of 50 mL. The PCRs were performed in a G-Storm GS1 thermocycler and the amplification program consisted of an initial denaturation step of 5 min, followed by 35 cycles of PCR (1 min at 94ºC, 40 s at annealing temperature (Ta) and 2.5 min at 72ºC), and a final extension step of 5 min at 72ºC. The PCR amplification results were visualized by electrophoresis in agarose gels (2%).

PCR products were purified using the JETQUICK PCR Purification Spin Kit (Genomed) and the mtDNA was fully sequenced in both strands using 62 primers (Table 2). Sequence reactions were carried out using the sequencing kit BigDye Terminator v.3 (Applied Biosystems) according to the manufacturer's specifications and were run in an ABI 3130XL sequencer (Servei de Genòmica, Universitat Autònoma de Barcelona).

Cloning

PCR products that, in the optimization process, show consistently more than a one size product were cloned into the pCR®4-TOPO® vector, using the TOPO TA Cloning® Kit for Sequencing (Invitrogen). Ten clones were sequenced using the previously explained methodology.

3.1.2 Results

Primer design and validation

From the primers reported in the literature for the complete mtDNA amplification, we evaluated the 11 primer pairs published by Torroni et al. (2001) since, compared with other works, the authors define a small number of fragments [see Table 2 in Torroni et al. (2001)]. These primers were tested using BLAST. After this analysis, four pairs of primers that show high similarity with nuclear regions were rejected.

The BLAST analysis allows detecting two major problematic regions for primer design. The first one, located between 3914 and 9074 mtDNA positions, presents a 98% similarity with a region within chromosome 1 (see BLAST results on Table I.1.1(1) of Annex I); the second one encompasses the region located between 9582 and 14479 mtDNA positions and it shows a 88% similarity with a region within chromosome 5 (see BLAST results on Table I.1.1 (2) of Annex I). For these regions new primers were designed and verified until the selective amplification of mtDNA was obtained. The 3914–9074 region (Figure 5) was split into two fragments (fragments 4 and 5 in Figure 5); For each primer pair, one

of the primers was designed outside the 3914–9074 region, preventing the amplification of the similar region of chromosome 1. For the 9582–14479 region (Figure 5), minimum of four fragments with 2000 bp were necessary to completely amplify it. Thus, the previous strategy could only be applied for flanking fragments (fragments 6 and 9 in Figure 5), and the primers used to amplify the two fragments that were placed completely within the region (fragments 7 and 8 in Figure 5) were designed in low-sequence similarity regions.

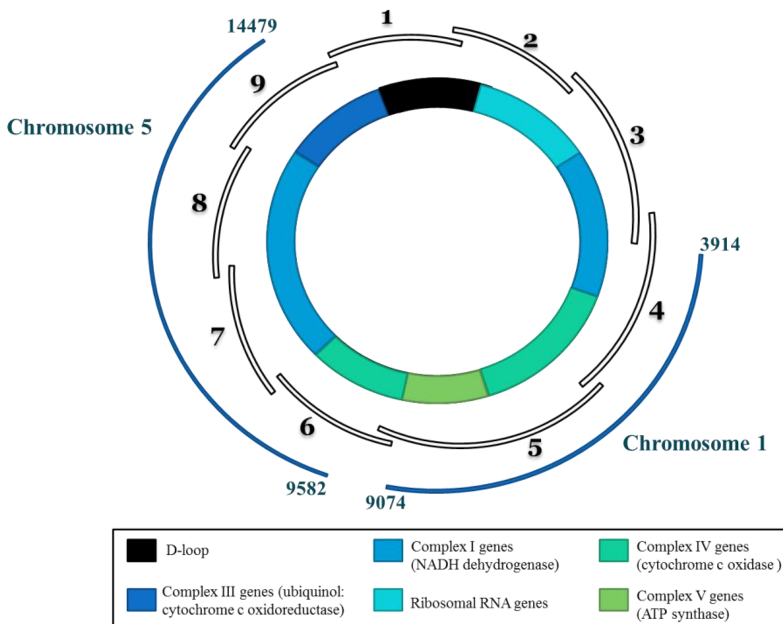


Figure 5. - Schematic representation of mtDNA. The nine overlapping fragments defined to PCR amplify the complete mtDNA genome are represented as well as the two nuclear regions with high similarity with mtDNA

After BLAST validation, PCR optimization was performed using three samples of total DNA extracted using JETQUICK Blood DNA Spin Kit (Genomed). Surprisingly, the outcome of the amplifications of fragments between positions 12012–13828 and 3798–6739 resulted in more than one PCR product. However, a new amplification with a different sample isolated with isoamilic phenol-chloroform (Sambrook 2001) was carried out and only one PCR fragment (with the expected size) was obtained (Figure 6). These results allow to hypothesize that: (i) the observed additional bands correspond to deletions in mtDNA; (ii) the amplification of an NUMT not detected by BLAST analysis; (iii) the extraction method influences the DNA quality and by consequence the PCR performance and specificity, resulting in unspecific amplification of nuclear or mtDNA. To test these hypotheses the fragment 12012–13828 was cloned into the pCR®4- TOPO® vector, using the TOPO TA Cloning® Kit for Sequencing (Invitrogen) and ten clones were sequenced. The resulting sequences were tested with BLAST showing a complete similarity with mtDNA. A careful analysis reveals that the PCR forward primer hybridized with two close regions of mtDNA: the target region and another one that present several differences with the primer. Therefore, the amplification result is an artifact that can be attributed to the poor quality of DNA. For this reason, the forward primer was redesigned and tested again as previously described.

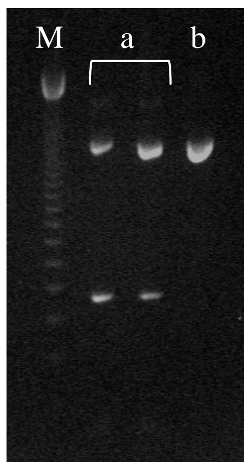


Figure 6- Gel electrophoresis showing amplification results between positions 12012-13828 using blood samples extracted with different methods. a) JETQUICK Blood DNA Spin Kit (Genomed); b) isoamilic phenol-chloroform. M: Size marker.

Finally, to discard co-amplifications of nDNA and mtDNA, which could not be deduced by BLAST analysis, PCR amplifications were performed using isolated nDNA obtained from sperm cells. The PCR primers were tested and no positive amplifications were observed.

PCR and sequencing primers to mtDNA analysis

For these primers, validated as previously explained, melting temperatures were adjusted as much as possible and the annealing temperatures were obtained by a PCR gradient being selected the highest temperature that shows positive amplification. In Table 1

PCR primers and annealing and melting temperatures to amplify the nine overlapping fragments that cover the entire mtDNA (Figure 5) are reported. Some of the primers were previously published by Torroni et al. (2001) and the remaining ones were designed for this study.

Since the size of fragments is of ~2000 bp, internal primers were designed to fully sequence both strands of each fragment. In Table 2 the sequence of the primers to sequence each fragment in both stands are shown. All the primers were tested to sequence three samples and a good efficiency of sequencing was obtained.

Table 1- Validated primers to amplify the complete mtDNA in 9 overlaping fragments. Melting temperatures (Tm) and annealing temperature (Ta) for each pair of primers are also presented.

Fragment	Fragment length	Name	Sequence (5'→ 3')	Primer pair length (pb)	Ta	Tm
1	1822	14898for 151rev	tagccatgcactactcaccaga ggatgaggcaggaatcaaagac	22	60	60.3
2	1758	16488for[a] 1676rev[a]	ctgtatccgacatctggttcct gtttagctcagagcggtcaagt	22	60	60.3
3	2543	1404for[a] 3947rev	acttaagggtcgaaggtggatt tcgatgttgaagcctgagacta	22	57	58.4
4	3005	3734for 6739rev	aagtcaccctagccatcattcta gatatcatagctcagaccatacc	23	61	58.9
5	2709	6511for 9220rev	ctgctggcatcactatactacta gattggtgggtcattatgtgttg	23	58	58.9
6	1738	8910for[a] 10648rev	cttaccacaaggcacacctaca ggcacaatattggctaagaggg	22	61	60.3
7	1866	10360for 12226rev	gtctggcctatgagtgactaca cagttcttgtgagctttctcgg	22	61	60.3
8	1853	11977for 13830rev	ctccctctacatatttaccacaac aagtcctaggaaagtgacagcga	24 23	63	59.3 60.6
9	1872	13477for[a] 15349rev[a]	gcaggaatacctttcctcacag gtgcaagaataggaggtggagt	22	63	60.3

[a] Oligonucleotides from Torroni et al. (2001)

Table 2- Primers designed to sequencing the whole mtDNA after PCR amplification in nine overlapping fragments.

Fragment	Name / Sequence (5'→3')				
1	14898for tagccatgcactactcaccaga	15416for tacacaatcaaagacgccctc	15966for agtctttaactccaccattag		
	151rev ggatgaggcaggaatcaaagac	16281rev gttggtatcctagtgggtgag	15825rev gtgaagtatagtacggatgct		
2	16488for ctgtatccgacatctggttcct	411for cggtatgcacttttaacagtc	909for gattaacccaagtcaatagaa		
	1677rev gtttagctcagagcggtcaagt	1159rev taagctgtggctcgtagtgt	638rev ggtgatgtgagcccgtctaaa		
3	1404for acttaagggtcgaaggtggatt	2028for gatagaatcttagttcaactt	2646for ggttcagctgtctcttacttt	3239for gcagagcccggtaatcgcata	
	3947rev tcgatgttgaagcctgagacta	3382rev ttcgttcggtaagcattagga	2801rev taatgcaggtttggtagttta	2263rev gggtgtgaggagttcagttat	
4	3734for aagtcaccctagccatcattcta	4346for gaacccatccctgagaatcca	4896for taccaaatctctccctcacta	5468for cacgctactcctacctatctc	5995for cagctctaagcctccttattc
	6739rev gatatcatagctcagaccatacc	6154rev ggaactagtcagttgccaaag	5571rev aagtattgcaacttactgagg	5017rev gaggagtatgctaagattttg	4421rev cttatttagctgaccttactt
5	6511for ctgctggcatcactatactacta	7111for acaccctagaccaaacctacg	7713for tcctaacactcacaacaaaac	8911for tagcattaaccttttaagtta	
	9220rev gattggtgggtcattatgtgttg	8600rev agaatgatcagtactgcggcg	8000rev caacgtcaaggagtcgcaggt	7359rev ctactattaggacttttcgct	
6	8910for cttaccacaaggcacacctaca	9393for cgagaaagcacataccaaggc	9874for taatatttcactttacatcca		
	10648rev ggcacaatattggctaagaggg	10154rev ttctatgtagccgttgagttg	9647rev agctcaggtgattgatactcc		
7	10360for gtctggcctatgagtgactaca	10892for atcaacaacaacctatttagc	11461for actcttaaaactaggcggcta		
	12226rev cagttcttgtgagctttctcgg	11673rev gtttggatgagaatggctgtt	11163rev cgggtgatgatagccaaggtg		
8	11977for ctccctctacatatttaccacaac	12500for tgtgcctagaccaagaagtta	12988for ctagcagcagcaggcaaatca		
	13830rev aagtcctaggaaagtgacagcga	13297rev ggttgatgccgattgtaacta	12763rev cgatgaacagttggaataggt		
9	13477for gcaggaatacctttcctcacag	13950for ctatctaggccttcttacgag	14440for atactcctcaatagccatcgc		
	15349rev gtgcaagaataggaggtggagt	14838rev catcatgcggagatgttggat	14325rev aactttaatagtgtaggaagc		

3.1.3 Discussion

3.1.3.1 Human mtDNA complete amplification and sequencing: a new validated primer set that prevents NUMTs co-amplification

The PCR amplification and sequencing of mtDNA is routinely used in many research fields; however, the interpretation of results, particularly of mtDNA heteroplasmy, could be complicated by the co-amplification of NUMTs (Parr et al. 2006b; Yao et al. 2008a). Therefore, an effort to design PCR primers that selectively amplify mtDNA must be performed. In this work we report a set of primers to amplify the whole mtDNA that were validated using the BLAST analysis and by the performance of PCR amplifications using nDNA isolated from sperm cells.

The BLAST search allowed the identification of two problematic regions in primer design (region 3914–9074 of the mtDNA present in chromosome 1 and region 9582–14479 present in chromosome 5). Region 3914–9074 was previously reported (Lascaro et al. 2008) as presenting a high sequence similarity with a region within chromosome 1, a result that is in accordance with our

results. However, the presence of an NUMT in chromosome 5, which shows a high identity with the 9582–14479 region of the mtDNA, was not reported in previous studies; accordingly, this region could represent a non-reported NUMT and this deserves further investigation.

According to Goios et al. (2008) the amplification of NUMTs can only be obtained when mtDNA is almost completely removed from the samples, such as those resulting from preferential semen lyses. Thus, to test the proposed set of primers for the selective amplification of mtDNA, we used a sperm sample that was submitted to DNA extraction using four steps of preferential lyses and for which Y-STRs were successfully amplified and no signs of amplification with the designed primers were detected. This result indicates that the proposed primers specifically amplify mtDNA and we can ensure that no NUMTs will be amplified if this set of nine pair primers is used.

The primers proposed amplification regions of about 2000 bp of mtDNA. As a consequence, in some samples (such as ancient and forensic ones) it would be almost impossible to obtain PCR amplifications with this set of primers. To overcome this problem, we suggest that additional primers within each fragment could be designed using a similar strategy to that applied for fragment 9582–14479, that is, locating primers in regions that show low similarity with nDNA.

3.1.3.2 Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: a revision based on the new human genome reference sequence (GRCh37)

The *in silico* primer validation presented in the present chapter and in the publication by Ramos et al. (2009) was performed using the human reference sequence hg18. However, a new human reference sequence -GRCh37- was recently (February 2009) generated by the Genome Reference Consortium. In order to guarantee the selective mtDNA amplification of presented primers, a new *in silico* validation using the new human reference sequence GRCh37 was performed.

Accordingly to the latest human reference sequence update, one of the most problematic NUMTs for the primer design appears to be larger than earlier reported. This nuclear sequence is located in chromosome 1 and presents a 98% of similarity with one third of the mtDNA. This redefined NUMT (detailed and described on table I.1.2 of Annex I) lays between positions 564461 and 570304 (in relation to the *Homo sapiens* chromosome 1, GRCh37 primary reference assembly - NC_000001), and presents a 98% similarity with the region 3911-9755 of human mtDNA (Andrews et al. 1999).

The new *in silico* analysis revealed that one of the nine previously described primer pairs, specifically the one designed to amplify the 2709 bp region of the mtDNA, located between positions 6511 and 9220 [fragment 5 in table 1 and reported by Ramos et al. (2009)], lays on the redefined similar region. Thus, an accidental co-amplification of mtDNA and nDNA could occur. The described NUMT appears to be Human specific since it is present in *Homo sapiens* but it is not present in other primates (see alignment results on table I.1.3 of Annex I). To date, its variability (presence/absence) in human population is unknown, and we cannot discard the possibility of nDNA co-amplification at least in some individuals. For this reason, a new design, description and *in silico* validation [according to the methods previously presented and reported by Ramos et al. (2009)] of the fragment 5 primer pair is presented.

The size of the redefined chromosome 1 NUMT does not allow locating a primer out of this region following a normal PCR amplification strategy. Consequently, a new approach to primer design was undertaken: (i) primers were relocated to areas with low degree of similarity with chromosome 1 and (ii) the 3' end of each primer were located in mismatch positions relatively to chromosome 1 sequence. Therefore, the possibility of co-amplification of nDNA and mtDNA is reduced. Optimized primer pair and PCR conditions for this concrete region of mtDNA are reported in Table 3. Internal sequence primers previously reported [table 2 and presented by Ramos et al. (2009)] are not affected and can still be used to sequence this region.

Table 3. New validated primer pair to amplify the mtDNA region affected by the new updated NUMT is reported. Melting temperatures (Tm) are also presented. In bold show the primer base located in position 3' that differs from the sequence of chromosome 1.

Fragment	Fragment length	Name	Sequence (5'-3')	Primers pair length (bp)	Tm
5	2664	6520for 9184rev	tcactatactactaacagaccg**c** gtagaggcttactagaagtgt**g**	23 22	58.9 58.4

In conclusion, we report a set of primers that guarantee the selective amplification and sequencing of the whole mtDNA in nine overlapping fragments. These primers could be a useful tool in future projects that deal with mtDNA complete sequencing since they represent the first published set of primers tested for the non-amplification of nDNA and therefore suitable for the screening of mtDNA heteroplasmy in both standard and pathological samples.

3.2 Capítol 2:

Updating of NUMTs database and usefulness in cancer studies

The described methodology as well as the reported database and their application in cancer studies are detailed below and in the following publication (Annex II):

Ramos, A; Barbena, E; Matieu, L; González, MM; Mairal, Q; Lima, M; Montiel, R; Aluja, MP; Santos, C. ***Nuclear insertions of mitochondrial origin: database updating and usefullness in cancer studies***. *Mitochondrion. 2011 Nov;11(6):946-53.*

3.2.1 Materials and methods

Updating of insertions of mitochondrial origin database

NUMTs 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 20bp 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.

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. (2008a) were used to evaluate the impact of NUMT contamination in cancer studies. First, mutations compiled by Santos et al. (2008a) 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 similarity (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.

3.2.2 Results

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.2.1 (Annex I).

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 (Figure 7).

```
Score =  149 bits (164),  Expect = 6e-33
 Identities = 223/311 (72%), Gaps = 17/311 (5%)
 Strand=Plus/Minus

Query  12032     AACATAAAACCCTCATTCACACGAGAAAACACCCTCATGTTCATACACCTATCCCCCATT  12091
                 ||||| ||||| |||||||||||||||||       || |||  | || |||      ||
Sbjct  94387832  AACATTAAACCATCATTCACACGAGAAAA-------ATATTCTAATACTTATATAT--TT  94387782

Query  12092     CTCCTCCTATCCCTCAACCCCGACATCATTACCGGGTTTTCCTCT--TGTAAATATAGTT  12149
                 || |||||||| |  |||||| | |  |||  |  ||   | | |  | ||||||| |||
Sbjct  94387781  CTTCTCCTATCAC--AACCCCCAAACTATT--CTAGTACCCATATATTATAAATATTGTT  94387726

Query  12150     TAACCAAAACATCAGATTGTGAATCTGACAACAGAGGCTTACGACCCCTTATTTACCGAG  12209
                  ||| ||||||| | ||||||| |   |||||||| || ||  || |||||||||| |||
Sbjct  94387725  AAAC-AAAACATTATATTGTGACTAAAACAACAGAAGCCTAAAACTCCTTATTTACTGAG  94387667

Query  12210     AAAGCTCACAAGAACTGCTAACTCATGCCCCCATGTCTAACAACATGGCTTTCTCAACTT  12269
                 ||||    ||| |||| |||||||||||| ||||||  || || || | ||| |||||||
Sbjct  94387666  AAAGTAAGCAAAAACTACTAACTCATGCCTCCATGTAGAATAATATTG-TTTTTCAACTT  94387608

Query  12270     TTAAAGGATAACAGCTATCCATTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAAT  12329
                 |||||| |||  |   ||||  |||  |||||  | |||||  |  ||| ||||||||||
Sbjct  94387607  TTAAAGCATAGAAATAATCCGCTGGAGTTAGGAGCTAAAAAAATGAGTGTAACTCCAAAT  94387548

Query  12330     AAAAGTAATAA  12340
                 ||||| |||||
Sbjct  94387547  AAAAGCAATAA  94387537
```

Figure 7. Alignment between mtDNA (region 12032-12340) and a NUMT located in Chromosome 1 (region 94387537-94387832). Examples of identical positions (box _ _), non-identical position (box ____) and gap region (box ……) are shown. Length in nDNA: 296bp; Length in mtDNA: 309bp; Total length of match: 311bp.

The 755 insertions are distributed throughout the genome with variable frequencies and sizes for each chromosome (Table 4 and Figure 8). 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.

Table 4. Number and length of insertions and their percentages for each chromosome. Length of each chromosome and their percentage are also reported.

Chromosome	Length chromosome (% in relation to the entire genome)	Number of insertions (% in relation to the total number of insertions)	Length of insertions (% in relation to the total length of insertions)
1	249250621 (8.05%)	51 (6.75%)	44798 (8.17%)
2	243199373 (7.86%)	106 (14.04%)	5268 (0.96%)
3	198022430 (6.40%)	44 (5.83%)	97489 (17.78%)
4	191154276 (6.17%)	41 (5.43%)	35103 (6.40%)
5	180915260 (5.84%)	35 (4.64%)	33053 (6.03%)
6	171115067 (5.53%)	31 (4.11%)	11738 (2.14%)
7	159138663 (5.14%)	43 (5.70%)	36328 (6.63%)
8	146364022 (4.73%)	47 (6.23%)	36917 (6.73%)
9	141213431 (4.56%)	34 (4.50%)	28498 (5.20%)
10	135534747 (4.38%)	35 (4.64%)	20926 (3.82%)
11	135006516 (4.36%)	31 (4.10%)	26640 (4.86%)
12	133851895 (4.32%)	42 (5.56%)	8264 (1.51%)
13	115169878 (3.72%)	22 (2.91%)	10691 (1.95%)
14	107349540 (3.47%)	9 (1.19%)	8899 (1.62%)
15	102531392 (3.31%)	14 (1.85%)	9395 (1.71%)
16	90354753 (2.92%)	15 (1.99%)	15965 (2.91%)
17	81195210 (2.62%)	23 (3.05%)	21990 (4.01%)
18	78077248 (2.52%)	6 (0.79%)	1334 0.24%)
19	59128983 (1.91%)	20 (2.65%)	9318 (1.70%)
20	63025520 (2.04%)	9 (1.19%)	7471 (1.36%)
21	48129895 (1.55%)	18 (2.38%)	6753 (1.23%)
22	51304566 (1.66%)	17 (2.25%)	23544 (4.29%)
X	155270560 (5.02%)	50 (6.62%)	38843 (7.08%)
Y	59373566 (1.92%)	12 (1.59%)	9025 (1.65%)
Total	3095677412 (100%)	755 (100%)	548250 (100%)

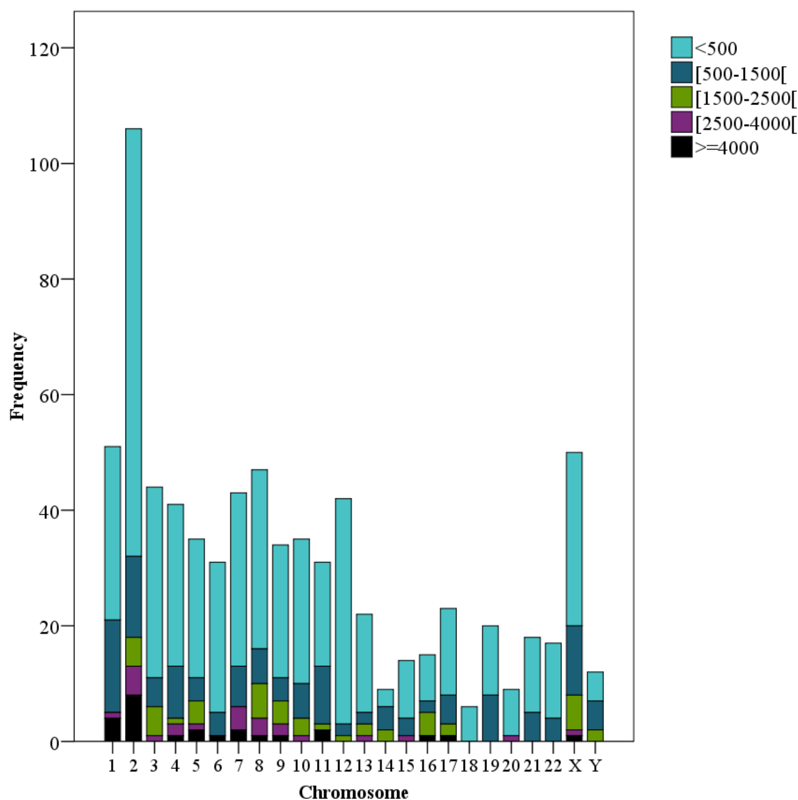
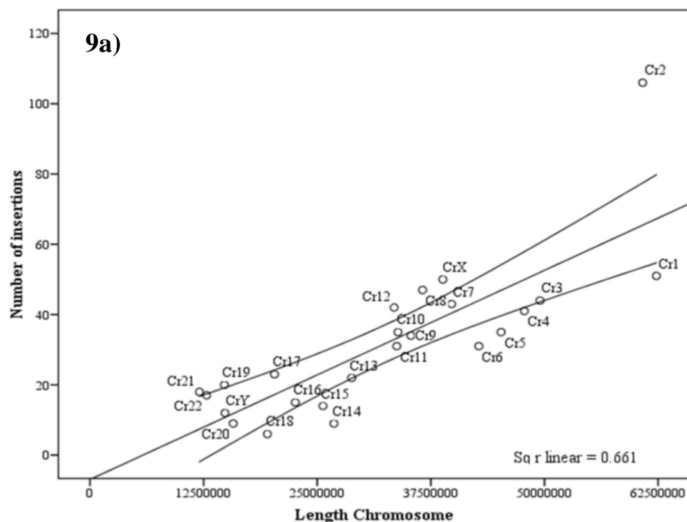


Figure 8. Frequency and size (bp) of insertions for each chromosome.

There is a positive correlation between the number of insertions and chromosome size (Spearman correlation: 0.822 $p<0.001$) (Figure 9a). Chromosome 2, the second largest chromosome in relation to the entire genome (7.86%), appears as an outlier. It encompasses 106 insertions (representing 14.04% of total number of insertions), nearly double the number of insertions that would be expected (Figure 9a).



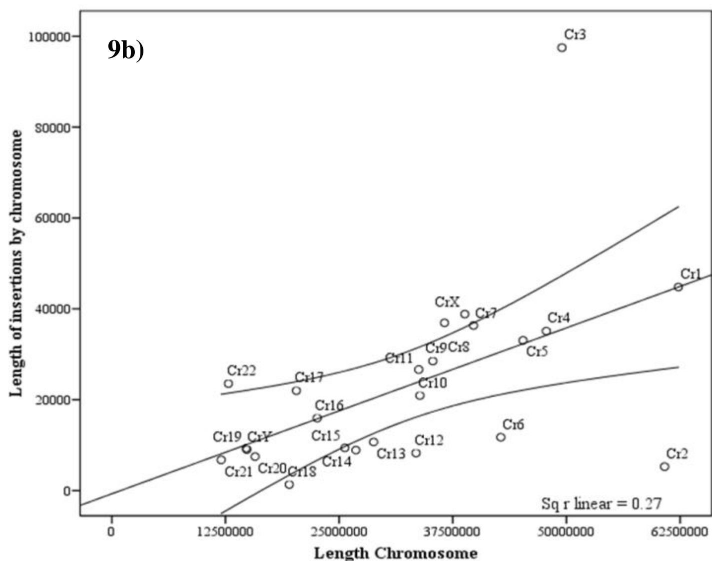


Figure 9. Correlation between chromosome length and number of insertions (figure 9a), and correlation between chromosome and insertion length (figure 9b). Coefficient of determination (Sq r linear) for the two correlations is shown.

The correlation between the size of insertions and their representation on each chromosome is also positive and significant (Spearman correlation: 0.587 p=0.003). However, two outliers, chromosome 2 and 3, were identified (Figure 9b). As before, chromosome 2 is an outlier containing the highest number of insertions. But the length of insertions is quite short, covering only 0.96% of the total length of insertions (Table 4). On the contrary, chromosome 3, which is also one of the largest (6.40% in relation to the entire genome) with a relatively high frequency of insertion (representing 5.83% of total number of insertion), presents the highest total length of insertions (17.78%). In both analyses, chromosome 2 appears as an outlier. This would suggest a differential evolution of this chromosome, a hypothesis that deserves further investigation (Figure 9b).

The mean percentage of identity between nuclear insertions and mtDNA sequence is 79.2%, ranging from 63.5% to 100%. When the analysis of identity is performed for each chromosome (Table 5), the similarity between inserted mitochondrial sequences and the original mitochondrial sequences presents almost the same pattern in all chromosomes, with a mean percentage of identity of ~80% along the chromosomes.

Table 5. Percentage identity for all chromosomes.

Chromosome		1	2	3	4	5	6	7	8	9	10	11	12	13
Mean		78.76	80.20	79.42	81.24	82.02	80.77	77.67	77.88	78.22	78.31	78.27	83.40	80.61
Median		75.73	78.93	79.80	80.43	80.73	81.74	75.83	76.32	77.30	76.98	75.76	88	80.72
Minimum		64.71	65.04	68.14	68.71	66.35	64.69	68.46	67.77	68.02	69.08	66.46	67.79	65.41
Maximum		100	98.48	97.44	96.77	97.67	93.33	97.09	98.81	90.20	92	98.61	98.88	99.22
Percentiles	25	71.43	74.58	73.16	76.19	76.47	77.78	72.22	72.37	73.97	74.14	74.44	77.58	72.41
	50	75.73	78.93	79.80	80.43	80.73	81.74	75.83	76.32	77.30	76.98	75.76	88	80.72
	75	84.03	85.47	83.56	84.86	88.36	85	82.02	82.35	80.71	82.08	81.67	88	88.59

Chromosome		14	15	16	17	18	19	20	21	22	X	Y
Mean		80.40	79.23	77.37	81.69	84.61	72.92	81.43	75.82	75.74	76.05	79.00
Median		79.13	77.65	76.72	80.41	83.96	69.65	77.88	74.05	72.16	73.83	74.57
Minimum		73.55	71.83	68.26	68.54	74.08	63.50	69.49	65.59	65.67	64.82	70.15
Maximum		93.07	95.35	85.45	97.06	95.57	86.59	100	98.75	100	96.30	100
Percentiles	25	74.75	75.15	73.53	77.06	75.89	67.43	72.40	69.09	69.32	69.78	71.38
	50	79.13	77.65	76.72	80.41	83.96	69.65	77.88	74.04	72.16	73.83	74.57
	75	86.02	83.37	81.73	83.42	94.08	79.04	94.10	79.71	80.31	80.95	82.35

The relationship between the size of the insertion and the percentage identity is shown in Table 6. Concerning the size of insertions, it appears that most refer to small insertions (Table 6 and Figure 8) of less than 500 bp. Only 12.85% of the insertions have more than 1500 bp. The smallest insertions found are 39 bp long and are located on chromosomes 1 and 3. On the other hand, the largest insertion (14,836 bp) is located on chromosome 4 (Table I.2.1, Annex I).

Table 6. Frequency and identity of insertions in relation to the size of insertion (bp).

Size (bp)	**Identity**		**Total**	**Percentage**
	≥80%	**<80%**		
<500	276	245	521	69.00
[500-1500[	21	116	137	18.15
[1500-2500[	5	43	48	6.36
[2500-4000[	3	21	24	3.18
≥4000	4	21	25	3.31
Total	309	446	755	100

About half of the small insertions have a percentage identity higher than 80% of the mitochondrial sequence (Table 6). As the size of the insertion increases, however, the proportion of insertions with a percentage identity equal to or higher than 80% decreases. There are 33 paralogous sequences with more than 500 bp, which present a similarity equal to or higher than 80% (Table 7). These sequences would represent the most problematic, in terms of mtDNA/nDNA co-amplification.

Table 7. Paralogous sequences with more than 500 bp and that present a similarity equal or higher than 80%.

Chromosome	% identity	nDNA length	mtDNA length	mtDNA region	nDNA region
1	98.53	5844	5845	3911-9755	564461-570304
1	84.03	517	521	987-1507	142792601-142793117
1	83.87	518	521	987-1507	143243224-143243741
1	83.78	515	519	987-1505	143344679-143345193
2	84.03	516	521	987-1507	95564754-95565269
2	83.95	728	729	6270-6998	50815826-50816553
2	81.78	1350	1349	4854-6202	156119971-156121320
3	95.39	1323	1323	1396-2718	96336032-96337354
3	83.31	623	623	7151-7773	120440870-120441492
3	80.38	1621	1622	1418-3039	40293638-40295258
5	95.05	2114	2121	577-2697	79945841-79947954
5	94.06	5219	5219	10269-15487	134258999-134264217
5	88.36	9108	9067	6117-15183	99381642-99390749
5	87.24	3463	3463	12662-16124	93903161-93906623
5	84.54	962	961	577-1537	123096499-123097460
6	90.70	527	527	2408-2934	62284008-62284534
7	83.66	2514	2517	577-3093	142373012-142375525
7	81.27	562	561	2740-3300	141504769-141505330
9	83.85	2517	2517	577-3093	33656612-33659128
10	84.62	533	533	8281-8813	101817140-101817672
11	94.24	2394	2396	577-2972	10529434-10531827
11	81.68	560	568	596-1163	87524440-87524999
14	93.07	1021	1024	5583-6606	32953304-32954324
17	96.17	653	653	6818-7470	51183094-51183746
17	83.42	10477	10536	577-11112	22021365-22031841
17	81.73	621	619	2656-3274	19504577-19505197
17	80.57	1514	1521	577-2097	19501874-19503387
17	80.41	2206	2205	14365-16569	22018521-22020726
17	80.24	575	582	2084-2665	19503699-19504273
21	83.78	515	519	987-1505	9735524-9736038
X	93.68	554	554	10606-11159	125606714-125607267
X	93.07	749	750	6553-7302	125605687-125606435
Y	83.08	517	521	987-1507	13290151-13290667

Applications to cancer

From the 271 mtDNA mutations previously described in cancer samples, and classified as having zero hits in the mtDNA phylogeny by Santos et al. (2008a), 15 have one or more hits in the current phylogeny by van Oven and Kayser (2009) (mit. Tree build 8). From the 256 positions with zero hits in the current mtDNA phylogeny (Table I.2.2 of Annex I), 220 could represent changes in one or more nuclear insertions of mtDNA (Table I.2.3 of Annex I includes the 220 positions and their matches in non-identical positions in the insertions and the paralogous mtDNA sequences), since they are non-identical between mtDNA and nDNA. Accordingly at that point, only 36 positions can be considered as authentic mtDNA mutations, since they were not found as non-identical in any of the nuclear insertions reported in our study (positions signalled with * in Table I.2.2 of Annex I).

The 220 positions that can represent false mtDNA mutations correspond to a total of 592 mutations described in the cancer studies, revised in Santos et al. (2008a). Of these, 186 mutations were once observed in different types of cancer and/or studies (Table I.2.2 of Annex I and references therein). The mutations that occur

more than once are mostly reported in the study of ovarian cancer by Aikhionbare et al. (2007). In this work, the D-loop and part of the coding region of mtDNA were characterized in 102 samples. The PCR reactions were performed with two primer pairs; after primer BLAST validation, the primer pair that amplifies the region located between positions 16453-16569 and 1-1693 reveals no risk of co-amplification. The *in silico* validation of the other primer pair, located between positions 7392 and 8921, reveals a risk of co-amplification between mtDNA and the region of chromosome 1 located between positions 46578 and 48102. However, in the NUMTs that can potentially be amplified by this primer set, the mtDNA mutations described by Aikhionbare et al. (2007) were not reported as non-identical positions in our NUMT database (Table I.2.3 of Annex I).

With regard the remaining studies, in five we were unable to assess their primers (Liu et al. 2001; Lorenc et al. 2003; Parr et al. 2006a; Petros et al. 2005; Zhou et al. 2007), since PCR primers were not reported in the original manuscripts, and after requesting them from the authors no response was obtained.

In the study by Allegra et al. (2006), the primer pair which amplifies the 10688-11500 mtDNA region also co-amplifies nDNA; specifically, it amplifies the region 99385306-99386137 on chromosome 5. Moreover, position 11203, reported as a mitochondrial mutation by Allegra et al. (2006), was reported as a non-identical position between mtDNA and the NUMT located on chromosome 5 in the region 99381642-99390749 (position signalled with † in Table I.2.3 of Annex I).

In ten different studies (Aikhionbare et al. 2007; Allegra et al. 2006; Fliss et al. 2000; Gasparre et al. 2007; Hervouet and Godinot 2006; Jin et al. 2007; Lièvre et al. 2005; Mithani et al. 2007; Polyak et al. 1998; Tan et al. 2006) the primer *in silico* validation also revealed a risk of co-amplification between mtDNA and nDNA. The twenty-seven positions described by the authors as mtDNA mutations, however, are not reported as non-identical position between mtDNA and the nuclear regions that might potentially be co-amplified.

Finally, 123 positions described in 15 studies (Aikhionbare et al. 2007; Fliss et al. 2000; Gasparre et al. 2007; Habano et al. 1999; Hervouet and Godinot 2006; Jin et al. 2007; Jones et al. 2001; Kassauei et al. 2006; Lièvre et al. 2005; Mithani et al. 2007; Nagy et al. 2003; Parrella et al. 2001; Polyak et al. 1998; Tan et al. 2006; Tzen et al. 2007) are more likely to be authentic, since the primers used to amplify the regions that encompass such mutations did not suggest the possibility of nDNA co-amplification (positions signalled with ** in Table I.2.2 of Annex I).

3.2.3 Discussion

In population and clinical genetics, as well as in the design of evolutionary, phylogenetic and epidemiological studies, it is important to have a complete overview of the total number of NUMTs, their location and variation. In this study, a NUMT database obtained using the latest version of the Human Genome – GRCh37 draft is reported. This is, to our knowledge, the first database that is based on this latest version of the human genome. Moreover, it is also the first where the non-identical positions between mtDNA and nDNA are listed for each NUMT.

In this study, 755 nuclear insertions of mtDNA origin were reported. A comparison of the total number of NUMTs in previously published works clearly shows significant discrepancies. The number of reported NUMTs ranges from 190 (Lascaro et al. 2008) to 1105 (Tourmen et al. 2002), with reports of 206 (Richly and Leister 2004), 211 (Ricchetti et al. 2004), 247 (Mishmar et al. 2004), 452 (Hazkani-Covo and Graur 2007), 612 (Woischnik and Moraes 2002) and 871 (Hazkani-Covo et al. 2010) insertions. The reasons for such discrepancies have been discussed by others previously (Lascaro et al. 2008). According to Lascaro et al. (2008), the incautious usage of

bioinformatics methods and the application of methods to an as yet not correctly assembled nuclear genome, are the main factors underlying the disagreement observed between studies. In this sense, the Human mtDNA Reference Sequence (NC_012920) was compared against the human RefSeq Genomic database at NCBI, using the RemoteBlast package in Bioperl bundle (Stajich et al. 2002) with the parameters set to restrict the search to the human organism and the E-value of 10^{-3}. Moreover, regions with less than 20 bp were excluded and only hits in the most recent version of the Human genome draft - GRCh37 primary reference assembly were considered.

Another factor that can induce differences in the number of reported insertions is whether or not the post-insertional processes were evaluated in order to differentiate original insertions from those that result from processes that alter the initially inserted sequence. In this study, post-insertional processes were not considered, since we aimed to obtain an accurate database with all the nuclear insertions and their specific variation (relatively to human mtDNA).

In accordance to other authors [for a review see Hazkani-Covo et al. (2010)], the reported frequency of insertions, as well as their length and distribution along the genome is variable. Most insertions have less than 500bp. This agrees with the most parsimonious explanation for the origin of NUMT which assumes that mtDNA would be used as a “patch” in the double-strand break repair from non-homologous end joining repair mechanism (Hazkani-Covo and Covo 2008; Hazkani-Covo et al. 2010; Leister 2005).

As previously mentioned, 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. 2008b). In this context, large insertions with a high percentage of identity with mtDNA would be particularly problematic. In fact, one of these insertions, located on chromosome 1, was previously reported to be a source of several errors and misinterpretations (Yao et al. 2008b). We characterized the 33 large insertions (with more than 500 bp) that present a percentage identity greater than 80%. Its size makes it difficult to perform primer designs to prevent co-amplification between nuclear insertions and mtDNA [chapter 1of the present thesis, Ramos et al. (2009) and Ramos et al. (2011b)].

This study reports for the first time all identical and non-identical positions between the NUMTs and the GRCh37 version of the nuclear genome and the mtDNA Reference Sequence. This information can be used as a tool in the authentication of mtDNA mutations reported in different fields. Notwithstanding, it is worth mentioning that, in this compilation, the variation of NUMTs was not considered since no studies of NUMT sequence variation had yet been performed to date. Moreover, the mtDNA variation was only taken into account for human mtDNA, although the original insertion could in fact represent mtDNA sequence from very different species.

The described NUMT database was used as a tool to assess the impact of NUMT contamination on cancer studies. Several authors detected innumerable deficiencies in the medical literature related to the analysis and interpretation of mtDNA data [Salas et al. (2005), Yao et al. (2004) among others]. Concerning cancer studies, Salas et

al. (2005) showed the consequences of poor experimental design, both in the misinterpretation of the role of mtDNA in the complex tumoral process, as well as in the comparison of results. Also, Yao et al. (2004) showed that researchers may still ignore the possibility of a NUMT contribution when a seemingly novel mtDNA sequence is encountered. Using the NUMT database, we evaluated 256 mtDNA positions that were previously reported to be mutated in cancer samples and that present zero hits in the updated mtDNA phylogeny (Van Oven and Kayser 2009). From those positions, 220 (identified in 21 studies) could represent changes in one or more nuclear insertions of mtDNA since they are non-identical between mtDNA and nDNA. Thus, we raise the hypothesis that these changes could represent nuclear variations instead of mtDNA variation. To test this hypothesis, we further validated *in silico* the primer pairs used in several original studies if they were accessible in the publication or if they were provided by the authors after request. After primer *in silico* validation, from a total of 16 works for which primers were available, only six studies (Habano et al. 1999; Jones et al. 2001; Kassauei et al. 2006; Nagy et al. 2003; Parrella et al. 2001; Tzen et al. 2007) did not evidence the possibility of nDNA co-amplification. In the remaining studies (Table I.2.3 of Annex I and references therein), some of the positions reported are located in regions that could be the result of co-amplification. However, with the exception of position 11203 reported by Allegra et al. (2006), these positions are not reported as non-identical between the potentially co-amplified mtDNA and nDNA regions. Even so, we must take into account the fact that the variability of NUMTs is unknown, and that the non-identical positions database was generated using the mtDNA

and nDNA Reference Sequences. Thus, the differences between mtDNA and nDNA reported in our database are surely underestimated and the above mentioned positions could be non-authentic mtDNA mutations. On the other hand, it is possible that these changes actually represent authentic mtDNA mutations; however, this hypothesis seems extremely difficult to test.

Using cancer studies as an example application, we aimed to highlight the applicability of the described NUMT database as a new tool to validate *in silico* mtDNA mutations described in different contexts; it is clear that there is a need to follow standardized protocols that avoid co-amplification between mitochondrial and nuclear paralogous sequences. Moreover, and due to the amount of information provided for each nuclear insertion, this database could be applied as a tool to design evolutionary, phylogenetic and epidemiological studies.

To our knowledge, this is the first work to report the non-identical positions between the mtDNA and nDNA. These positions are the basis for the application to cancer studies reported in this work, and knowledge of their distribution is useful for the authentication of mtDNA mutations reported in different kinds of study, especially those occurring in heteroplasmy.

3.3 Capítol 3:

Frequency and pattern of human mitochondrial heteroplasmy

3.3.1 Materials and methods

Samples and DNA Extraction

From two hundred and fourteen samples collected and analysed by Alvarez et al. (2010), a total of one hundred and one were used in the present study. Alvarez et al. (2010) characterized the genetic structure of a Northwest Iberia population using the mtDNA variation. The authors sequenced the HVRI of mtDNA and analysed phylogenetically informative coding region polymorphisms in order to assign each sample to a mtDNA haplogroup. To the present study, the selection of samples encompasses an equal representation of both sex and a mtDNA haplogroup distribution similar to that of an European population. All of the samples were from maternally unrelated Spanish individuals sampled in regional health centres. For all voluntary donors, appropriate informed consent and the birth places of all their known maternal ancestors up to the third generation were obtained under strictly confidential circumstances.

Total DNA from blood and mouth scrap samples was extracted using JETQUICK Blood DNA Spin Kit (Genomed, Löhne, Germany) according to the manufacturer's specifications.

MtDNA Analysis

The whole mitochondrial genome was amplified for all samples using a set of 9 primers pairs that prevents NUMTs co-amplification [chapter 1 of the present thesis, Ramos et al. (2009) and Ramos et al. (2011b)] PCR was performed as described in chapter 1 of the present thesis and by Ramos et al. (2009; 2011b). PCR products were purified using the JETQUICK PCR Purification Spin Kit (Genomed, Löhne, Germany).

All samples were fully sequenced using 32 internal primers in forward direction as well as 32 primers in reverse direction [chapter 1 of the present thesis, Ramos et al. (2009) and Ramos et al. (2011b)]. Sequence reactions were carried out using the sequencing kit Big Dye Terminator v.3 (Applied Biosystems, Foster City, USA) according to the manufacturer's specifications. Purification of DNA sequencing reactions was carried out with BigDye XTerminator® Purification Kit (Applied Biosystems, Foster City, USA) and sequences were run in an ABI Prism 3100 sequencer (Servei de Genòmica, Universitat Autònoma de Barcelona).

Detection and Authentication of Heteroplasmy

Sequences obtained were analysed using the Sequencing Analysis 5.2.0 software (Applied Biosystems, Foster City, USA), considering a value of 2% in the Mixed Base Identification option. Moreover, all sequences were visually verified and compared with others from the same run in order to detect the presence of

heteroplasmy. Only sequences with a good intensity and with background almost imperceptible were used.

Sequences were subsequently compiled and aligned in relation to the revised Cambridge Reference Sequence (CRS) (Andrews et al. 1999), using SeqScape software (Applied Biosystems, Foster City, USA).

Levels of heteroplasmy were determined as described in Santos et al. (2005b) using the height of the peaks in the electropherograms. About 6 sequences of every point heteroplasmy detected were used to calculate the average of heteroplasmy levels.

The authentication of mtDNA heteroplasmy was performed following a strategy similar to that used by Santos et al. (Santos et al. 2005b; Santos et al. 2008c), following the three main steps:

1. DNA extraction, PCR amplification, and sequencing of total mitochondrial genome were carried out; if necessary, the amplification and sequencing process was repeated to ensure sequences with a good intensity and with almost imperceptible backgrounds.

2. Individuals that appeared to present heteroplasmy in step 1 were confirmed by a second PCR amplification and sequencing.

3. To authenticate results, an independent DNA extraction, PCR amplification, and sequencing were subsequently performed for all individuals showing heteroplasmy in step 2.

Data analysis

Statistical analyses

In order to increase the total number of point heteroplasmies analysed and to perform a more robust heteroplasmy study, results published by Li et al. (2010) were also pooled with our data.

The point heteroplasmy frequencies detected both in the present study and that reported by Li et al. (2010) were determined by counting and the Bayesian 0.95 credible region (95% CR) was calculated using the SAMPLING software (V. Macaulay, personal communication).

To know if the distribution of point heteroplasmy across the mtDNA genome shows differences, a proportion test was calculated using the OpenStat version 2008. To test if there is a relation between haplogroup and the presence of point heteroplasmy, Fisher exact test was performed using the Struc program by Genepop 3.3 (Raymond and Rousset 2001). The remaining statistical tests were performed using the program SPSS ver. 15.0.1 software (SPSS 1989-2006).

Hits in the phylogeny, Population database and Conservation index (CI)

The number of hits in the phylogeny or number of occurrences was compiled from the updated mtDNA phylogeny – mit. Tree build 8 – (Van Oven and Kayser 2009) and from Soares et al. (2009).

Using the number of occurrences for each mutation reported by Soares et al. (2009), the probability of mutation was calculated as: number of hits/total number of hits. A mtDNA position was considered as a hotspot if it presents a probability of mutation ten times higher than the mean value of the probability of mutation.

In order to calculate the frequency of each variant for a particular nucleotide position, a population database of 3880 mtDNA complete sequences available from mtDNA phylogeny –mit. Tree build 8 – (Van Oven and Kayser 2009) was created. The database is in SPSS format (SPSS 1989-2006), and each mtDNA position represents a variable that allows calculate the frequency of nucleotides for each mtDNA position.

Nucleotide and amino acid conservation index (CI) were estimated for all heteroplasmic point positions calculating the percentage of variation between reference sequences of different species of the reign of metazoa. A total of 1491 nucleotid sequences and 1628 amino acid sequences were used respectively for nucleotides and amino acid CI (for specific list of species and accession number see Table I.3.1 of Annex I). Sequences were aligned using clustal W (Thompson et al. 1994) and formatting for further frequency analyses in SPSS software (SPSS 1989-2006).

Due to the difficulty to obtain a good alignment for the D-loop, an independent alignment was performed using 48 mitochondrial reference sequences of primates.

Structure prediction

To know if any implication of mitochondrial point heteroplasmy exists in the secondary structure of tRNAs and rRNA or in the protein interactions of mitochondrial genes, an analysis of structure prediction was performed.

The secondary structure prediction were developed using the RNAfold web server. Base-pair probabilities and positional entropy of any tRNA predicted structures were used to estimate the implication in the molecule. Human reference tRNA models were compiled (Putz et al. 2007), whereas rRNA models were not available. In order to know if rRNA heteroplasmic positions are located in stem or loop regions, additional analysis with software mtDNA-GeneSyn version 1.0 (Pereira et al. 2009) was performed.

The implications in the tertiary structures of proteins were performed using available crystal structures. After BLAST search, only bovine crystal structures of mitochondrial complex bc1 and IV were actually available (Iwata et al. 1998; Tsukihara et al. 1996).

The human sequence of COXI, COXII and COXIII (complex IV) and CYB (bc1 complex) were used as templates to search for the highest E-value pdb using BlastP analysis (Altschul et al. 1997). The resulting codes of bovine structures (pdb 1occ and 1bgy for complex IV and bc1, respectively) were then used in MODELLER (Sali and Blundell 1993) in order to build human structural models. The accuracy of the predicted 3D human models was estimated using three model quality: Verify 3D, SOLVX and ANOLEA (Holm and Sander 1992; Luthy et al. 1992; Melo and Feytmans 1998) .

All residue-residue bonds were calculated using Pymol software and all the structures were visualized using the same software.

3.3.2 Results

Frequency and type of heteroplasmy

From 101 complete mtDNA genome sequences, an exhaustive analysis of frequency and pattern of heteroplasmy was performed. The complete mutation report of all individuals is available on Table I.3.2 in Annex I.

For analysed samples, 63 individuals (62.37%) presented point and/or length heteroplasmy, and the remaining 37.63% were fully homoplasmic (Table 8). The 24.75% of the individuals present point heteroplasmy. There were found 4 individuals with more than one point heteroplasmy (Table 8). Twenty-eight different positions were involved; with the exception of positions 152 and 16189 that appeared to be heteroplasmic in two individuals, each position is heteroplasmic in a single individual.

The frequency of heteroplasmy did not present significant differences between sexes (χ^2=0.601, df=1 p=0.438). Moreover, the haplogroup distribution of the heteroplasmic samples is similar to that of the original sample (Fisher exact test p=0.8091).

Concerning length heteroplasmy, a 48.51% of the individuals were heteroplasmic. In this case, 33 individuals presented only one length heteroplasmy and the remaining carried more than one length and/or one point heteroplasmy (Table 8).

Table 8. Classification of the analysed individuals depending on the heteroplasmic presence (PH: point heteroplasmy, LH: length heteroplasmy, CI: confidence interval).

Classification	Number of individuals		Frequency (95% CI)
Homoplasmy	38		37.63 (28.8-47.4)
Heteroplasmy	63		62.37 (52.6-71.2)
1 PH		12	11.88 (7-19.6)
1 LH		33	32.67 (24.3-42.3)
>1 PH		2	1.98 (0.6-6.9)
>1 LH		5	4.95 (2.2-11.1)
One or more PH+LH		11	10.89 (6.2-18.5)
Total PH		25	24.75 (17.4-34)
Total LH		49	48.51 (39-58.2)
Total	101		

In table 9 the regions of the mtDNA genome showing length heteroplasmy are presented. The highest frequency of length heteroplasmy was located in the poly(C) tract of HVRII. The poly(C) tract of HVRI, and the poly(C) tract of HVRIII located between positions 568-573 of the mtDNA present a frequency of 9.1%. A 5.45% of individuals showed length heteroplasmy in poly (AC) between positions 514-525 of the HVRIII region. Finally, three individuals present length heteroplasmy in different regions of mtDNA: between positions 8272-8278, 956-965 and in position 8289 (Table 9).

Table 9. Distribution of length heteroplasmy along the mtDNA genome.

Length heteroplasmy region	Number of heteroplasmies	Frequency (95% CI)
HVRI poly(C) (16189-16193)	5	9.1 (4-19.6)
HVRII poly(C) (303-309)	39	70.9 (57.8-81.2)
HVRIII		
poly(C) (514-525)	3	5.4 (2-14.9)
poly(AC) (568-573)	5	9.1 (4-19.6)
Coding region		
956-965	1	1.8 (0.4-9.6)
8272-8278	1	1.8 (0.4-9.6)
8289	1	1.8 (0.4-9.6)
Total	55	

Pattern and distribution of point heteroplasmy along the mtDNA genome

A total of 30 point heteroplasmic positions were confirmed in our study. To obtain a more accurate pattern of point heteroplasmy, results published by Li et al. (2010) were pooled together with our data (for a detailed list of positions see table I.3.3 in Annex I).

The distribution of point heteroplasmy observed along the mitochondrial genome is showed in table 10, both for the present study and for the study of Li et al. (2010). As it can be observed, the distribution of heteroplasmy across the mtDNA genome is similar in both studies.

Combining the information of both studies, the prevalent region of mtDNA that presents point heteroplasmy was the control region, with a frequency of 31.34% of the total of detected heteroplasmies (Table 10). The remaining 68.66% were located in tRNA (7.46%) and rRNA (5.97%), and coding positions for the respiratory chain of mtDNA (55.22%) (Table 10).

In order to know if point heteroplasmy followed a differential distribution along mitochondrial genome, tests for difference between two independent proportions were performed. Statistical significant differences (after Bonferroni correction) were observed between control and coding region, being the control region the most frequent location of mtDNA heteroplasmy ($z=-4.319$, $p<0.0001$). Also, statistical differences were observed between RNA regions (rRNA and tRNA) and protein coding genes presenting heteroplasmy (complex I, III, IV and V) ($z=-5.838$, $p<0.0001$) as

well as between complex I and the rest of others complex, being the complex I the most common region in the protein coding genes zone (Complex I vs. Complex IV: z=-3.622, p=0.0001; Complex I vs. Complex III and Complex I vs. Complex V: z=-3.896, p<0.0001).

From 46 point heteroplasmies located in 45 different positions of the coding region, 37 of them were located in protein coding genes: 21 were located in positions that encode for different subunits of the NADH dehydrogenase (56.75%); 6 point heteroplasmic positions were in the cytochrome c oxidase subunits I, II and III (16.21%); 5 point heteroplasmic positions were in the ATP synthase F0 subunit 6 (13.51%); and the remaining 5 were located in the cytochrome B (13.51%) (Table 10).

Table 10. Distribution of point heteroplasmy along the mtDNA genome in the present study and reported by Li et al. (2010).

mtDNA region	Present study		Li et al. (2010)		Pooled Data	
	Number of heteroplasmies	Frequency (95% CI)	Number of heteroplasmies	Frequency (95% CI)	Number of heteroplasmies	Frequency (95% CI)
D-loop	8	26.7 (14.2-44.6)	13	35.14 (21.8-51.4)	21	31.34 (21.5-43.3)
HVRI	3	10 (3.6-25.8)	5	13.51 (6-28.1)	8	11.93 (6.2-21.9)
HVRII	5	16.7 (7.5-33.7)	8	21.63 (11.4-37.3)	13	19.40 (11.7-30.5)
Coding region	22	73.3 (55.4-85.8)	24	64.86 (48.6-78.2)	46	68.66 (56.7-78.5)
tRNA	4	13.3 (5.5-29.8)	1	2.7 (0.6-13.8)	5	7.46 (3.3-16.3)
rRNA	2	6.7 (2-21.4)	2	5.4 (1.7-17.7)	4	5.97 (2.4-14.4)
Complex I	7	23.3 (3.4-13.6)	14	37.8 (24-54)	21	31.34 (21.5-43.3)
Complex III	2	6.7 (2-21.4)	3	8.1 (2.9-21.4)	5	7.46 (3.3-16.3)
Complex IV	4	13.3 (55.4-85.8)	2	5.4 (1.7-17.7)	6	8.95 (4.2-18.2)
Complex V	3	10 (3.6-25.8)	2	5.4 (1.7-17.7)	5	7.46 (3.3-16.3)
Total	30		37		67	

These 37 point heteroplasmies placed in protein coding genes are located in 36 different positions and 16 of them represent nonsynonymous mutations (43.24% of frequency), giving a nonsynonymous:synonymous ratio of 1:3 (Table 11).

Concerning the kind of mutation, 34 of point mutations detected in heteroplasmy were purine transitions and 29 were pyrimidine transitions, only 4 transversions were found. The transition:transversion and purine:pyrimidine ratios are also presented in table 11.

Table 11. Comparison of results obtained in present study with heteroplasmy detected by Santos et al. (2008c) and with the whole mtDNA genome mutational spectrum at population level (Pereira et al. 2009).

Type of mutation	Heteroplasmy in present study[a]	Heteroplasmy in D-loop[b]	Mutational spectrum at population level[c]
Transition:Transversion			
Whole mtDNA genome	15.75:1	-	7.5:1
Control Region	21:0	57.5:1	20.8:1
Purine:Pyrimidine			
Whole mtDNA genome	1.17:1	-	1.28:1
Control Region	1:3.2	1:2.93	1:1.7
Nonsynonimous:Synonymous	1:3	-	1:1.97

[a] Pooled data from present work and from Li et al. (2010).
[b] Santos et al. (2008c).
[c] Computed using data of mtDNA mutation fixed at the individual level and polymorphic at population level, by Pereira et al. (2009).

Stability of point heteroplasmic positions

For the 67 point heteroplasmies found in pooled data, and located in 61 different positions of mtDNA genome, several analyses were performed to know the impact of mutation in the involved positions. The number of hits in the mtDNA phylogeny, the probability of mutation, the frequency in the population database, and the conservation index (CI) both at nucleotide and at amino acid level, were calculated (Table I.3.3 in Annex I).

A total of 21 point heteroplasmies were located in 16 different positions of the non-coding region (Figure 10 and Table I.3.3 in Annex I). These heteroplasmic positions have typical characteristics of non-stable position: high number of hits in the phylogeny, high frequency representation of the minority variant in database population and low percentages of conservation index (Table I.3.3 in Annex I).

Concerning coding region, 9 point heteroplasmies (in 9 positions) were located in stem regions of mitochondrial rRNA and tRNA (Table 10). From that, 2 heteroplasmic positions (1552 and 3014) are related to a high stability and conservation. These positions do not have any hit in mtDNA phylogeny, present a null representation of minority variant on database population and have a nucleotide CI over 70% (Figure 10 and Table I.3.3 in Annex I).

Thirty seven point heteroplasmies were located in 36 different positions of protein coding genes (Figure 10 and Table I.3.3 in Annex I). There are a total of 11 point heteroplasmies located in 11 stable positions (3532, 6054, 7697, 7754, 8603, 9029, 11253, 13604,

14561, 14770 and 15046). All these 11 positions present a maximum of 2 hits in the phylogeny, a minimum of 99.6% of representation in database population concerning to the majority variant and a minimum nucleotide CI of 77.25%. Moreover, nine of stable positions represent nonsynonymous mutations (3532, 8603, 6054, 7697, 7754, 9029, 11253, 13604 and 14561). From the total of 11 point heteroplasmies, six of them can be considered as highly stable positions (6054, 7697, 7754, 9029, 13604, and 15046). Besides the characteristics mentioned above (low number of hits, low representation in database population and low CI at nucleotide level), these positions present CI ≥92.9% at amino acid level (Table I.3.3 in Annex I).

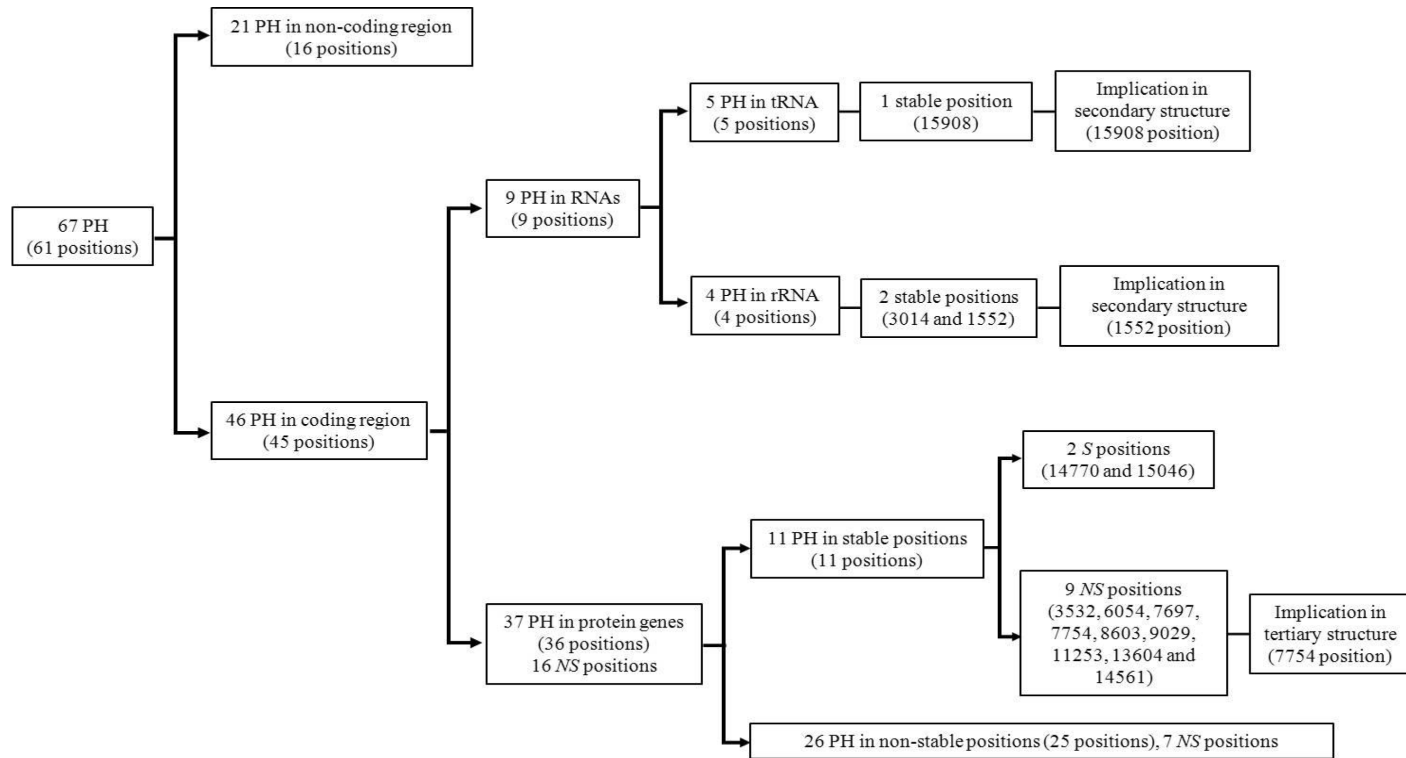


Figure 10. Schematic information of point heteroplasmies analysed in present study and reported by Li et al. (2010). Information about distribution along the mtDNA genome, kind of mutation (synonymous/nonsynonymous), stability[a] of positions and implication in secondary and tertiary structure is reported (PH:point heteroplasmy, *S*:synonymous, *NS*:nonsynonymous).

[a]Defined by: distribution in database population, number of hits in the phylogeny and nucleotide and amino acid conservation index.

In table 12 the distribution of point heteroplasmy and the number of hits in the phylogeny, considering the stability of positions, is presented. Taking the distribution of mutations at the population level, as the expected mutation distribution, there is an excess of point heteroplasmies located in positions that present 0 hits in the phylogeny. The frequency of point heteroplasmies located in no hotspot positions with a probability of mutation lower than the mean seems to be quite different compared to mutations at the population level. On the other hand, the proportion of heteroplasmies in hotspots is similar to the observed at population level.

Table 12. Distribution of point heteroplasmies and of hits in the phylogeny considering the stability of position (Mean value of the probability of mutation[a] = 6.034×10^{-5}). *N*: total number.

mtDNA positions	Point heteroplasmy		Hits in the phylogeny[c]	
	N	**Frequency (95% IC)**	**N**	**Frequency (95% IC)**
0 Hits	13	19.40 (11.7-30.5)	0	0 (0.00-0.00)
Hotspot[b]	29	43.28 (32.1-55.2)	4089	38.29 (37.8-38.7)
No hotspot ≥ mean	8	11.94 (6.5-21.9)	2872	26.89 (26.1-27.7)
No hotspot < mean	17	25.37 (16.5-37)	3719	34.82 (33.9-35.7)

[a] Probability of mutation: number of hits/total number of hits.
[b] Hotspot: probability of mutation ten times higher than the mean value.
[c] Soares et al. (2009).

In order to know the impact of mutation in the stability of secondary and tertiary conformation of tRNA, rRNA and proteins encoded by mtDNA, a prediction of different structures with the wildtype and mutant variant were performed and afterwards are presented those that imply changes in secondary or tertiary structures (Figure I.3.1 of Annex I).

In the case of tRNAs, it seems that point heteroplasmic position 15908 located in threonine tRNA implies a reduction in the number of residue-residue bond that evolve one of the stems of this tRNA (Figure 11). Moreover, this position presents two hits in the phylogeny, a representation of minor variant of 0.1% in population database and a CI of 17.78%, representing a stable position. Concerning rRNA, half of positions located in 12S and 16S are in stem regions, and the rest of them in loops. The heteroplasmy located in 1552 position imply a reduction in the stability of bonds of this stem region. Moreover, entropy information of bonds in this position is reduced (Figure 12).

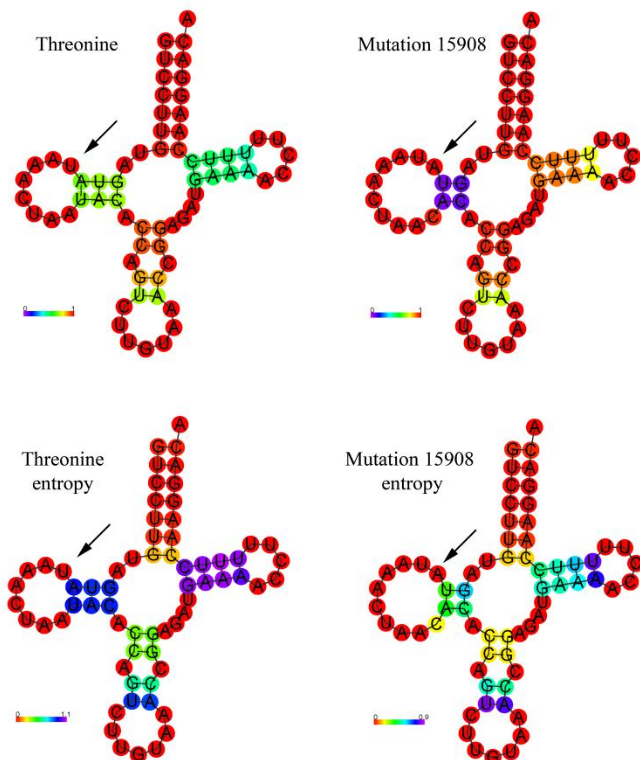


Figure 11. Secondary structure prediction of threonine tRNA. Implication of point heteroplasmic position 15908. The structure below is colored by base-pairing probabilities, for unpaired regions the colour denotes the probability of being unpaired. Structure drawing encoding positional entropy is also reported.

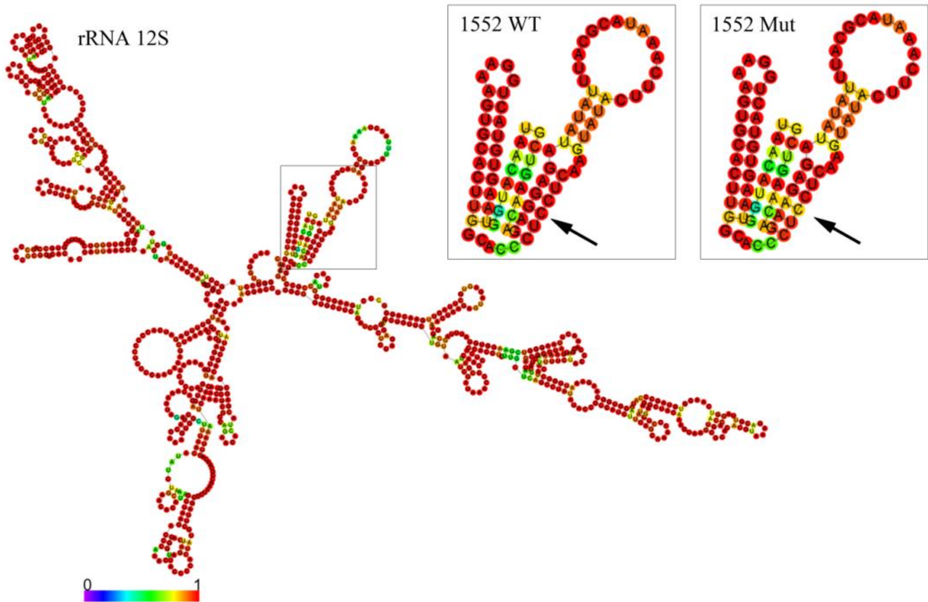


Figure 12. Secondary structure prediction of 12S RNA. Implication of point heteroplasmic position 3014. Detailed view of wildtype and mutated position is showed. The structure below is colored by base-pairing probabilities, for unpaired regions the colour denotes the probability of being unpaired. Structure drawing encoding positional entropy is reported in Figure I.3.1 in Annex I. (WT:wildtype, Mut:mutated).

Concerning the implications in the tertiary structures of proteins, only information of complex IV and bc1 was available. Four nonsynonimous heteroplasmic positions (6052, 7697, 7754 and 15314) were located in these regions (Figure I.3.1 in Annex I). It seems that heteroplasmic position 7754, which implies an amino acid substitution (Asp57Asn) in COXII region, has some implication in the interactions between amino acids in tertiary structure. The mutated amino acid is located in the centre of an α-helix and the

substitution of an acidic by a polar amino acid implies the loss of interactions with the neighbours amino acid 56S and 58A (Figure 13). Moreover, as reported above, the heteroplasmic position 7754 is one of the most conserved and stable positions (Table I.3.3 in Annex I).

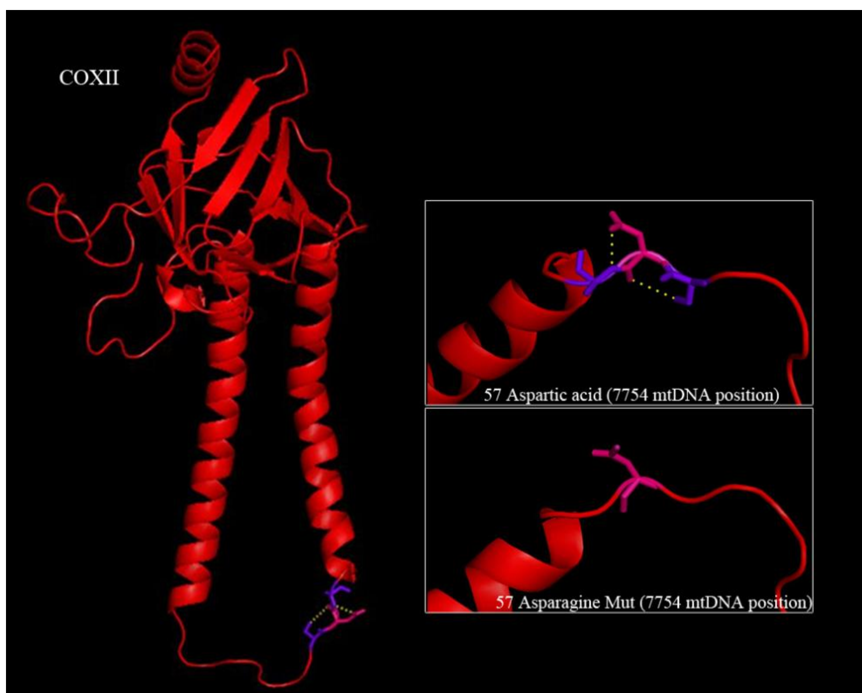


Figure 13. Three-dimensional model of human COXII complex. Implication of amino acid substitution Asp57Asn due to a point heteroplasmic position 7754. Detailed view of wildtype (Asp) and mutated (Asn) is also showed.

3.3.3 Discussion

Frequency and type of heteroplasmy

The frequency of heteroplasmy at the population level has been estimated for the hypervariable regions of mtDNA (De Camargo et al. 2011; Irwin et al. 2009; Santos et al. 2008c) and at the moment only exists one previous study that use deep sequencing of the complete mitochondrial genomes to analyse mtDNA heteroplasmy (Li et al. 2010).

In the present study, 62.37% of individuals are heteroplasmic at the population level. The 48.51% of analysed individuals presented length heteroplasmy and 24.75% of individuals carry out point heteroplasmy. The frequency of length heteroplasmy obtained in the control region [4.95% for HVRI (95% CI, 2.2-11.1%) and 38.61% for HVRII (95% CI, 29.7-48.4%)] was significantly lower than that reported by Santos et al. (2008c) [17.145% for HVRI (95% CI, 12.3-22.93%) and 64.76% for HVRII (95% CI, 57.89-71.21%)]. These differences could not be due to methodological approaches since both studies were performed using the same methodology and analytic conditions. It is possible that the differences could be related to the population origin of analysed samples since other authors

reported differences in point heteroplasmy frequencies in different populations (Irwin et al. 2009).

Concerning the frequency of individuals with point heteroplasmy in the control region, our results [7.9% (95% CI, 4.1-14.9%)] are slightly higher than those reported by Santos el at. (2008c) [3.81% (95% CI, 1.66-7.37%)] however differences are not significant. For the total mtDNA genome, comparing our results of point heteroplasmy frequency [24.75% (95% CI, 17.4-34%)] to those reported by Li et al. (2010) [24.42% (95% CI, 17.9-32.5%)] no differences were observed. Our results put in evidence the high frequency of heteroplasmy at the population level. Taking into account the high frequency of heteroplasmy and the high level of individuals that present more than one heteroplasmy, it appears that mitochondrial recombination could be a source of new mtDNA variants, increasing thus the variability in the mitochondrial genome at individual level and ultimately, at population level. On the other hand, recombination could avoid the accumulation of deleterious mutations along maternal lineages preventing their extinction as a result of Muller's ratchet (Wallace 2007).

Different technologies have been used for heteroplasmy detection. The majority of work published so far used the automated sequencing, the same method used in the present study. On the other hand, next-generation sequencing has been used by Li et al. (2010). Although the sequencing method is different, both studies report similar point heteroplasmy frequencies and similar distribution of point heteroplasmy across the mtDNA genome. In the present study, special attention was taken to avoid co-amplification between

mtDNA and nDNA, as well as to use an efficient protocol to validate point heteroplasmy. In spite of it, and in accordance with others (Santos et al. 2008c), our study demonstrates that heteroplasmy levels of 5% can be detected with confidence using the automated sequencing system if a good sequencing strategy and an accurate procedure of heteroplasmy validation are used. And comparing our results to those reported by Li et al. (2010), the methodology used in the present study could be as precise as massive sequencing.

No significant results have been found in the relation between sex or haplogroup and the distribution of point heteroplasmy in human populations, a result similar to that obtained by Camargo et al. (2011). These results suggest that evolutionary forces that control the levels of heteroplasmy act independently from sex and mtDNA haplogroups.

Several investigations have noted an association between increases in human mitochondrial heteroplasmy and aging. This includes pathogenic heteroplasmies (Cortopassi et al. 1992) but also seemingly benign deletions or polymorphic variants (Calloway et al. 2000; Da Costa et al. 2007; Michikawa et al. 1999; Wang et al. 2001). Some studies have suggested that some heteroplasmies may play a role in longevity (Rose et al. 2007; Salvioli et al. 2008). In our study a total of 30 point heteroplasmies have been detected in blood samples of 101 individuals. The age of these individuals ranges between 9 and 92 years, with an average age of 58.9 years. The average age of heteroplasmic individuals (58.6 years) and non heteroplasmic individuals (59 years) did not present significant

differences. Thus, it seems that our data do not show any relation between mitochondrial heteroplasmy and aging.

The heteroplasmic state of an individual could have arisen either via transmission from the mother (i.e., through the germline) or via a new somatic mutation occurring within the individual. In order to know the origin of point heteroplasmies detected in present study, further analyses were performed. Mouth scrap samples of the detected heteroplasmic individuals were used as a reference of another tissue with different embryonic origin (blood as a reference tissue of mesodermic origin and mouth scrap of ectodermic origin). From the 24 available epithelial samples, 21 presented once again the point heteroplasmy, whereas only 3 did not show any signal of mix variant. The presence of heteroplasmy in more than one tissue with different blastodermic origin indicates that probably this heteroplasmy is germinal. Thus, 87.5% of the point heteroplasmies detected in the present study are germinal and only 12.5% are somatic. Our data are in accordance with that reported by Santos et al. (2005b). In that work, the authors confirmed the point heteroplasmic origin by the analysis of the mother and other relatives of the heteroplasmic individuals and a frequency of 61.6% germinal heteroplasmies was reported. The low representation of somatic point heteroplasmies in our data may have biased the detection of a possible relation between heteroplasmy and aging previously presented. In fact, the age average of individuals presenting somatic heteroplasmy (72.2 years) is higher than the reported for the individuals presenting germinal heteroplasmy (56.8 years). These observations put forward that most of heteroplasmies are germinal and only a small fraction are somatic. Notwithstanding, somatic

heteroplasmies appear to be related to aging/longevity since these heteroplasmies appear in older individuals. On the other hand, the high levels of germinal point heteroplasmies would imply that individuals carried out a mutational background that would increase the risk or protect from mitochondrial related pathologies. Moreover, this genetic variation would be selected for the sequence evolution, implicating the final fixation or loss of a specific mtDNA variant.

Pattern and stability of point heteroplasmic positions along the mtDNA genome

Concerning the kind of mutations detected, the 94% of point heteroplasmies involved transitional variants, yielding a transition:transversion ratio of 15.75:1 (Table 11). This high representation of transitions are in accordance to the mutational spectrum at population level reported by Pereira et al. (2009) (ratio 7.5:1). However, in the present study there are twice more transitions in heteroplasmy than at the population level, a similar result to that obtained for D-loop in previous studies (Pereira et al. 2009). These provide evidences of an excess of transitions (or deficit of transversions) in the mutational spectrum in heteroplasmy.

A frequency of 64% of point mutations detected in heteroplasmy in the whole mitochondrial genome involved purines, representing a ratio purine:pyrimidine of 1.17:1, according well with the ratio 1.28:1 at population level (Pereira et al. 2009). Classifying the reported results by coding and non-coding region and according with others (Irwin et al. 2009; Pereira et al. 2009; Santos et al.

2008c), high proportion of mutations involving pyrimidines (76.2%) was obtained for the control region, both considering mutations in heteroplasmy and at the population level (Table 11). This differential result between mitochondrial regions could not be due to a differential base composition, since there is a similar ratio 1:1.31 of purine:pyrimidine in both regions, and the causes of such difference deserves further investigations.

As previously reported, mutations are the root cause of heteroplasmy and also provide the basis for sequence evolution, so it is reasonable to expect a correlation between sites that experience high rates of heteroplasmy and those that evolve quickly within populations (Irwin et al. 2009). However, it seems that a different pattern is established for point heteroplasmy, since there is an excess of point heteroplasmies located in positions that present 0 hits in the phylogeny.

A total of 16 different point heteroplasmies has been detected in nonsynonymous positions, representing a ratio nonsynonymous:synonymous of 1:3. These results are in accordance with values reported by others at the population level (ratio 1:1.97) (Pereira et al. 2009). Pereira et al. (2009) found high levels of nonsynonymous polymorphisms in ATP6 and ATP8 regions of mtDNA and low levels in COXI, ND4 and ND5 regions. In the present work, the distribution between regions is different, probably due to the low number of nonsynonymous positions detected. The same authors show that the group of neutral amino acid V, I, A, M, T present the higher proportion of all possible changes observed in human population. Our data show that from the 16 nonsynonymous

heteroplasmic positions, 7 of these are located in amino acid of the reported group, and also all of them imply an amino acid change with the same polarity/acidity (neutral to neutral), suggesting that the new residue could maintain the interactions with the other residues in order to preserve the tertiary structure of the protein.

Concerning tRNAs, one of the most conserved positions where heteroplasmy was found is the 15908. This change implies a reduction in the stability of the molecule due to a length reduction of one of the stems of the threonine tRNA. There has been reported that some polymorphisms in the human population are located in positions identified as 100% and >90% conserved in mammalian tRNAs (Kivisild et al. 2006). Moreover, Pereira et al. (2009) and according to Kivisild et al (2006), found the threonine as the tRNA with the highest amount of polymorphisms. Along with the point heteroplasmic position 15908 there are three more heteroplasmies located in other tRNAs, all of them found in stem regions. These results are in contrast with Pereira et al. (2009) that found a ratio of stem:loop of 0.99, according to the principal function of stem regions in the maintenance of the secondary structure.

Eleven point heteroplasmies detected in protein coding genes are located in stable positions (3532, 6054, 7697, 7754, 8603, 9029, 11253, 13604, 14561, 14770 and 15046), presenting a representation in mtDNA phylogeny near zero, low or null levels of minor variant in population database and high levels of CI. Moreover, from those heteroplasmic positions located in complexes for which the mitochondrial crystal structures are known, the 7754 position, which implicates an amino acid substitution Asp70Asn, have a direct

implication in the tertiary structure of COXII, implying the loss of interactions with nearby amino acids.

In the present study, two heteroplasmies located in stable positions (7679 and 11253) have been related to mitochondrial pathologies. The point heteroplasmy located in 7697 position and its amino acid change implication (valine to isoleucine) have been reported as a genetic factor that indicate a susceptibility to Hypertrophic cardiomyopathy (HCM) in Chinese Han ethnic population (Wei et al. 2009). Moreover, the point heteroplasmy located in 11253 position (amino acid change isoleucine to threonine) has been related to have a direct implication on Leber's hereditary optic neuropathy (LHON) (Fauser et al. 2002; Leo-Kottler et al. 2002). Although the mutant variant frequency of this individual is relatively high (76%), in the majority of LHON patients and family members, the pathogenic mtDNA mutation is homoplasmic; thus, at this stage the mutation observed in this individuals does not have phenotypic implications (Howell et al. 2000; Smith et al. 1993).

Taking up the previously mentioned about the excess of transitions in the mutational spectrum in heteroplasmy, it seems that some evolutionary force may be acting to lower these at population level. Most probably these mutations will be finally eliminated by genetic drift or by selection. In fact, all of the point heteroplasmies located in stable positions could have a greater impact in the viability of mitochondria survival, suggesting that purifying selection must be operating on some heteroplasmies to prevent their fixation within individuals. Although purifying selection has been proposed by others (Elson et al. 2004; Rand 2001; Ruiz-Pesini and Wallace 2006;

Santos et al. 2008c), no previous evidence involving heteroplasmy has been found, but in accordance with Li et al (2010), actually there are not enough heteroplasmic studies to reflect the purifying selection actuation.

This is the first study that analyses the frequency of heteroplasmy in healthy population, carrying out an evolutionary evaluation of the detected changes. This work provides a new perspective which would be important for medical, evolutionary, and forensic proposes.

4. DISCUSSIÓ GENERAL

Disseny i validació de primers per a l'amplificació i seqüenciació del DNAmt

L'amplificació i seqüenciació del DNAmt s'utilitza rutinàriament en una gran diversitat d'estudis. Tot i així, la interpretació dels resultats, sobretot pel que fa a l'estudi de l'heteroplàsmia mitocondrial, es pot complicar per la possibilitat de co-amplificar DNAmt i DNAn (Parr et al. 2006b; Yao et al. 2008b). Per aquest motiu, i donat que no existeixen a la bibliografia *primers* que permetin amplificar la totalitat del genoma mitocondrial, evitant alhora l'amplificació de NUMTs, la primera tasca duta a terme va ser dissenyar i validar uns *primers* que permetessin amplificar el genoma mitocondrial, a la vegada que evitaven la possible co-amplificació amb NUMTs.

El disseny de *primers* es va fer tenint en compte l'esborrany hg18 del genoma humà, el més actual en l'inici de la Tesi. Durant el disseny de les parelles de *primers* es van detectar dues regions del DNAn de gran mida i altament similars amb el DNAmt. Aquestes regions es localitzen al cromosoma 1 (regió 3914–9074 del DNAmt) i al cromosoma 5 (regió 9582–14479 del DNAmt). La regió que se situa al cromosoma 1 ja havia estat descrita anteriorment (Lascaro et

al. 2008) com una regió nuclear altament similar al DNAmt. Però pel que fa a la regió que es localitza al cromosoma 5, aquesta no ha estat reportada amb anterioritat, podent representar una inserció nuclear d'origen mitocondrial no descrita en treballs anteriors sobre NUMTs. De fet, aquesta seqüència s'ha pogut identificar i caracteritzar en el moment en que es va dur a terme la compilació de NUMTs.

La validació dels *primers* d'amplificació dibuixats, mitjançant una mostra d'esperma sotmesa a una lisi preferencial, ha permès demostrar que les nou parelles de *primers* dissenyades amplifiquen específicament DNAmt i no permeten la co-amplificació amb DNAn.

Donat que no hi havia *primers* descrits a la bibliografia que permetessin una amplificació selectiva del DNAmt, evitant alhora possibles co-amplificacions entre el DNAmt i el DNAn, aquest treball va publicar-se a inicis de l'any 2009 (Ramos et al. 2009). Els *primers* presentats poden ser una eina molt útil per a qualsevol projecte on calgui realitzar una amplificació i seqüenciació completa del DNAmt, sent molt adient per al *screening* de l'heteroplàsmia mitocondrial tant en mostres d'individus sans com malalts, així com en determinacions forenses. D'altra banda, l'estricta metodologia emprada per dissenyar-los i validar-los pot ser utilitzada com a referència per a qualsevol estudi de DNAmt que requereixi un disseny de *primers*.

Com s'ha dit anteriorment, aquests *primers* [capítol 1 de la present Tesi i Ramos et al. (2009)] van ser validats *in silico*

(mitjançant BLAST) utilitzant la seqüència de referència hg18, però al febrer de 2009, va generar-se un nou esborrany del genoma humà GRCh37. Com a conseqüència, s'han hagut de validar, *in silico,* novament els *primers*, utilitzant aquesta vegada la nova seqüència de referència del genoma humà GRCh37.

D'acord amb l'actualització de l'esborrany del genoma humà, un dels NUMTs més problemàtics (situat al cromosoma 1) sembla ser més gran del que s'havia descrit. Amb l'antiga referència del genoma humà aquest NUMT tenia un 98% de similitud amb la regió del DNAmt compresa entre les posicions 3914–9074, però degut a l'actualització, aquesta regió s'amplia en més de 700 pb (regió 3911-9755 del DNAmt), conservant la mateixa similitud (Andrews et al. 1999). Aquesta modificació provoca que un dels *primers* dissenyats (situat entre les posicions 6511-9220 del DNAmt) perdi la capacitat d'amplificació exclusiva del DNAmt i per tant, possibiliti la co-amplificació amb el DNAn.

L'estudi d'aquest NUMT ha permès identificar-lo com un NUMT específic d'espècie, ja que no el comparteixen la resta de primats. Malauradament però, no es coneix encara la variabilitat (presència/absència) que pot tenir un NUMT a la població humana i per tant, no es pot descartar la possible co-amplificació entre el DNAmt i DNAn. Per aquest motiu s'ha redissenyat una de les parelles de *primers*. Degut a la gran mida del NUMT, aquesta parella de *primers* s'ha hagut de localitzar en una zona de baixa similitud entre el DNAmt i el DNAn, col·locant l'extrem 3' de cadascun dels *primers* en una posició no idèntica entre el DNAmt i el DNAn.

Aquesta actualització va publicar-se l'any 2011 (Ramos et al. 2011b).

Actualització de la base de dades de NUMTs i aplicació en estudis de càncer

Els NUMTs poden ser una eina molt útil en estudis evolutius i de poblacions. D'altra banda, degut a l'elevada similitud amb el DNAmt, els NUMTs poden ser una font de contaminació en estudis basats en l'amplificació per PCR del DNAmt. Per tant, conèixer de primera mà el nombre total d'aquestes insercions, la seva localització i variació és important per a dur a terme un bon disseny de qualsevol tipus d'estudi, ja sigui filogenètic, epidemiològic, evolutiu, poblacional o clínic. Actualment però, tot i existir una gran quantitat de bases de dades que recullen aquesta informació referent als NUMTs, totes elles s'han realitzat mitjanant una versió desactualitzada del genoma humà de referència. Per aquests motius, un dels objectius de la present Tesi, ha estat realitzar una exhaustiva compilació de les insercions nuclears d'origen mitocondrial basada en la darrera versió de l'esborrany del genoma humà (GRCh37). El treball resultant s'ha publicat recentment a la revista Mitochondrion (Ramos et al. 2011a).

En aquesta base de dades es presenten 755 insercions. Si comparem el número total de NUMTs amb els reportats en estudis publicats prèviament, aquest nombre pot variar des de 190 (Lascaro et al. 2008) a 1105 (Tourmen et al. 2002), amb publicacions que reporten 206 insercions (Richly and Leister 2004), 211 (Ricchetti et

al. 2004), 247 (Mishmar et al. 2004), 452 (Hazkani-Covo and Graur 2007), 585 (Simone et al. 2011), 612 (Woischnik and Moraes 2002) i 871 (Hazkani-Covo et al. 2010).

D'acord amb Lascaro et al. (2008), els motius principals d'aquestes discrepàncies són la disparitat que hi ha en els mètodes bioinformàtics utilitzats, així com en els paràmetres tinguts o no en compte i a l'aplicació de genomes nuclears que encara no estan acoblats correctament.

Un altre factor que por induir diferències en el nombre total de NUMTs és el fet de considerar, o no, els processos post-insercionals que pot patir una inserció. Aquest factor pot servir per saber si un NUMT representa una inserció original o si és el resultat d'un procés que ha alterat la seqüència inicial com ara la duplicació. En la base de dades que es presenta en aquesta Tesi, i a fi de poder obtenir una base de dades que contingués totes les insercions nuclears, així com la seva pròpia variabilitat relativa al DNAmt, els processos post-insercionals no s'han tingut en compte.

La freqüència de les insercions presentades, així com la seva mida i distribució al llarg del genoma, és variable, destacant el fet que la majoria d'aquestes insercions són menors de 500 pb. Aquests resultats estan d'acord amb altres autors (Hazkani-Covo et al. 2010) i amb l'explicació més parsimoniosa sobre l'origen del NUMTs, que assumeix que el DNAmt podria ser utilitzat com a pegat en la reparació dels trencament de doble cadena (DSBR) pel mecanisme d'unió d'extrems no homòlegs (NHEJ) (Hazkani-Covo and Covo 2008).

D'altra banda, i per primera vegada, juntament amb la base de dades d'insercions nuclears, es presenten les posicions no idèntiques entre NUMTs (mitjançant el darrer esborrany del genoma humà GRCh37) i la seqüència de referència del DNAmt. Aquesta informació pot ser molt útil per autenticar mutacions del DNAmt que han estat publicades en diferents àrees d'estudi. Cal mencionar però, que en aquest recull no s'ha tingut en compte la variació de presència/absència que pot tenir un NUMT en la població. A més, pel que fa a la variació del DNAmt, només s'ha tingut present a nivell humà, però una inserció original podria representar la seqüència d'un DNAmt d'una espècie molt diferent.

Com un exemple d'aplicació d'aquesta base de dades, s'ha avaluat l'impacte de la contaminació per NUMTs en estudis de càncer. Aquest fet és degut a l'elevada similitud que tenen els NUMTs amb el DNAmt, sent una font de contaminació en els treballs de DNAmt basats en la seva amplificació per PCR (Parr et al. 2006b; Yao et al. 2008b). De fet, actualment estan publicats un seguit d'estudis que posen de manifest innumerables deficiències en la literatura mèdica relacionada amb la interpretació de les dades en anàlisis del DNAmt [Salas et al. (2005), Yao et al. (2004), entre d'altres]. Aquests treballs demostren les conseqüències que pot tenir un pobre disseny experimental i el desconeixement de l'existència dels NUMTs com a possible font de contaminació.

De les 256 posicions del DNAmt avaluades, que han estat prèviament reportades com a mutacions relacionades amb càncer i que no apareixen com a polimòrfiques a l'espècie humana, 220 poden representar canvis en una o més insercions nuclears, ja que

estan localitzades en posicions no idèntiques entre el DNAmt i DNAn. Després de dur a terme la validació *in silico* dels *primers* utilitzats en els diferents estudis de càncer (*primers* disponibles en 16 estudis d'un total de 21 treballs analitzats), només en 6 treballs (Habano et al. 1999; Jones et al. 2001; Kassauei et al. 2006; Nagy et al. 2003; Parrella et al. 2001; Tzen et al. 2007) no es van observar evidències de possibles co-amplificacions. De tota manera, amb l'excepció de la posició 11203 descrita per Allegra et al. (2006), la resta de posicions no estan reportades com a posicions no idèntiques entre el DNAmt i el DNAn, en les insercions que podrien ser potencialment co-amplificades.

Utilitzant els estudis de càncer com a exemple, s'ha evidenciat l'aplicabilitat de la base de dades de NUMTs com una nova eina per validar *in silico* les mutacions de DNAmt descrites en diferents contextos. D'altra banda, l'exemple posa de manifest la necessitat d'estandarditzar protocols que evitin la co-amplificació de DNAmt i DNAn.

Tenir a l'abast una base de dades de NUMTs, així com dur a terme un correcte disseny de *primers* com el que s'ha descrit en aquesta Tesi són dues eines a tenir en compte en el moment d'iniciar un treball centrat en el DNAmt, ja que proporcionen una font d'informació molt útil. Existeix la possibilitat però, que no es pugui evitar la co-amplificació, com és el cas dels estudis de DNA antic o forenses on la mida dels fragments amplificats no permet situar els *primers* fora de les regions conflictives entre DNAmt i DNAn. En aquests casos seria important tenir com a mínim constància de les posicions no idèntiques que es presenten a la base de dades de

NUMTs i així poder controlar la possible contaminació en el moment d'interpretar els resultats.

Freqüència i patró de l'heteroplàsmia mitocondrial humana

L'estudi de l'heteroplàsmia mitocondrial s'ha realitzat utilitzant els *primers* dissenyats en la present Tesi per amplificar i seqüenciar el genoma mitocondrial de 101 individus. L'estudi posterior de l'heteroplàsmia mitocondrial a permès reafirmar l'amplificació selectiva del DNAmt, ja que no s'ha detectat co-amplificació entre DNAmt i DNAn en cap de les mostres.

La freqüència d'heteroplàsmia a nivell poblacional ha estat estimada en diverses ocasions pel que fa a la regió control del DNAmt (De Camargo et al. 2011; Irwin et al. 2009; Santos et al. 2008c), actualment però, només existeix un estudi que ho realitzi per tot el genoma mitocondrial (Li et al. 2010).

La metodologia utilitzada per dur a terme l'estudi de l'heteroplàsmia mitocondrial humana ha estat la seqüenciació automàtica, sent la tècnica més utilitzada en la gran majoria de treballs centrats aquest tema. En comparació amb la tècnica de seqüenciació massiva aplicada per Li et al. (2010) i donat que tots dos treballs reporten freqüències similars d'heteroplàsmia puntual, s'ha fet palès que les dues metodologies presenten una sensibilitat similar. Val a dir que això ha estat gràcies a l'exhaustiu disseny de *primers* així com a l'aplicació d'un eficient protocol de validació de l'heteroplàsmia puntual. Conforme això i d'acord amb Santos et al.

(2008c), el present estudi demostra la possibilitat de detectar amb seguretat les heteroplàsmies puntuals amb un nivell d'un 5%.

En el present estudi s'ha determinat que un 62.37% de la població presenta heteroplàsmia mitocondrial, concretament, s'han detectat un 58.51% d'individus amb heteroplàsmia de longitud i un 24.75% amb heteroplàsmia puntual. Aquests resultats posen de manifest l'elevada freqüència d'heteroplàsmia mitocondrial a nivell poblacional. Tenint en compte aquesta elevada freqüència i l'elevat nombre d'individus que presenten més d'una heteroplàsmia mitocondrial, sembla ser que la recombinació mitocondrial podria actuar com a font generadora de noves variants de DNAmt, augmentant així la variabilitat en el genoma mitocondrial a nivell individual i en última instància, a nivell poblacional. D'altra banda, la recombinació podria impedir l'acumulació de mutacions perjudicials al llarg dels llinatges materns, prevenint-ne l'extinció com a resultat del trinquet de Muller (Wallace 2007).

Pel que fa a la relació entre el sexe o l'haplogrup de l'individu i l'heteroplàsmia mitocondrial en població humana, no s'han trobat resultats significatius. Aquests resultats estan d'acord amb altres autors (De Camargo et al. 2011), suggerint que les forces evolutives que controlen els nivells d'heteroplàsmia actuen independentment del sexe i de l'haplogrup mitocondrial.

Un seguit de treballs han detectat una associació entre l'augment d'heteroplàsmies mitocondrials i l'edat (Calloway et al. 2000; Cortopassi et al. 1992; Da Costa et al. 2007; Michikawa et al. 1999; Wang et al. 2001). D'altres, han suggerit que certes heteroplàsmies podrien jugar un paper fonamental en la longevitat

(Rose et al. 2007; Salvioli et al. 2008). L'edat dels individus analitzats en el present estudi varia entre 9 i 92 anys, amb una mitjana d'edat de 58.9 anys. La mitjana d'edat dels individus que tenen heteroplàsmia mitocondrial (58.6 anys) i dels que no en presenten (59 anys) no té diferències significatives. Per tant, els nostres resultats no semblen mostrar cap relació entre la presència d'heteroplàsmia i l'edat.

L'heteroplàsmia mitocondrial d'un individu pot tenir origen matern (i per tant, venir a través de la línia germinal) o pot resultar de l'aparició de mutacions somàtiques que es donen en el sí de l'individu. Per conèixer quin és l'origen de les heteroplàsmies puntuals detectades en el present estudi, van realitzar-se anàlisis posteriors dels individus que presentaven heteroplàsmia mitocondrial. Es van utilitzar mostres de raspat bucal d'aquests individus com a referència d'un teixit amb diferent origen embrionari (la sang seria el teixit de referència d'origen mesodèrmic i l'epiteli bucal seria el d'origen ectodèrmic). De les 24 mostres de raspat bucal disponible, 21 van presentar de nou l'heteroplàsmia puntual detectada prèviament en sang, i només en 3 no hi va haver senyal d'una mescla de variants. La presència d'una heteroplàsmia mitocondrial en més d'una fulla blastodèrmica indicaria que l'origen d'aquesta heteroplàsmia és probablement germinal. D'aquesta manera, el 87.5% de les heteroplàsmies puntuals analitzades tindrien un origen germinal, mentre que només un 12.5% serien somàtiques. Aquesta elevada freqüència d'heteroplàsmies d'origen germinal ja ha estat reportada per altres autors (Santos et al. 2005b), mitjançant l'anàlisi de la mare i altres familiars de l'individu heteroplàsmic.

La baixa representació d'heteroplàsmies d'origen somàtic, podria esbiaixar la possible relació entre l'heteroplàsmia i l'edat. De fet, la mitjana d'edat dels individus que presenten heteroplàsmia d'origen somàtic (72.2 anys) és força superior a la dels individus amb heteroplàsmia germinal (56.8 anys), la qual cosa podria relacionar l'heteroplàsmia somàtica amb processos d'envelliment i/o longevitat.

Aquestes observacions posen de manifest que la majoria de les heteroplàsmies mitocondrials són germinals i només una petita fracció són somàtiques. L'elevat nivell d'heteroplàsmia germinal, implica que els individus porten un *background* mutacional, que podria augmentar el risc, o protegir-los, enfront a patologies relacionades amb disfunció mitocondrial.

El 94% de les heteroplàsmies mitocondrials detectades en el genoma mitocondrial comprometen transicions, donant una raó transició:transversió de 15.75:1. Aquesta elevada representació està en consonància amb el que es presenta a nivell poblacional, reportat per Pereira et al. (2009) (ratio7.5:1). Val a dir, que les freqüències en heteroplàsmia doblen a les que es presenten a nivell poblacional, evidenciant un excés de transicions (o dèficit de transversions) en l'espectre mutacional de l'heteroplàsmia.

D'altra banda, el 64% de les heteroplàsmies detectades impliquen canvis purínics, representant una raó purina:pirimidina de 1.17:1, molt en consonància amb la raó de 1.28:1 reportada a nivell poblacional (Pereira et al. 2009). Si només tenim en compte els resultats obtinguts pel que fa la regió control, i d'acord amb altres autors (Irwin et al. 2009; Pereira et al. 2009; Santos et al. 2008c), es

detecten elevades freqüències de pirimidines (76.2% present treball). Cal destacar que la composició nucleotídica és la mateixa en tot el genoma mitocondrial (raó purina:pirimidina 1.31:1), per tant, la diferent composició de bases entre regions del DNAmt no podria ser la causa d'aquesta inversió que es detecta tant en l'espectre mutacional de l'heteroplàsmia com a nivell poblacional, per tant, les causes d'aquesta diferència haurien de ser investigades en futurs estudis.

Les mutacions són la causa fonamental de l'heteroplàsmia alhora que proporcionen la base per l'evolució del genoma, per tant és raonable esperar una correlació entre les posicions que experimenten elevades taxes d'heteroplàsmia amb aquelles que evoluciones amb rapidesa dintre de les poblacions (Irwin et al. 2009). De tota manera, sembla que hi ha un patró diferencial per a l'heteroplàsmia puntual, ja que s'ha detectat un excés d'aquestes en posicions que presenten zero *hits* a la filogènia.

S'han detectat un total de 16 heteroplàsmies puntuals en posicions no sinònimes, representant una raó no-sinònim:sinònim de 1:2.2. Aquests resultats estan d'acord amb altres descrits a nivell poblacional (raó de 1:1.97) (Pereira et al. 2009). Pereira et al. (2009), reporta que els nivells més elevats de mutacions no sinònimes es troben en les regions ATP6 i ATP8 del DNAmt i els més baixos en les regions del COXI, ND4 i ND5. Respecte això, en el present treball no es detecta la mateixa distribució, probablement degut al baix nombre de posicions no sinònimes analitzades. Així, caldria ampliar el nombre d'estudis d'heteroplàsmia de tot el genoma. Aquests mateixos autors mostren que els aminoàcids neutres V, I, A,

M i T manifesten les proporcions més elevades pel que fa les mutacions no sinònimes. En els nostres resultats, de les 16 posicions heteroplàsmiques no sinònimes que s'han detectat, 7 d'elles estan localitzades en el grup d'aquests aminoàcids; a més, totes impliquen un canvi aminoacídic amb una polaritat similar (neutre a neutre), suggerint que el nou residu aminoacídic podria seguir mantenint les interaccions amb els aminoàcids veïns per preservar l'estructura terciària de la proteïna.

Respecte dels tRNAs, una de les posicions més conservades on s'ha detectat heteroplàsmia, és a la posició 15908. Aquest canvi provoca una reducció en l'estabilitat de l'estructura secundària del tRNA de la treonina degut a una disminució en la mida d'un dels braços. Alguns estudis han descrit que certs polimorfismes de la població humana es localitzen en posicions altament conservades dels tRNAs (Kivisild et al. 2006). Tanmateix, Periera et al. (2009) i d'acord amb Kivisild et al. (2006), troben que el tRNA de la treonina és un dels que conté major nombre de polimorfismes. A més de l'heteroplàsmia puntal 15908, s'han detectat 3 posicions heteroplàsmiques més situades en altres tRNAs, totes ells localitzades en les braços dels tRNAs. Aquests resultats difereixen dels detectats a nivell poblacional, on es troba una raó, pel que fa a la localització en el tRNA, braç-bucle de 0.99 (Pereira et al. 2009), alhora que contrasta amb la funció principal dels braços dels tRNAs de mantenir l'estructura secundària.

S'han detectat un total d'onze heteroplàsmies puntuals localitzades en posicions estables de gens que codifiquen per proteïnes (3532, 6054, 7697, 7754, 8603, 9029, 11253, 13604,

14561, 14770 i 15046). Aquestes posicions tenen entre dos i zero *hits* a la filogènia, presenten freqüències nul·les o quasi nul·les a la població pel que fa a la variant minoritària, així com uns índex de conservació molt elevats. A més, d'aquelles posicions heteroplàsmiques localitzades en complexos mitocondrials pels quals ha estat descrita la seva estructura cristal·lina, la posició 7754, que implica una substitució Asp70Asn, té una implicació directa en l'estructura terciària del complex COXII, provocant la pèrdua de les interaccions amb els aminoàcids propers.

Dues d'aquestes posicions heteroplàsmiques estables (7679 i 11253) estan relacionades amb patologies mitocondrials. La posició heteroplàsmica 7697 i el conseqüent canvi aminoacídic (valina a isoleucina) ha estat reportat com un factor genètic indicatiu de successibilitat a cardiomiopatia hipertròfica (HCM) (Wei et al. 2009). Pel que fa a l'heteroplàsmia puntual localitzada a la posició 11253 (canvi aminoacídic isoleucina a treonina), s'ha relacionat directament amb neuropatia òptica hereditària de Leber (LHON) (Fauser et al. 2002; Leo-Kottler et al. 2002).

Recuperant l'anteriorment exposat pel que fa a l'excés de transicions en l'espectre mutacional de l'heteroplàsmia, sembla ser que les forces evolutives podrien estar actuant per disminuir aquests nivells a nivell poblacional. El més probable és que aquestes mutacions siguin finalment eliminades per deriva genètica o per selecció. De fet, les heteroplàsmies puntuals detectades en aquest treball en posicions estables, podrien tenir un gran impacte en la supervivència de la mitocòndria, suggerint que la selecció

purificadora deu estar actuant per prevenir la fixació de mutacions perjudicials en els individus.

Aquest és el primer estudi on s’analitza la freqüència i el patró de l’heteroplàsmia mitocondrial humana en individus sans. Tenint en compte l'elevada quantitat de dades que es presenta, unit a la poca informació prèvia disponible, aquest treball proporciona una nova perspectiva per a l'estudi de les malalties associades al DNAmt, així com en treballs a nivell poblacional, evolutiu i forense.

5. CONCLUSIONS

CONCLUSIONS DE LA TESI

1. S'han dissenyat, validat i descrit per primera vegada un conjunt de *primers* que permeten amplificar i seqüenciar selectivament la totalitat del genoma mitocondrial. Aquesta metodologia pot ser utilitzada com a referència per al correcte disseny de *primers*.

2. La publicació del nou genoma humà de referència GRCh37 ha provocat el redisseny i validació d'una de les parelles de *primers* per poder seguir garantint l'amplificació selectiva de DNAmt; posant de manifest la necessitat de realitzar una actualització tant de les bases de dades de NUMTs com dels *primers* dissenyats per amplificar selectivament DNAmt.

3. S'ha generat, per primera vegada, una base de dades d'insercions nuclears d'origen mitocondrial basada en el nou esborrany del genoma humà GRCh37. Tenint en compte la quantitat d'informació aportada per a cada

inserció nuclear, la base de dades pot ser usada per al disseny d'estudis evolutius, filogenètics i epidemiològics.

4. És el primer treball en que es reporten les posicions no-idèntiques entre el DNAmt i els NUMTs. El coneixement de la distribució d'aquestes posicions és una eina útil per autenticar mutacions del DNAmt presentades en diferents tipus d'estudis, sobretot aquelles detectades en heteroplàsmia.

5. Utilitzant els estudis de càncer com a exemple, s'ha evidenciat l'aplicabilitat de la base de dades de NUMTs com una nova eina per validar *in silico* les mutacions de DNAmt descrites en diferents contextos. També es posa de manifest la necessitat d'estandarditzar protocols que evitin la co-amplificació entre DNAmt i DNAn, com per exemple el disseny de *primers* que es presenta en aquesta Tesi.

6. L'estudi de l'heteroplàsmia mitocondrial realitzat revela que un 62.37% de la població presenta heteroplàsmia mitocondrial (de longitud i/o puntual) i un 24.75% d'individus presenta heteroplàsmia puntual, posant de manifest l'elevada freqüència d'heteroplàsmia mitocondrial a nivell poblacional.

7. La comparació dels resultats obtinguts amb els descrits per Li et. al (2010) apunta que les dues metodologies utilitzades en la seqüenciació completa del DNAmt,

presenten una sensibilitat similar. A més, el present estudi demostra la possibilitat de detectar amb seguretat les heteroplàsmies puntuals amb un nivell d'un 5%.

8. La gran majoria de les heteroplàsmies puntuals tenen un origen germinal i només una petita fracció té un origen somàtic. Les heteroplàsmies d'origen somàtic semblen tenir una certa relació amb l'envelliment i/o la longevitat, ja que les presenten els individus d'edat avançada.

9. L'elevada freqüència d'heteroplàsmies d'origen germinal implica que els individus porten un *background* mutacional que augmentaria el risc o protegiria enfront a patologies relacionades amb disfunció mitocondrial.

10. La distribució de l'heteroplàsmia mitocondrial comprèn tot el genoma mitocondrial, localitzant-se tant en posicions definides com a *hotspot* com en posicions altament estables i conservades. Comparant el patró mutacional de l'heteroplàsmia amb el descrit a nivell poblacional, sembla haver-hi un excés d'heteroplàsmia puntual localitzada en posicions que presenten zero *hits* en la filogènia.

11. L'excés de transicions detectat en l'espectre mutacional de l'heteroplàsmia, en comparació amb el descrit a nivell poblacional, implica que determinades forces evolutives, com la deriva o la selecció, han hagut d'actuar perquè finalment siguin eliminades. En aquest context, les

heteroplàsmies puntuals localitzades en posicions estables podrien tenir un gran impacte en la supervivència de la pròpia mitocòndria, suggerint que la selecció purificadora hauria d'estar actuant per prevenir la seva fixació en els individus.

12. Aquest és el primer treball on s'analitza la freqüència i el patró de l'heteroplàsmia mitocondrial humana en individus sans. Tenint en compte l'elevada quantitat de dades que es presenta, unit a la poca informació prèvia disponible, aquest treball proporciona una nova perspectiva per a l'estudi de les malalties associades al DNAmt, així com en treballs a nivell poblacional, evolutiu i forense.

6. BIBLIOGRAFIA

Abnet CC, Huppi K, Carrera A, Armistead D, McKenney K, Hu N, Tang ZZ, Taylor PR, Dawsey SM (2004) Control region mutations and the 'common deletion' are frequent in the mitochondrial DNA of patients with esophageal squamous cell carcinoma. BMC Cancer 4:30

Afonso Costa H, Carvalho M, Lopes V, Balsa F, Bento AM, Serra A, Andrade L, Anjos MJ, Vide MC, Pantoja S, Vieira DN, Corte-Real F (2010) Mitochondrial DNA sequence analysis of a native Bolivian population. J Forensic Leg Med 17:247

Aikhionbare FO, Mehrabi S, Kumaresan K, Zavareh M, Olatinwo M, Odunsi K, Partridge E (2007) Mitochondrial DNA sequence variants in epithelial ovarian tumor subtypes and stages. J Carcinog 6:1

Alston CL, Lowe J, Turnbull DM, Maddison P, Taylor RW (2010) A novel mitochondrial tRNAGlu (MTTE) gene mutation causing chronic progressive external ophthalmoplegia at low levels of heteroplasmy in muscle. J Neurol Sci 298:140

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389

Alvarez L, Santos C, Ramos A, Pratdesaba R, Francalacci P, Aluja MP (2010) Mitochondrial DNA patterns in the Iberian Northern plateau: population dynamics and substructure of the Zamora province. Am J Phys Anthropol 142:531

Allegra E, Garozzo A, Lombardo N, De Clemente M, Carey TE (2006) Mutations and polymorphisms in mitochondrial DNA in head and neck cancer cell lines. Acta Otorhinolaryngol Ital 26:185

Ambrosini-Spaltro A, Morandi L, Spagnolo DV, Cavazza A, Brisigotti M, Damiani S, Jain S, Eusebi V (2010) Nasal seromucinous hamartoma (microglandular adenosis of the nose): a morphological and molecular study of five cases. Virchows Arch 457:727

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981)

Sequence and organization of the human mitochondrial genome. Nature 290:457

Anderson S, de Bruijn MH, Coulson AR, Eperon IC, Sanger F, Young IG (1982) Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

Ansorge WJ (2009) Next-generation DNA sequencing techniques. N Biotechnol 25:195

Arctander P (1999) Mitochondrial recombination? Science 284:2090

Avise JC (1991) Ten unorthodox perspectives on evolution prompted by comparative population genetic findings on mitochondrial DNA. Annu Rev Genet 25:45

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, Reeb CA, Saunders NC (1987) Intraspecific phylogeography - the mitochondrial DNA bridge between population genetics and systematics. Ann Rev Ecol Syst 18:489

Awadalla P, Eyre-Walker A, Smith JM (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524

Awadalla P, Eyre-Walker A, Smith JM (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931a

Bai RK, Wong LJ (2004) Detection and quantification of heteroplasmic mutant mitochondrial DNA by real-time amplification refractory mutation system quantitative PCR analysis: a single-step approach. Clin Chem 50:996

Ballana E, Govea N, de Cid R, Garcia C, Arribas C, Rosell J, Estivill X (2008) Detection of unrecognized low-level mtDNA heteroplasmy may explain the variable phenotypic expressivity of apparently homoplasmic mtDNA mutations. Hum Mutat 29:248

Ballard JW (2000) Comparative genomics of mitochondrial DNA in members of the Drosophila melanogaster subgroup. J Mol Evol 51:48

Ballard JW, Kreitman M (1994) Unraveling selection in the mitochondrial genome of Drosophila. Genetics 138:757

Bendall KE, Macaulay VA, Baker JR, Sykes BC (1996) Heteroplasmic point mutations in the human mtDNA control region. Am J Hum Genet 59:1276

Bendall KE, Macaulay VA, Sykes BC (1997) Variable levels of a heteroplasmic point mutation in individual hair roots. Am J Hum Genet 61:1303

Bensasson D, Zhang D, Hartl DL, Hewitt GM (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol 16:314

Betts J, Jaros E, Perry RH, Schaefer AM, Taylor RW, Abdel-All Z, Lightowlers RN, Turnbull DM (2006) Molecular neuropathology of MELAS: level of heteroplasmy in individual neurones and evidence of extensive vascular involvement. Neuropathol Appl Neurobiol 32:359

Bi R, Zhang AM, Yu D, Chen D, Yao YG (2010) Screening the three LHON primary mutations in the general Chinese population by using an optimized multiplex allele-specific PCR. Clin Chim Acta 411:1671

Bidooki S, Jackson MJ, Johnson MA, Chrzanowska-Lightowlers ZM, Taylor RW, Venables G, Lightowlers RN, Turnbull DM, Bindoff LA (2004) Sporadic mitochondrial myopathy due to a new mutation in the mitochondrial tRNASer(UCN) gene. Neuromuscul Disord 14:417

Biggin A, Henke R, Bennetts B, Thorburn DR, Christodoulou J (2005) Mutation screening of the mitochondrial genome using denaturing high-performance liquid chromatography. Mol Genet Metab 84:61

Blouin MS (2000) Neutrality tests on mtDNA: unusual results from nematodes. J Hered 91:156

Brandstatter A, Parson W (2003) Mitochondrial DNA heteroplasmy or artefacts--a matter of the amplification strategy? Int J Legal Med 117:180

Briggs AW, Good JM, Green RE, Krause J, Maricic T, Stenzel U, Lalueza-Fox C, Rudan P, Brajkovic D, Kucan Z, Gusic I, Schmitz R, Doronichev VB, Golovanova LV, de la Rasilla M, Fortea J, Rosas A, Paabo S (2009) Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325:318

Brinckmann A, Ruther K, Williamson K, Lorenz B, Lucke B, Nurnberg P, Trijbels F, Janssen A, Schuelke M (2007) De novo double mutation in PAX6 and mtDNA tRNA(Lys)

associated with atypical aniridia and mitochondrial disease. J Mol Med (Berl) 85:163

Bromham L, Eyre-Walker A, Smith NH, Smith JM (2003) Mitochondrial Steve: paternal inheritance of mitochondria in humans. Trends Ecol Evol 18:2

Brown DT, Samuels DC, Michael EM, Turnbull DM, Chinnery PF (2001) Random genetic drift determines the level of mutant mtDNA in human primary oocytes. Am J Hum Genet 68:533

Brown WM (1981) Mechanisms of evolution in animal mitochondrial DNA. Ann N Y Acad Sci 361:119

Brown WM (1985) The mitochondrial genome of animals. In: R.J. eM (ed) Molecular evolutionary genetics, (Plenum, New York), pp. 95-130

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225

Calloway CD, Reynolds RL, Herrin GL, Jr., Anderson WW (2000) The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age. Am J Hum Genet 66:1384

Cann RL, Brown WM, Wilson AC (1982) Evolution of human mitochondrial DNA: a preliminary report. Prog Clin Biol Res 103 Pt A:157

Cann RL, Brown WM, Wilson AC (1984) Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics 106:479

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31

Canter JA, Eshaghian A, Fessel J, Summar ML, Roberts LJ, Morrow JD, Sligh JE, Haines JL (2005) Degree of heteroplasmy reflects oxidant damage in a large family with the mitochondrial DNA A8344G mutation. Free Radic Biol Med 38:678

Cao Y, Ma Y, Zhang Y, Li Y, Fang F, Wang S, Bu D, Xu Y, Pei P, Li L, Xiao Y, Wu H, Yang Y, Zou L, Qi Y (2010) Detection of eight frequently encountered point mutations in mitochondria in Chinese patients suggestive of mitochondrial encephalomyopathies. Mitochondrion 10:330

Cardaioli E, Da Pozzo P, Malfatti E, Gallus GN, Rubegni A, Malandrini A, Gaudiano C, Guidi L, Serni G, Berti G, Dotti MT, Federico A (2008) Chronic progressive external

ophthalmoplegia: a new heteroplasmic tRNA(Leu(CUN)) mutation of mitochondrial DNA. J Neurol Sci 272:106

Cassandrini D, Calevo MG, Tessa A, Manfredi G, Fattori F, Meschini MC, Carrozzo R, Tonoli E, Pedemonte M, Minetti C, Zara F, Santorelli FM, Bruno C (2006) A new method for analysis of mitochondrial DNA point mutations and assess levels of heteroplasmy. Biochem Biophys Res Commun 342:387

Clark AG, Lyckegaard EM (1988) Natural selection with nuclear and cytoplasmic transmission. III. Joint analysis of segregation and mtDNA in Drosophila melanogaster. Genetics 118:471

Claros MG (1995) MitoProt, a Macintosh application for studying mitochondrial proteins. Comput Appl Biosci 11:441

Claros MG, Perea J, Shu Y, Samatey FA, Popot JL, Jacq C (1995) Limitations to in vivo import of hydrophobic proteins into yeast mitochondria. The case of a cytoplasmically synthesized apocytochrome b. Eur J Biochem 228:762

Clayton DA (1996) Mitochondrial DNA gets the drift. Nat Genet 14:123

Cortopassi GA, Shibata D, Soong NW, Arnheim N (1992) A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues. Proc Natl Acad Sci U S A 89:7370

Coskun PE, Beal MF, Wallace DC (2004) Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. PNAS 101:10726

Chen H, Vermulst M, Wang YE, Chomyn A, Prolla TA, McCaffery JM, Chan DC (2010) Mitochondrial fusion is required for mtDNA stability in skeletal muscle and tolerance of mtDNA mutations. Cell 141:280

Chinnery PF (2002) Inheritance of mitochondrial disorders. Mitochondrion 2:149

Chong MD, Calloway CD, Klein SB, Orrego C, Buoncristiani MR (2005) Optimization of a duplex amplification and sequencing strategy for the HVI/HVII regions of human mitochondrial DNA for forensic casework. Forensic Sci Int 154:137

D'Aurelio M, Vives-Bauza C, Davidson MM, Manfredi G (2010) Mitochondrial DNA background modifies the bioenergetics of NARP/MILS ATP6 mutant cells. Hum Mol Genet 19:374

Da Costa CK, Kiyomoto BH, Schmidt B, Oliveira AS, Gabbai AA, Tengan CH (2007) Age-related mitochondrial DNA point

mutations in patients with mitochondrial myopathy. J Neurol Sci 263:139

Da Cruz S, Parone PA, Martinou JC (2005) Building the mitochondrial proteome. Expert Rev Proteomics 2:541

De Camargo MA, Paneto GG, de Mello AC, Martins JA, Barcellos W, Cicarelli RM (2011) No relationship found between point heteroplasmy in mitochondrial DNA control region and age range, sex and haplogroup in human hairs. Mol Biol Rep 38:1219

Debray FG, Lambert M, Lortie A, Vanasse M, Mitchell GA (2007) Long-term outcome of Leigh syndrome caused by the NARP-T8993C mtDNA mutation. Am J Med Genet A 143A:2046

DiMauro S, Schon EA (2008) Mitochondrial disorders in the nervous system. Annu Rev Neurosci 31:91

Dobrowolski SF, Hendrickx AT, Van den Bosch BJ, Smeets HJ, Gray J, Miller T, Sears M (2009) Identifying sequence variants in the human mitochondrial genome using high-resolution melt (HRM) profiling. Hum Mutat 30:891

Elson JL, Andrews RM, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (2001) Analysis of European mtDNAs for recombination. Am J Hum Genet 68:145

Elson JL, Turnbull DM, Howell N (2004) Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. Am J Hum Genet 74:229

Excoffier L (1990) Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. J Mol Evol 30:125

Excoffier L, Yang Z (1999) Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. Mol Biol Evol 16:1357

Eyre-Walker A, Smith NH, Smith JM (1999) How clonal are human mitochondria? Proc Biol Sci 266:477

Fauser S, Luberichs J, Besch D, Leo-Kottler B (2002) Sequence analysis of the complete mitochondrial genome in patients with Leber's hereditary optic neuropathy lacking the three most common pathogenic DNA mutations. Biochem Biophys Res Commun 295:342

Fendt L, Niederstatter H, Huber G, Zelger B, Dunser M, Seifarth C, Rock A, Schafer G, Klocker H, Parson W (2011) Accumulation of mutations over the entire mitochondrial

genome of breast cancer cells obtained by tissue microdissection. Breast Cancer Res Treat 128:327

Filosto M, Mancuso M, Vives-Bauza C, Vila MR, Shanske S, Hirano M, Andreu AL, DiMauro S (2003) Lack of paternal inheritance of muscle mitochondrial DNA in sporadic mitochondrial myopathies. Ann Neurol 54:524

Fliss MS, Usadel H, Caballero OL, Wu L, Buta MR, Eleff SM, Jen J, Sidransky D (2000) Facile detection of mitochondrial DNA mutations in tumors and bodily fluids. Science 287:2017

Forner F, Foster LJ, Campanaro S, Valle G, Mann M (2006) Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. Mol Cell Proteomics 5:608

Fos M, Dominguez MA, Latorre A, Moya A (1990) Mitochondrial DNA evolution in experimental populations of Drosophila subobscura. Proc Natl Acad Sci U S A 87:4198

Frederiksen AL, Jeppesen TD, Vissing J, Schwartz M, Kyvik KO, Schmitz O, Poulsen PL, Andersen PH (2009) High prevalence of impaired glucose homeostasis and myopathy in asymptomatic and oligosymptomatic 3243A>G mitochondrial DNA mutation-positive subjects. J Clin Endocrinol Metab 94:2872

Gabaldon T, Huynen MA (2004) Shaping the mitochondrial proteome. Biochim Biophys Acta 1659:212

Gambello MJ, Bai RK, Chen TJ, Dimachkie M, Wong LJ (2006) Exercise intolerance associated with a novel 8300T > C mutation in mitochondrial transfer RNAlys. Muscle Nerve 34:437

Garcia-Martinez J, Castro JA, Ramon M, Latorre A, Moya A (1998) Mitochondrial DNA haplotype frequencies in natural and experimental populations of Drosophila subobscura. Genetics 149:1377

Gasparre G, Porcelli AM, Bonora E, Pennisi LF, Toller M, Iommarini L, Ghelli A, Moretti M, Betts CM, Martinelli GN, Ceroni AR, Curcio F, Carelli V, Rugolo M, Tallini G, Romeo G (2007) Disruptive mitochondrial DNA mutations in complex I subunits are markers of oncocytic phenotype in thyroid tumors. Proc Natl Acad Sci U S A 104:9001

Genasetti A, Valentino ML, Carelli V, Vigetti D, Viola M, Karousou EG, Melzi d'Eril GV, De Luca G, Passi A, Pallotti F (2007) Assessing heteroplasmic load in Leber's hereditary optic neuropathy mutation 3460G->A/MT-ND1 with a real-time PCR quantitative approach. J Mol Diagn 9:538

Gigarel N, Ray PF, Burlet P, Frydman N, Royer G, Lebon S, Bonnefont JP, Frydman R, Munnich A, Steffann J (2005) Single cell quantification of the 8993T>G NARP mitochondrial DNA mutation by fluorescent PCR. Mol Genet Metab 84:289

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of human mitochondrial DNA. Proc Natl Acad Sci U S A 77:6715

Gilkerson RW, Schon EA, Hernandez E, Davidson MM (2008) Mitochondrial nucleoids maintain genetic autonomy but allow for functional complementation. J Cell Biol 181:1117

Goios A, Prieto L, Amorim A, Pereira L (2008) Specificity of mtDNA-directed PCR-influence of NUclear MTDNA insertion (NUMT) contamination in routine samples and techniques. Int J Legal Med 122:341

Goto H, Dickins B, Afgan E, Paul IM, Taylor J, Makova KD, Nekrutenko A (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biol 12:R59

Gray MW (1992) The endosymbiont hypothesis revisited. Int Rev Cytol 141:233

Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. Science 283:1476

Gray MW, Doolittle WF (1982) Has the endosymbiont hypothesis been proven? Microbiol Rev 46:1

Grzybowski T (2000) Extremely high levels of human mitochondrial DNA heteroplasmy in single hair roots. Electrophoresis 21:548

Grzybowski T, Malyarchuk BA, Czarny J, Miscicka-Sliwka D, Kotzbach R (2003) High levels of mitochondrial DNA heteroplasmy in single hair roots: reanalysis and revision. Electrophoresis 24:1159

Gyllensten U, Wharton D, Josefsson A, Wilson AC (1991) Paternal inheritance of mitochondrial DNA in mice. Nature 352:255

Habano W, Sugai T, Yoshida T, Nakamura S (1999) Mitochondrial gene mutation, but not large-scale deletion, is a feature of colorectal carcinomas with mitochondrial microsatellite instability. Int J Cancer 83:625

Hagelberg E (2003) Recombination or mutation rate heterogeneity? Implications for Mitochondrial Eve. Trends Genet 19:84

Hagelberg E, Goldman N, Lio P, Whelan S, Schiefenhovel W, Clegg JB, Bowden DK (1999) Evidence for mitochondrial DNA

recombination in a human population of island Melanesia. Proc Biol Sci 266:485

Hagelberg E, Goldman N, Lio P, Whelan S, Schiefenhovel W, Clegg JB, Bowden DK (2000) Evidence for mitochondrial DNA recombination in a human population of island Melanesia: correction. Proc Biological Sci 267:1595

Harrison RG (1989) Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. Trends Ecol Evol 4:6

Hazkani-Covo E (2009) Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. Mol Biol Evol 26:2175

Hazkani-Covo E, Covo S (2008) Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet 4:e1000237

Hazkani-Covo E, Graur D (2007) A Comparative Analysis of numt Evolution in Human and Chimpanzee. Mol. Biol. Evol. 24:13

Hazkani-Covo E, Sorek R, Graur D (2003) Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J Mol Evol 56:169

Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 6:e1000834

He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE, Diaz LA, Jr., Kinzler KW, Vogelstein B, Papadopoulos N (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 464:610

Hervouet E, Godinot C (2006) Mitochondrial disorders in renal tumors. Mitochondrion 6:105

Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. Am J Hum Genet 69:1113

Hirano M, Shtilbans A, Mayeux R, Davidson MM, DiMauro S, Knowles JA, Schon EA (1997) Apparent mtDNA heteroplasmy in Alzheimer's disease patients and in normals due to PCR amplification of nucleus-embedded mtDNA pseudogenes. Proc Natl Acad Sci U S A 94:14894

Hogeboon GH, Schneider WC, Palade GE (1948) Cytochemical studies of mammalian tissues; isolation of intact mitochondria from rat liver; some biochemical properties of mitochondria and submicroscopic particulate material. J Biol Chem 172:619

Holm L, Sander C (1992) Evaluation of protein models by atomic solvation preference. J Mol Biol 225:93

Holt IJ, Harding AE, Morgan-Hughes JA (1988) Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. Nature 331:717

Howell N, Ghosh SS, Fahy E, Bindoff LA (2000) Longitudinal analysis of the segregation of mtDNA mutations in heteroplasmic individuals. J Neurol Sci 172:1

Howell N, Halvorson S, Kubacka I, McCullough DA, Bindoff LA, Turnbull DM (1992) Mitochondrial gene segregation in mammals: is the bottleneck always narrow? Hum Genet 90:117

Howell N, Mackey D (1997) mtDNA mutation rates - No need to panic - Reply. Am J Hum Genet 61:986

Howell N, Smejkal CB (2000) Persistent heteroplasmy of a mutation in the human mtDNA control region: hypermutation as an apparent consequence of simple-repeat expansion/contraction. Am J Hum Genet 66:1589

Howell N, Smejkal CB, Mackey DA, Chinnery PF, Turnbull DM, Herrnstadt C (2003) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. Am J Hum Genet 72:659

Hutter CM, Rand DM (1995) Competition between mitochondrial haplotypes in distinct nuclear genetic environments: Drosophila pseudoobscura vs. D. persimilis. Genetics 140:537

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

Irwin JA, Saunier JL, Niederstatter H, Strouss KM, Sturk KA, Diegoli TM, Brandstatter A, Parson W, Parsons TJ (2009) Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. J Mol Evol 68:516

Iwata S, Lee JW, Okada K, Lee JK, Iwata M, Rasmussen B, Link TA, Ramaswamy S, Jap BK (1998) Complete structure of

the 11-subunit bovine mitochondrial cytochrome bc1 complex. Science 281:64

Jacobs HT, Hutchin TP, Kappi T, Gillies G, Minkkinen K, Walker J, Thompson K, Rovio AT, Carella M, Melchionda S, Zelante L, Gasparini P, Pyykko I, Shah ZH, Zeviani M, Mueller RF (2005) Mitochondrial DNA mutations in patients with postlingual, nonsyndromic hearing impairment. Eur J Hum Genet 13:26

Jacobs L, Gerards M, Chinnery P, Dumoulin J, de Coo I, Geraedts J, Smeets H (2007) mtDNA point mutations are present at various levels of heteroplasmy in human oocytes. Mol Hum Reprod 13:149

Jakupciak JP, Maragh S, Markowitz ME, Greenberg AK, Hoque MO, Maitra A, Barker PE, Wagner PD, Rom WN, Srivastava S, Sidransky D, O'Connell CD (2008) Performance of mitochondrial DNA mutations detecting early stage cancer. BMC Cancer 8:285

Jansen RP, de Boer K (1998) The bottleneck: mitochondrial imperatives in oogenesis and ovarian follicular fate. Mol Cell Endocrinol 145:81

Jazin E, Soodyall H, Jalonen P, Lindholm E, Stoneking M, Gyllensten U (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism. Nat Genet 18:109

Jazin EE, Cavelier L, Eriksson I, Oreland L, Gyllensten U (1996) Human brain contains high levels of heteroplasmy in the noncoding regions of mitochondrial DNA. Proc Natl Acad Sci U S A 93:12382

Jensen-Seaman MI, Wildschutte JH, Soto-Calderón ID, Anthony NM (2009) A Comparative Approach Shows Differences in Patterns of Numt Insertion During Hominoid Evolution. J Mol Evol 68

Jin X, Zhang J, Gao Y, Ding K, Wang N, Zhou D, Jen J, Cheng S (2007) Relationship between mitochondrial DNA mutations and clinical characteristics in human lung cancer. Mitochondrion 7:347

Johnson DT, Harris RA, Blair PV, Balaban RS (2007) Functional consequences of mitochondrial proteome heterogeneity. Am J Physiol Cell Physiol 292:C698

Jones JB, Song JJ, Hempen PM, Parmigiani G, Hruban RH, Kern SE (2001) Detection of mitochondrial DNA mutations in pancreatic cancer offers a "mass"-ive advantage over detection of nuclear DNA mutations. Cancer Res 61:1299

Jorde LB, Bamshad M (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931

Kaare M, Gotz A, Ulander VM, Ariansen S, Kaaja R, Suomalainen A, Aittomaki K (2009) Do mitochondrial mutations cause recurrent miscarriage? Mol Hum Reprod 15:295

Kambhampati S, Rai KS, Verleye DM (1992) Frequencies of mitochondrial DNA haplotypes in laboratory cage populations of the mosquito, Aedes albopictus. Genetics 132:205

Kaneda H, Hayashi J, Takahama S, Taya C, Lindahl KF, Yonekawa H (1995) Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis. Proc Natl Acad Sci U S A 92:4542

Kaplanova V, Zeman J, Hansikova H, Cerna L, Houst'kova H, Misovicova N, Houstek J (2004) Segregation pattern and biochemical effect of the G3460A mtDNA mutation in 27 members of LHON family. J Neurol Sci 223:149

Karppa M, Mahjneh I, Karttunen A, Tolonen U, Majamaa K (2004) Muscle computed tomography patterns in patients with the mitochondrial DNA mutation 3243A>G. J Neurol 251:556

Kassauei K, Habbe N, Mullendore ME, Karikari CA, Maitra A, Feldmann G (2006) Mitochondrial DNA mutations in pancreatic cancer. Int J Gastrointest Cancer 37:57

Kilpatrick ST, Rand DM (1995) Conditional hitchhiking of mitochondrial DNA: frequency shifts of Drosophila melanogaster mtDNA variants depend on nuclear genetic background. Genetics 141:1113

Kirches E, Michael M, Warich-Kirches M, Schneider T, Weis S, Krause G, Mawrin C, Dietzmann K (2001) Heterogeneous tissue distribution of a mitochondrial DNA polymorphism in heteroplasmic subjects without mitochondrial disorders. J Med Genet 38:312

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

Kleine T, Maier UG, Leister D (2009) DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu Rev Plant Biol 60:115

Kloss-Brandstatter A, Schafer G, Erhart G, Huttenhofer A, Coassin S, Seifarth C, Summerer M, Bektic J, Klocker H, Kronenberg F (2010) Somatic mutations throughout the entire mitochondrial genome are associated with elevated PSA levels in prostate cancer patients. Am J Hum Genet 87:802

Kondo R, Satta Y, Matsuura ET, Ishiwa H, Takahata N, Chigusa SI (1990) Incomplete maternal transmission of mitochondrial DNA in Drosophila. Genetics 126:657

Krakauer DC, Mira A (1999) Mitochondria and germ-cell death. Nature 400:125

Kraytsberg Y, Schwartz M, Brown TA, Ebralidse K, Kunz WS, Clayton DA, Vissing J, Khrapko K (2004) Recombination of human mitochondrial DNA. Science 304:981

Kumar S, Hedrick P, Dowling T, Stoneking M (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931

Kvist L, Martens J, Nazarenko AA, Orell M (2003) Paternal leakage of mitochondrial DNA in the great tit (Parus major). Mol Biol Evol 20:243

Laloi-Michelin M, Meas T, Ambonville C, Bellanne-Chantelot C, Beaufils S, Massin P, Vialettes B, Gin H, Timsit J, Bauduceau B, Bernard L, Bertin E, Blickle JF, Cahen-Varsaux J, Cailleba A, Casanova S, Cathebras P, Charpentier G, Chedin P, Crea T, Delemer B, Dubois-Laforgue D, Duchemin F, Ducluzeau PH, Bouhanick B, Dusselier L, Gabreau T, Grimaldi A, Guerci B, Jacquin V, Kaloustian E, Larger E, Lecleire-Collet A, Lorenzini F, Louis J, Mausset J, Murat A, Nadler-Fluteau S, Olivier F, Paquis-Flucklinger V, Paris-Bockel D, Raynaud I, Reznik Y, Riveline JP, Schneebeli S, Sonnet E, Sola-Gazagnes A, Thomas JL, Trabulsi B, Virally M, Guillausseau PJ (2009) The clinical variability of maternally inherited diabetes and deafness is associated with the degree of heteroplasmy in blood leukocytes. J Clin Endocrinol Metab 94:3025

Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C, Attimonelli M (2008) The RHNumtS compilation: features and bioinformatics approaches to locate and quantify Human NumtS. BMC Genomics 9:267

Lehninger (1965) The Mitochondrion — Molecular Basis of Structure and Function. W. A. Benjamin , New York

Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet 21:655

Leo-Kottler B, Luberichs J, Besch D, Christ-Adler M, Fauser S (2002) Leber's hereditary optic neuropathy: clinical and molecular genetic results in a patient with a point mutation at np T11253C (isoleucine to threonine) in the ND4 gene and spontaneous recovery. Graefes Arch Clin Exp Ophthalmol 240:758

Leshinsky-Silver E, Lev D, Tzofi-Berman Z, Cohen S, Saada A, Yanoov-Sharav M, Gilad E, Lerman-Sagie T (2005) Fulminant neurological deterioration in a neonate with Leigh syndrome due to a maternally transmitted missense mutation in the mitochondrial ND3 gene. Biochem Biophys Res Commun 334:582

Li M, Schonberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet 87:237

Li R, Ishikawa K, Deng JH, Heman-Ackah S, Tamagawa Y, Yang L, Bai Y, Ichimura K, Guan MX (2005) Maternally inherited nonsyndromic hearing loss is associated with the T7511C mutation in the mitochondrial tRNASerUCN gene in a Japanese family. Biochem Biophys Res Commun 328:32

Lièvre A, Chapusot C, Bouvier AM, Zinzindohoue F, Piard F, Roignot P, Arnould L, Beaune P, Faivre J, Laurent-Puig P (2005) Clinical value of mitochondrial mutations in colorectal cancer. J Clin Oncol 23:3517

Lightowlers RN, Chinnery PF, Turnbull DM, Howell N (1997) Mammalian mitochondrial genetics: heredity, heteroplasmy and disease. Trends Genet 13:450

Lindberg C, Moslemi AR, Oldfors A (2008) MELAS syndrome in a patient with a point mutation in MTTS1. Acta Neurol Scand 117:128

Liu VW, Shi HH, Cheung AN, Chiu PM, Leung TW, Nagley P, Wong LC, Ngan HY (2001) High incidence of somatic mitochondrial DNA mutations in human ovarian carcinomas. Cancer Res 61:5998

Lo MC, Lee HM, Lin MW, Tzen CY (2005) Analysis of heteroplasmy in hypervariable region II of mitochondrial DNA in maternally related individuals. Ann N Y Acad Sci 1042:130

Lorenc A, Bryk J, Golik P, Kupryjanczyk J, Ostrowski J, Pronicki M, Semczuk A, Szolkowska M, Bartnik E (2003) Homoplasmic MELAS A3243G mtDNA mutation in a colon cancer sample. Mitochondrion 3:119

Lu SY, Nishio S, Tsukada K, Oguchi T, Kobayashi K, Abe S, Usami S (2009) Factors that affect hearing level in individuals with the mitochondrial 1555A.G mutation. Clin Genet 75:480

Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356:83

Lutz-Bonengel S, Sanger T, Parson W, Muller H, Ellwart JW, Follo M, Bonengel B, Niederstatter H, Heinrich M, Schmidt U (2008) Single lymphocytes from two healthy individuals with mitochondrial point heteroplasmy are mainly homoplasmic. Int J Legal Med 122:189

Macaulay V, Richards M, Sykes B (1999) Mitochondrial DNA recombination-no need to panic. Proc Biol Sci 266:2037

Macaulay VA, Richards MB, Forster P, Bendall KE, Watson E, Sykes B, Bandelt HJ (1997) mtDNA mutation rates--no need to panic. Am J Hum Genet 61:983

MacRae AF, Anderson WW (1988) Evidence for non-neutrality of mitochondrial DNA haplotypes in Drosophila pseudoobscura. Genetics 120:485

Manfredi G, Thyagarajan D, Papadopoulou LC, Pallotti F, Schon EA (1997) The fate of human sperm-derived mtDNA in somatic cells. Am J Hum Genet 61:953

Marchington D, Malik S, Banerjee A, Turner K, Samuels D, Macaulay V, Oakeshott P, Fratter C, Kennedy S, Poulton J (2011) Information for genetic management of mtDNA disease: sampling pathogenic mtDNA mutants in the human germline and in placenta. J Med Genet 47:257

Marchington DR, Macaulay V, Hartshorne GM, Barlow D, Poulton J (1998) Evidence from human oocytes for a genetic bottleneck in an mtDNA disease. Am J Hum Genet 63:769

Margulis L (1981) Symbiosis in Cell Evolution: Life and its Environment on the Early Earth. San Francisco: W. H. Freeman and Co.

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. PNAS 74:560

Meierhofer D, Mayr JA, Ebner S, Sperl W, Kofler B (2005) Rapid screening of the entire mitochondrial DNA for low-level heteroplasmic mutations. Mitochondrion 5:282

Meierhofer D, Mayr JA, Fink K, Schmeller N, Kofler B, Sperl W (2006) Mitochondrial DNA mutations in renal cell carcinomas revealed no general impact on energy metabolism. Br J Cancer 94:268

Melo F, Feytmans E (1998) Assessing protein structures with a non-local atomic interaction energy. J Mol Biol 277:1141

Merriweather DA, Kaestle FA (1999) Mitochondrial recombination? (continued). Science 285:837

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31

Michikawa Y, Mazzucchelli F, Bresolin N, Scarlato G, Attardi G (1999) Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. Science 286:774

Mishmar D, Ruiz-Pesini E, Brandon M, Wallace DC (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Hum Mutat 23:125

Mithani SK, Taube JM, Zhou S, Smith IM, Koch WM, Westra WH, Califano JA (2007) Mitochondrial mutations are a late event in the progression of head and neck squamous cell cancer. Clin Cancer Res 13:4331

Monnot S, Gigarel N, Samuels DC, Burlet P, Hesters L, Frydman N, Frydman R, Kerbrat V, Funalot B, Martinovic J, Benachi A, Feingold J, Munnich A, Bonnefont JP, Steffann J (2011) Segregation of mtDNA throughout human embryofetal development: m.3243A>G as a model system. Hum Mutat 32:116

Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial-DNA- relevance for population biology and systematics. Ann Rev Ecol Sys 18:269

Mourier T, Hansen AJ, Willerslev E, Arctander P (2001) The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol Biol Evol 18:1833

Nachman MW (1998) Deleterious mutations in animal mitochondrial DNA. Genetica 102-103:61

Nachman MW, Boyer SN, Aquadro CF (1994) Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice. Proc Natl Acad Sci U S A 91:6364

Nachman MW, Brown WM, Stoneking M, Aquadro CF (1996) Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics 142:953

Nagy A, Wilhelm M, Kovacs G (2003) Mutations of mtDNA in renal cell tumours arising in end-stage renal disease. J Pathol 199:237

Nakada K, Inoue K, Hayashi J (2001) Interaction theory of mammalian mitochondria. Biochem Biophys Res Commun 288:743

Nakada K, Sato A, Hayashi J (2009) Mitochondrial functional complementation in mitochondrial DNA-based diseases. Int J Biochem Cell Biol 41:1907

Nakahara H, Sekiguchi K, Imaizumi K, Mizuno N, Kasai K (2008) Heteroplasmies detected in an amplified mitochondrial DNA control region from a small amount of template. J Forensic Sci 53:306

Nigro L, Prout T (1990) Is there selection on RFLP differences in mitochondrial DNA? Genetics 125:551

Paabo S (1996) Mutational hot spots in the mitochondrial microcosm. Am J Hum Genet 59:493

Pakendorf B, Stoneking M (2005) Mitochondrial DNA and human evolution. Annu Rev Genomics Hum Genet 6:165

Palmer JD (1997) Organelle genomes: going, going, gone! Science 275:790

Paneto GG, Longo LV, Martins JA, de Camargo MA, Costa JC, de Mello AC, Chen B, Oliveira RN, Hirata MH, Cicarelli RM (2010) Heteroplasmy in hair: study of mitochondrial DNA third hypervariable region in hair and blood samples. J Forensic Sci 55:715

Parfait B, Rustin P, Munnich A, Rotig A (1998) Co-amplification of nuclear pseudogenes and assessment of heteroplasmy of mitochondrial DNA mutations. Biochem Biophys Res Commun 247:57

Parr RL, Dakubo GD, Crandall KA, Maki J, Reguly B, Aguirre A, Wittock R, Robinson K, Alexander JS, Birch-Machin MA, Abdel-Malak M, Froberg MK, Diamandis EP, Thayer RE (2006a) Somatic mitochondrial DNA mutations in prostate cancer and normal appearing adjacent glands in comparison to age-matched prostate samples without malignant histology. J Mol Diagn 8:312

Parr RL, Maki J, Reguly B, Dakubo GD, Aguirre A, Wittock R, Robinson K, Jakupciak JP, Thayer RE (2006b) The pseudo-

mitochondrial genome influences mistakes in heteroplasmy interpretation. BMC Genomics 7:185

Parrella P, Xiao Y, Fliss M, Sanchez-Cespedes M, Mazzarelli P, Rinaldi M, Nicol T, Gabrielson E, Cuomo C, Cohen D, Pandit S, Spencer M, Rabitti C, Fazio VM, Sidransky D (2001) Detection of mitochondrial DNA mutations in primary breast cancer and fine-needle aspirates. Cancer Res 61:7623

Parsons TJ, Irwin JA (2000) Questioning evidence for recombination in human mitochondrial DNA. Science 288:1931

Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, Berry DL, Holland KA, Weedn VW, Gill P, Holland MM (1997) A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet 15:363

Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, Maximo V, Macaulay V, Rocha R, Samuels DC (2009) The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet 84:628

Petros JA, Baumann AK, Ruiz-Pesini E, Amin MB, Sun CQ, Hall J, Lim S, Issa MM, Flanders WD, Hosseini SH, Marshall FF, Wallace DC (2005) mtDNA mutations increase tumorigenicity in prostate cancer. Proc Natl Acad Sci U S A 102:719

Phasukkijwatana N, Chuenkongkaew WL, Suphavilai R, Luangtrakool K, Kunhapan B, Lertrit P (2006) Transmission of heteroplasmic G11778A in extensive pedigrees of Thai Leber hereditary optic neuropathy. J Hum Genet 51:1110

Picornell A, Gimenez P, Castro JA, Ramon MM (2006) Mitochondrial DNA sequence variation in Jewish populations. Int J Legal Med 120:271

Piganeau G, Eyre-Walker A (2004) A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. Heredity (Edinb) 92:282

Poe BG, Navratil M, Arriaga EA (2007) Absolute quantitation of a heteroplasmic mitochondrial DNA deletion using a multiplex three-primer real-time PCR assay. Anal Biochem 362:193

Polyak K, Li Y, Zhu H, Lengauer C, Willson JK, Markowitz SD, Trush MA, Kinzler KW, Vogelstein B (1998) Somatic mutations of the mitochondrial genome in human colorectal tumours. Nat Genet 20:291

Popot JL, de Vitry C (1990) On the microassembly of integral membrane proteins. Annu Rev Biophys Biophys Chem 19:369

Poulton J (1995) Transmission of mtDNA: cracks in the bottleneck. Am J Hum Genet 57:224

Poulton J, Marchington DR (2002) Segregation of mitochondrial DNA (mtDNA) in human oocytes and in animal models of mtDNA disease: clinical implications. Reproduction 123:751

Prokisch H, Andreoli C, Ahting U, Heiss K, Ruepp A, Scharfe C, Meitinger T (2006) MitoP2: the mitochondrial proteome database--now including mouse data. Nucleic Acids Res 34:D705

Putz J, Dupuis B, Sissler M, Florentz C (2007) Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures. RNA 13:1184

Quesada H, Wenne R, Skibinski DO (1999) Interspecies transfer of female mitochondrial DNA is coupled with role-reversals and departure from neutrality in the mussel Mytilus trossulus. Mol Biol Evol 16:655

Ramos A, Barbena E, Mateiu L, Gonzalez MM, Mairal Q, Lima M, Montiel R, Aluja MP, Santos C (2011a) Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies. Mitochondrion 11:946

Ramos A, Santos C, Alvarez L, Nogués R, Aluja MP (2009) Human Mitochondrial DNA complete amplification and sequencing: A new validated primer set that prevents nuclear DNA sequences of mitochondial origin co-amplification. Electrophoresis 30:1

Ramos A, Santos C, Barbena E, Mateiu L, Alvarez L, Nogues R, Aluja MP (2011b) Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: a revision based on the New Human Genome Reference Sequence (GRCh37). Electrophoresis 32:782

Rand DM (1994) Thermal habit, metabolic rate and the evolution of mitochondrial DNA. Trends Ecol Evol 9:125

Rand DM (2001) The units of selection on mitochondrial dna. Annu. Rev. Ecol. Syst 32:415

Rand DM, Dorfsman M, Kann LM (1994) Neutral and non-neutral evolution of Drosophila mitochondrial DNA. Genetics 138:741

Rand DM, Kann LM (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol Biol Evol 13:735

Rand DM, Kann LM (1998) Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. Genetica 102-3:393

Rand DM, Weinreich DM, Cezairliyan BO (2000) Neutrality tests of conservative-radical amino acid changes in nuclear- and mitochondrially-encoded proteins. Gene 261:115

Raymond M, Rousset F (2001) GENEPOP 3.3: population genetics software for exact tests and ecumenicism. , Institut des Sciences de l'Evolution. Université de Montpellier II, Montpellier, France

Ricchetti M, Tekaia F, Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. PLoS Biol 2:E273

Richly E, Leister D (2004) NUMTs in sequenced eukaryotic genomes. Mol Biol Evol 21:1081

Robin ED, Wong R (1988) Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol 136:507

Roche (2006) PCR Applications Manual. Roche Diagnostics GmbH, Mannheim

Rose G, Passarino G, Scornaienchi V, Romeo G, Dato S, Bellizzi D, Mari V, Feraco E, Maletta R, Bruni A, Franceschi C, De Benedictis G (2007) The mitochondrial DNA control region shows genetically correlated levels of heteroplasmy in leukocytes of centenarians and their offspring. BMC Genomics 8:293

Rose G, Romeo G, Dato S, Crocco P, Bruni AC, Hervonen A, Majamaa K, Sevini F, Franceschi C, Passarino G (2010) Somatic point mutations in mtDNA control region are influenced by genetic background and associated with healthy aging: a GEHA study. PLoS One 5:e13395

Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V, Wallace DC (2004) Effects of purifying and adaptive selection on regional variation in human mtDNA. Science 303:223

Ruiz-Pesini E, Wallace DC (2006) Evidence for adaptive selection acting on the tRNA and rRNA genes of human mitochondrial DNA. Hum Mutat 27:1072

Salas A, Carracedo A, Macaulay V, Richards M, Bandelt HJ (2005) A practical guide to mitochondrial DNA error prevention in

clinical, forensic, and population genetics. Biochem Biophys Res Commun 335:891

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779

Salvioli S, Capri M, Santoro A, Raule N, Sevini F, Lukas S, Lanzarini C, Monti D, Passarino G, Rose G, De Benedictis G, Franceschi C (2008) The impact of mitochondrial DNA on human lifespan: a view from studies on centenarians. Biotechnol J 3:740

Sambrook J (2001) Molecular cloning : a laboratory manual Cold Spring Harbor : Cold Spring Harbor Laboratory

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441

Santos C, Abade A, Cantons J, Mayer FM, Aluja MP, Lima M (2005a) Genetic structure of Flores island (Azores, Portugal) in the 19th century and in the present day: evidence from surname analysis. Hum Biol 77:317

Santos C, Martinez M, Lima M, Hao YJ, Simoes N, Montiel R (2008a) Mitochondrial DNA mutations in cancer: a review. Curr Top Med Chem 8:1351

Santos C, Montiel R, Arruda A, Alvarez L, Aluja MP, Lima M (2008b) Mutation patterns of mtDNA: empirical inferences for the coding region. BMC Evol Biol 8:167

Santos C, Montiel R, Sierra B, Bettencourt C, Fernandez E, Alvarez L, Lima M, Abade A, Aluja MP (2005b) Understanding differences between phylogenetic and pedigree-derived mtDNA mutation rate: a model using families from the Azores Islands (Portugal). Mol Biol Evol 22:1490

Santos C, Sierra B, Alvarez L, Ramos A, Fernandez E, Nogues R, Aluja MP (2008c) Frequency and pattern of heteroplasmy in the control region of human mitochondrial DNA. J Mol Evol 67:191

Sarzi E, Brown MD, Lebon S, Chretien D, Munnich A, Rotig A, Procaccio V (2007) A novel recurrent mitochondrial DNA mutation in ND3 gene is associated with isolated complex I deficiency causing Leigh syndrome and dystonia. Am J Med Genet A 143:33

Schatz G (1995) Mitochondria: beyond oxidative phosphorylation. Biochim Biophys Acta 1271:123

Scheffler I (2008) Mitochondria. John Wiley & Sons

Schriner SE, Ogburn CE, Smith AC, Newcomb TG, Ladiges WC, Dolle ME, Vijg J, Fukuchi K, Martin GM (2000) Levels of DNA damage are unaltered in mice overexpressing human catalase in nuclei. Free Radic Biol Med 29:664

Schwartz M, Vissing J (2002) Paternal inheritance of mitochondrial DNA. N Engl J Med 347:576

Schwartz M, Vissing J (2003) New patterns of inheritance in mitochondrial disease. Biochem Biophys Res Commun 310:247

Schwartz M, Vissing J (2004) No evidence for paternal inheritance of mtDNA in patients with sporadic mtDNA mutations. J Neurol Sci 218:99

Shigenaga MK, Hagen TM, Ames BN (1994) Oxidative damage and mitochondrial decay in aging. Proc Natl Acad Sci U S A 91:10771

Shitara H, Hayashi JI, Takahama S, Kaneda H, Yonekawa H (1998) Maternal inheritance of mouse mtDNA in interspecific hybrids: segregation of the leaked paternal mtDNA followed by the prevention of subsequent paternal leakage. Genetics 148:851

Sigurgardottir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P (2000) The mutation rate in the human mtDNA control region. Am J Hum Genet 66:1599

Simone D, Calabrese FM, Lang M, Gasparre G, Attimonelli M (2011) The reference human nuclear mitochondrial sequences compilation validated and implemented on the UCSC genome browser. BMC Genomics 12:517

Singh R, Ellard S, Hattersley A, Harries LW (2006) Rapid and sensitive real-time polymerase chain reaction method for detection and quantification of 3243A>G mitochondrial point mutation. J Mol Diagn 8:225

Slate J, Gemmell NJ (2004) Eve 'n' Steve: recombination of human mitochondrial DNA. Trends Ecol Evol 19:561

Smith KH, Johns DR, Heher KL, Miller NR (1993) Heteroplasmy in Leber's hereditary optic neuropathy. Arch Ophthalmol 111:1486

Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE (1985) The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. Nucleic Acids Res 13:2399

Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321:674

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740

Sondheimer N, Glatz CE, Tirone JE, Deardorff MA, Krieger AM, Hakonarson H (2011) Neutral mitochondrial heteroplasmy and the influence of aging. Hum Mol Genet 20:1653

SPSS I (1989-2006) SPSS 15.0.1 for Windows. SPSS Inc., Chicago.

Stajich J, Block D, Boulez K, Brenner S, Chervitz S, Dagdigian C, Fuellen G, Gilbert J, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall C, Osborne B, Pocock M, Schattner P, Senger M, Stein L, Stupka E, Wilkinson M, Birney E (2002) The BioPerl toolkit: Perl modules for the life sciences. Genome Res. 12:1611

Steinert M, Steinert G (1960) Inhibition de la synthese de lacide desoxyribonucleique de trypanosoma-mega par luree a faible concentration. Experimental Cell Research 19:421

Stoneking M (1993) DNA and recent human evolution. Evol. Anthropol. 2:60

Stoneking M (2000) Hypervariable sites in the mtDNA control region are mutational hotspots. Am J Hum Genet 67:1029

Stoneking M, Soodyall H (1996) Human evolution and the mitochondrial genome. Curr Opin Genet Dev 6:731

Suen DF, Narendra DP, Tanaka A, Manfredi G, Youle RJ (2010) Parkin overexpression selects against a deleterious mtDNA mutation in heteroplasmic cybrid cells. PNAS 107:11835

Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G (1999) Ubiquitin tag for sperm mitochondria. Nature 402:371

Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G (2000) Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos. Biol Reprod 63:582

Tan D, Goerlitz DS, Dumitrescu RG, Han D, Seillier-Moiseiwitsch F, Spernak SM, Orden RA, Chen J, Goldman R, Shields PG (2008) Associations between cigarette smoking and mitochondrial DNA abnormalities in buccal cells. Carcinogenesis 29:1170

Tan DJ, Chang J, Liu LL, Bai RK, Wang YF, Yeh KT, Wong LJ (2006) Significance of somatic mutations and content alteration of mitochondrial DNA in esophageal cancer. BMC Cancer 6:93

Tang S, Huang T (2010) Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. Biotechniques 48:287

Taylor FJR (1974) Implications and Extensions of the Serial Endosymbiosis Theory of the Origin of Eukaryotes. Taxon 23:229

Taylor RW, McDonnell MT, Blakely EL, Chinnery PF, Taylor GA, Howell N, Zeviani M, Briem E, Carrara F, Turnbull DM (2003) Genotypes from patients indicate no paternal mitochondrial DNA contribution. Ann Neurol 54:521

Templeton AR (1996) Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. Genetics 144:1263

Thalmann O, Serre D, Hofreiter M, Lukas D, Eriksson J, Vigilant L (2005) Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. Mol Ecol 14:179

Theves C, Keyser-Tracqui C, Crubezy E, Salles JP, Ludes B, Telmon N (2006) Detection and quantification of the age-related point mutation A189G in the human mitochondrial DNA. J Forensic Sci 51:865

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673

Thyagarajan B, Padua RA, Campbell C (1996) Mammalian mitochondria possess homologous DNA recombination activity. J Biol Chem 271:27536

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 69:1348

Tourmen Y, Baris O, Dessen P, Jacques C, Malthiery Y, Reynier P (2002) Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics 80:71

Tsukihara T, Aoyama H, Yamashita E, Tomizaki T, Yamaguchi H, Shinzawa-Itoh K, Nakashima R, Yaono R, Yoshikawa S (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 A. Science 272:1136

Tzen CY, Mau BL, Wu TY (2007) ND4 mutation in transitional cell carcinoma: Does mitochondrial mutation occur before tumorigenesis? Mitochondrion 7:273

Uusimaa J, Moilanen JS, Vainionpaa L, Tapanainen P, Lindholm P, Nuutinen M, Lopponen T, Maki-Torkko E, Rantala H, Majamaa K (2007) Prevalence, segregation, and phenotype of the mitochondrial DNA 3243A>G mutation in children. Ann Neurol 62:278

Van der Giezen M, Tovar J (2005) Degenerate mitochondria. EMBO Rep 6:525

Van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386

Vo TD, Palsson BO (2007) Building the power house: recent advances in mitochondrial studies through proteomics and systems biology. Am J Physiol Cell Physiol 292:C164

Wai T, Teoli D, Shoubridge EA (2008) The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes. Nat Genet 40:1484

Wallace DC (2007) Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine. Annu Rev Biochem 76:781

Wallace DC, Brown MD, Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. Gene 238:211

Wallace DC, Zheng XX, Lott MT, Shoffner JM, Hodge JA, Kelley RI, Epstein CM, Hopkins LC (1988) Familial mitochondrial encephalomyopathy (MERRF): genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease. Cell 55:601

Wang Q, Boles RG (2006) Individual human hair mitochondrial DNA control region heteroplasmy proportions in mothers and children. Mitochondrion 6:37

Wang Y, Michikawa Y, Mallidis C, Bai Y, Woodhouse L, Yarasheski KE, Miller CA, Askanas V, Engel WK, Bhasin S, Attardi G (2001) Muscle-specific mutations accumulate with aging in critical human mtDNA control sites for replication. Proc Natl Acad Sci U S A 98:4022

Wei YL, Yu CA, Yang P, Li AL, Wen JY, Zhao SM, Liu HX, Ke YN, Campbell W, Zhang YG, Li XH, Liao WQ (2009) Novel mitochondrial DNA mutations associated with Chinese familial hypertrophic cardiomyopathy. Clin Exp Pharmacol Physiol 36:939

Weinreich DM, Rand DM (2000) Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. Genetics 156:385

White HE, Durston VJ, Seller A, Fratter C, Harvey JF, Cross NC (2005) Accurate detection and quantitation of heteroplasmic mitochondrial point mutations by pyrosequencing. Genet Test 9:190

Whittaker RG, Schaefer AM, McFarland R, Taylor RW, Walker M, Turnbull DM (2007) Prevalence and progression of diabetes in mitochondrial disease. Diabetologia 50:2085

Whittam TS, Clark AG, Stoneking M, Cann RL, Wilson AC (1986) Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. Proc Natl Acad Sci U S A 83:9611

Wise CA, Sraml M, Easteal S (1998) Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. Genetics 148:409

Woischnik M, Moraes CT (2002) Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res 12:885

Wong LJ, Wladyka C, Mardach-Verdon R (2004) A mitochondrial DNA mutation in a patient with an extensive family history of Duchenne muscular dystrophy. Muscle Nerve 30:118

Yao J, Chang C, Salmi ML, Hung YS, Loraine A, Roux SJ (2008a) Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. BMC Bioinformatics 9:288

Yao YG, Bravi CM, Bandelt HJ (2004) A call for mtDNA data quality control in forensic science. Forensic Sci Int 141:1

Yao YG, Kong QP, Salas A, Bandelt HJ (2008b) Pseudomitochondrial genome haunts disease studies. J Med Genet 45:769

Zaragoza MV, Fass J, Diegoli M, Lin D, Arbustini E (2010) Mitochondrial DNA variant discovery and evaluation in human Cardiomyopathies through next-generation sequencing. PLoS One 5:e12295

Zhadanov SI, Grechanina EY, Grechanina YB, Gusar VA, Fedoseeva NP, Lebon S, Munnich A, Schurr TG (2007) Fatal manifestation of a de novo ND5 mutation: Insights into the pathogenetic mechanisms of mtDNA ND5 gene defects. Mitochondrion 7:260

Zhou S, Kachhap S, Sun W, Wu G, Chuang A, Poeta L, Grumbine L, Mithani SK, Chatterjee A, Koch W, Westra WH, Maitra A, Glazer C, Carducci M, Sidransky D, McFate T, Verma A, Califano JA (2007) Frequency and phenotypic implications of mitochondrial DNA mutations in human squamous cell cancers of the head and neck. Proc Natl Acad Sci U S A 104:7540

Zouros E, Freeman KR, Ball AO, Pogson GH (1992) Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel Mytilus. Nature 359:412

Zsurka G, Hampel KG, Nelson I, Jardel C, Mirandola SR, Sassen R, Kornblum C, Marcorelles P, Lavoue S, Lombes A, Kunz WS (2011) Severe epilepsy as the major symptom of new mutations in the mitochondrial tRNA(Phe) gene. Neurology 74:507

Zsurka G, Kraytsberg Y, Kudina T, Kornblum C, Elger CE, Khrapko K, Kunz WS (2005) Recombination of mitochondrial DNA in skeletal muscle of individuals with multiple mitochondrial DNA heteroplasmy. Nat Genet 37:873

7. ANNEX

7.1 Annex 1

Els annexos d'aquesta Tesi estan organitzats segons el capítol on es citen, i es presenten de la següent manera:

Capítol 1: Disseny i validació de *primers* per a l'amplificació i seqüenciació del DNAmt

Taula I.1.1. Resultat del BLAST de les regions del DNAmt: **1**) 3914-9074; i **2**) 9582-14479.

Table I.1.2 Descripció detallada del NUMT actualitzat.

Table I.1.3 Alineament de la seqüència d'Homo sapiens amb cinc primats i el cavall.

Capítol 2: Actualització de la base de dades de NUMTs i aplicació en estudis de càncer

Taula I.2.1 Base de dades de NUMTs basada en el genoma humà de referència GRCh37.

Taula I.2.2 Posicions compilades de Santos et al. (2008) que segueixen tenint 0 hits a la filogènia actual (Van Oven et al. 2009) (mit. Tree build 8).

Taula 1.2.3 Posicions amb zero hits en la filogènia del DNAmt que coincideixen amb posicions no idèntiques entre el DNAmt i les insercions nuclears.

Aquesta informació es troba disponible en el CD adjunt dins de la carpeta anomenda "Annex I_Capítol 2" i dins l'arxiu Excel anomenat "Annex_Capítol 2".

Capítol 3: Freqüència i patró de l'heteroplàsmia mitocondrial humana

Figura I.3.1 Implicació de l'heteroplàsmia puntual del DNAmt en l'estructura secundària dels tRNAs, rRNAs i en l'estructura terciària dels complexos COXI, COXII i CitB.

Taula I.3.1 Llistat dels números d'accés i les espècies utilitzades per calcular els índex de conservació a nivell nucleotídic i a nivell aminoacídic.

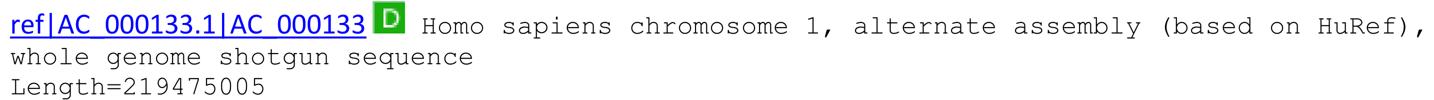
Aquesta informació es troba disponible en el CD adjunt dins de la carpeta anomenada "Annex I_Capítol 3" i dins l'arxiu Excel anomenat "Annex_Capítol 3".

Taula I.3.2 Compilació de les mutacions presents en el genoma mitocondrial de 101 individus. Les heteroplàsmies puntuals i de longitud estan representades en negreta. El sexe i l'haplogrup de cada individu també es troba representat.

Taula I.3.3 Compilació dels resultats per a cada posició heteroplàsmica analitzada (localització, nom de la mostra, tipus d'heteroplàsmia, origen de l'heteroplàsmia, mitjana de l'alçada dels pics, distribució en la base de dades, número de hits en la filogènia del DNAmt (PhyloTree.org) i la reportada per Soares et al. (2009), probabilitat de mutació i índex de conservació a nivell de nucleòtid i d'aminoàcid).

Table I.1.1

1) Blast result for 3914-9074 mtDNA region.



ref|AC_000133.1|AC_000133 Homo sapiens chromosome 1, alternate assembly (based on HuRef), whole genome shotgun sequence
Length=219475005

Max score: 9118
Total score: 1.475e+04
Query coverage: 68%
E value 0.00
Max ident 100%

```
 Score = 9118 bits (4937),  Expect = 0.0
 Identities = 5087/5161 (98%), Gaps = 3/5161 (0%)
 Strand=Plus/Plus

Query  3914  GGGAGTCCGAACTAGTCTCAGGCTTCAACATCGAATACGCCGCAGGCCCCTTCGCCCTAT  3973
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  719   GGGAGTCCGAACTAGTCTCAGGCTTCAACATCGAATACGCCGCAGGCCCCTTCGCCCTAT  778

Query  3974  TCTTCATAGCCGAATACACAAACATTATTATAATAAACACCCTCACCACTACAATCTTCC  4033
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  779   TCTTCATAGCCGAATACACAAACATTATTATAATAAACACCCTCACCACTACAATCTTCC  838
```

```
Query  4034  TAGGAACAACATATGACGCACTCTCCCCTGAACTCTACACAACATATTTTGTCACCAAGA  4093
             |||||||||||||| |||||||||||||||||||||||||||||||||||||||||||||
Sbjct  839   TAGGAACAACATATAACGCACTCTCCCCTGAACTCTACACAACATATTTTGTCACCAAGA  898

Query  4094  CCCTACTTCTAACCTCCCTGTTCTTATGAATTCGAACAGCATACCCCCGATTCCGCTACG  4153
             |||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  899   CCCTACTTCTGACCTCCCTGTTCTTATGAATTCGAACAGCATACCCCCGATTCCGCTACG  958

Query  4154  ACCAACTCATACACCTCCTATGAAAAAACTTCCTACCACTCACCCTAGCATTACTTATAT  4213
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  959   ACCAACTCATACACCTCCTATGAAAAAACTTCCTACCACTCACCCTAGCATTACTTATAT  1018

Query  4214  GATATGTCTCCATACCCATTACAATCTCCAGCATTCCCCCTCAAACCTAAGAAATATGTC  4273
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1019  GATATGTCTCCATACCCATTACAATCTCCAGCATTCCCCCTCAAACCTAAGAAATATGTC  1078

Query  4274  TGATAAAAGAGTTACTTTGATAGAGTAAATAATAGGAGCTTAAACCCCCTTATTTCTAGG  4333
             |||||||||||||||||||||||||||||||||||||| ||||| |||||||||||||||
Sbjct  1079  TGATAAAAGAGTTACTTTGATAGAGTAAATAATAGGAGTTTAAATCCCCTTATTTCTAGG  1138

Query  4334  ACTATGAGAATCGAACCCATCCCTGAGAATCCAAAATTCTCCGTGCCACCTATCACACCC  4393
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1139  ACTATGAGAATCGAACCCATCCCTGAGAATCCAAAATTCTCCGTGCCACCTATCACACCC  1198

Query  4394  CATCCTAAAGTAAGGTCAGCTAAATAAGCTATCGGGCCCATACCCCGAAAATGTTGGTTA  4453
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1199  CATCCTAAAGTAAGGTCAGCTAAATAAGCTATCGGGCCCATACCCCGAAAATGTTGGTTA  1258
```

```
Query  4454  TACCCTTCCCGTACTAATTAATCCCCTGGCCCAACCCGTCATCTACTCTACCATCTTTGC  4513
             || |||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1259  TATCCTTCCCGTACTAATTAATCCCCTGGCCCAACCCGTCATCTACTCTACCATCTTTGC  1318

Query  4514  AGGCACACTCATCACAGCGCTAAGCTCGCACTGATTTTTTACCTGAGTAGGCCTAGAAAT  4573
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1319  AGGCACACTCATCACAGCGCTAAGCTCGCACTGATTTTTTACCTGAGTAGGCCTAGAAAT  1378

Query  4574  AAACATGCTAGCTTTTATTCCAGTTCTAACCaaaaaaaTAAACCCTCGTTCCACAGAAGC  4633
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1379  AAACATGCTAGCTTTTATTCCAGTTCTAACCAAAAAAATAAACCCTCGTTCCACAGAAGC  1438

Query  4634  TGCCATCAAGTATTTCCTCACGCAAGCAACCGCATCCATAATCCTTCTAATAGCTATCCT  4693
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1439  TGCCATCAAGTATTTCCTCACGCAAGCAACCGCATCCATAATCCTTCTAATAGCTATCCT  1498

Query  4694  CTTCAACAATATACTCTCCGGACAATGAACCATAACCAATACTACCAATCAATACTCATC  4753
             |||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||
Sbjct  1499  CTTCAACAATATACTCTCCGGACAATGAACCATAACCAATACCACCAATCAATACTCATC  1558

Query  4754  ATTAATAATCATAATAGCTATAGCAATAAAACTAGGAATAGCCCCCTTTCACTTCTGAGT  4813
             ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1559  ATTAATAATCATAATGGCTATAGCAATAAAACTAGGAATAGCCCCCTTTCACTTCTGAGT  1618

Query  4814  CCCAGAGGTTACCCAAGGCACCCCTCTGACATCCGGCCTGCTTCTTCTCACATGACAAAA  4873
             |||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||
Sbjct  1619  CCCAGAGGTTACCCAAGGCACCCCTCTGACATCCGGCCTGCTCCTTCTCACATGACAAAA  1678
```

```
Query  4874  ACTAGCCCCCATCTCAATCATATACCAAATCTCTCCCTCACTAAACGTAAGCCTTCTCCT  4933
             |||||||||||||||||||||||||||||| ||||||||| |||||||||||||||||||
Sbjct  1679  ACTAGCCCCCATCTCAATCATATACCAAATTTCTCCCTCATTAAACGTAAGCCTTCTCCT  1738

Query  4934  CACTCTCTCAATCTTATCCATCATAGCAGGCAGTTGAGGTGGATTAAACCAAACCCAGCT  4993
             |||||| ||||||||||||||||| |||||||||||||||||||||||||||||||| ||
Sbjct  1739  CACTCTTTCAATCTTATCCATCATGGCAGGCAGTTGAGGTGGATTAAACCAAACCCAACT  1798

Query  4994  ACGCAAAATCTTAGCATACTCCTCAATTACCCACATAGGATGAATAATAGCAGTTCTACC  5053
             |||||||||||||||||||||||||||||||||||||||||||||| |||||||||||||
Sbjct  1799  ACGCAAAATCTTAGCATACTCCTCAATTACCCACATAGGATGAATAACAGCAGTTCTACC  1858

Query  5054  GTACAACCCTAACATAACCATTCTTAATTTAACTATTTATATTATCCTAACTACTACCGC  5113
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1859  GTACAACCCTAACATAACCATTCTTAATTTAACTATTTATATTATCCTAACTACTACCGC  1918

Query  5114  ATTCCTACTACTCAACTTAAACTCCAGCACCACGACCCTACTACTATCTCGCACCTGAAA  5173
             ||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  1919  ATTCCTACTACTCAACTTAAACTCCAGCACCACAACCCTACTACTATCTCGCACCTGAAA  1978

Query  5174  CAAGCTAACATGACTAACACCCTTAATTCCATCCACCCTCCTCTCCCTAGGAGGCCTGCC  5233
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1979  CAAGCTAACATGACTAACACCCTTAATTCCATCCACCCTCCTCTCCCTAGGAGGCCTGCC  2038

Query  5234  CCCGCTAACCGGCTTTTTGCCCAAATGGGCCATTATCGAAGAATTCACAAAAAACAATAG  5293
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2039  CCCGCTAACCGGCTTTTTGCCCAAATGGGCCATTATCGAAGAATTCACAAAAAACAATAG  2098
```

```
Query  5294  CCTCATCATCCCCACCATCATAGCCACCATCACCCTCCTTAACCTCTACTTCTACCTACG  5353
             |||||||||||||||||||||||||| |||||||||||||||||||||||||||||| ||
Sbjct  2099  CCTCATCATCCCCACCATCATAGCCATCATCACCCTCCTTAACCTCTACTTCTACCTGCG  2158

Query  5354  CCTAATCTACTCCACCTCAATCACACTACTCCCCATATCTAACAACGTAAAAATAAAATG  5413
             ||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||||
Sbjct  2159  CCTAATCTACTCCACCTCAATCACACTACTCCCTATATCTAACAACGTAAAAATAAAATG  2218

Query  5414  ACAGTTTGAACATACAAAACCCACCCCATTCCTCCCCACACTCATCGCCCTTACCACGCT  5473
             |||||||||||| |||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  2219  ACAGTTTGAACACACAAAACCCACCCCATTCCTCCCCACACTCATCGCCCTTACCACACT  2278

Query  5474  ACTCCTACCTATCTCCCCTTTTATACTAATAATCTTATAGAAATTTAGGTTAAATACAGA  5533
              ||||||||||||||||||||||| |||||||||||||||||||||||||||||||||||
Sbjct  2279  GCTCCTACCTATCTCCCCTTTTATGCTAATAATCTTATAGAAATTTAGGTTAAATACAGA  2338

Query  5534  CCAAGAGCCTTCAAAGCCCTCAGTAAGTTGCAATACTTAATTTCTGTAACAGCTAAGGAC  5593
             |||||||||||||||||||||||||||||||||||||||||||||| |||||||||||||
Sbjct  2339  CCAAGAGCCTTCAAAGCCCTCAGTAAGTTGCAATACTTAATTTCTGCAACAGCTAAGGAC  2398

Query  5594  TGCAAAACCCCACTCTGCATCAACTGAACGCAAATCAGCCACTTTAATTAAGCTAAGCCC  5653
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2399  TGCAAAACCCCACTCTGCATCAACTGAACGCAAATCAGCCACTTTAATTAAGCTAAGCCC  2458

Query  5654  TTACTAGACCAATGGGACTTAAACCCACAAACACTTAGTTAACAGCTAAGCACCCTAATC  5713
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2459  TTACTAGACCAATGGGACTTAAACCCACAAACACTTAGTTAACAGCTAAGCACCCTAATC  2518
```

```
Query  5714  AACTGGCTTCAATCTACTTCTCCCGCCGCCGGGAAAAAAGGCGGGAGAAGCCCCGGCAGG  5773
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2519  AACTGGCTTCAATCTACTTCTCCCGCCGCCGGGAAAAAAGGCGGGAGAAGCCCCGGCAGG  2578

Query  5774  TTTGAAGCTGCTTCTTCGAATTTGCAATTCAATATGAAAATCACCTCGGAGCTGGTAAAA  5833
             |||||||||||||||||||||||||||||||||||||||||||||||| |||||||||||
Sbjct  2579  TTTGAAGCTGCTTCTTCGAATTTGCAATTCAATATGAAAATCACCTCAGAGCTGGTAAAA  2638

Query  5834  AGAGGCCTAACCCCTGTCTTTAGATTTACAGTCCAATGCTTCACTCAGCCATTTTACCTC  5893
             |||||| |||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2639  AGAGGCTTAACCCCTGTCTTTAGATTTACAGTCCAATGCTTCACTCAGCCATTTTACCTC  2698

Query  5894  ACCCCCACTGATGTTCGCCGACCGTTGACTATTCTCTACAAACCACAAAGACATTGGAAC  5953
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2699  ACCCCCACTGATGTTCGCCGACCGTTGACTATTCTCTACAAACCACAAAGACATTGGAAC  2758

Query  5954  ACTATACCTATTATTCGGCGCATGAGCTGGAGTCCTAGGCACAGCTCTAAGCCTCCTTAT  6013
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2759  ACTATACCTATTATTCGGCGCATGAGCTGGAGTCCTAGGCACAGCTCTAAGCCTCCTTAT  2818

Query  6014  TCGAGCCGAGCTGGGCCAGCCAGGCAACCTTCTAGGTAACGACCACATCTACAACGTTAT  6073
             ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2819  TCGAGCCGAACTGGGCCAGCCAGGCAACCTTCTAGGTAACGACCACATCTACAACGTTAT  2878

Query  6074  CGTCACAGCCCATGCATTTGTAATAATCTTCTTCATAGTAATACCCATCATAATCGGAGG  6133
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2879  CGTCACAGCCCATGCATTTGTAATAATCTTCTTCATAGTAATACCCATCATAATCGGAGG  2938
```

```
Query  6134  CTTTGGCAACTGACTAGTTCCCCTAATAATCGGTGCCCCCGATATGGCGTTTCCCCGCAT  6193
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2939  CTTTGGCAACTGACTAGTTCCCCTAATAATCGGTGCCCCCGATATGGCGTTTCCCCGCAT  2998

Query  6194  AAACAACATAAGCTTCTGACTCTTACCTCCCTCTCTCCTACTCCTGCTCGCATCTGCTAT  6253
             ||||||||||||||||||||||||||| |||||||||||||||||||| |||||||||||
Sbjct  2999  AAACAACATAAGCTTCTGACTCTTACCCCCCTCTCTCCTACTCCTGCTTGCATCTGCTAT  3058

Query  6254  AGTGGAGGCCGGAGCAGGAACAGGTTGAACAGTCTACCCTCCCTTAGCAGGGAACTACTC  6313
             |||||||||||| |||||||||||||||||||||||||||||||| ||||||||||||||
Sbjct  3059  AGTGGAGGCCGGCGCAGGAACAGGTTGAACAGTCTACCCTCCCTTGGCAGGGAACTACTC  3118

Query  6314  CCACCCTGGAGCCTCCGTAGACCTAACCATCTTCTCCTTACACCTAGCAGGTGTCTCCTC  6373
             |||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct  3119  CCACCCTGGAGCCTCCGTAGACCTAACCATCTTCTCCTTACACCTAGCAGGTATCTCCTC  3178

Query  6374  TATCTTAGGGGCCATCAATTTCATCACAACAATTATCAATATAAAACCCCCTGCCATAAC  6433
             ||||||||| |||||||||||||||||||||||||| |||||||||||||||||||||||
Sbjct  3179  TATCTTAGGAGCCATCAATTTCATCACAACAATTATTAATATAAAACCCCCTGCCATAAC  3238

Query  6434  CCAATACCAAACGCCCCTCTTCGTCTGATCCGTCCTAATCACAGCAGTCCTACTTCTCCT  6493
             |||||||||||||||||| |||||||||||||||||||||||||||||| ||||||||||
Sbjct  3239  CCAATACCAAACGCCCCTTTTCGTCTGATCCGTCCTAATCACAGCAGTCTTACTTCTCCT  3298

Query  6494  ATCTCTCCCAGTCCTAGCTGCTGGCATCACTATACTACTAACAGACCGCAACCTCAACAC  6553
             |||||||||||||||||| |||||||||||||||||||||||||||||| |||||||||||
Sbjct  3299  ATCTCTCCCAGTCCTAGCCGCTGGCATCACTATACTACTAACAGACCGTAACCTCAACAC  3358
```

```
Query  6554  CACCTTCTTCGACCCCGCCGGAGGAGGAGACCCCATTCTATACCAACACCTATTCTGATT  6613
             ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3359  CACCTTCTTCGACCCAGCCGGAGGAGGAGACCCCATTCTATACCAACACCTATTCTGATT  3418

Query  6614  TTTCGGTCACCCTGAAGTTTATATTCTTATCCTACCAGGCTTCGGAATAATCTCCCATAT  6673
             |||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
Sbjct  3419  TTTCGGTCACCCTGAAGTTTATATTCTCATCCTACCAGGCTTCGGAATAATCTCCCATAT  3478

Query  6674  TGTAACTTACTACTCCGGaaaaaaaGAACCATTTGGATACATAGGTATGGTCTGAGCTAT  6733
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3479  TGTAACTTACTACTCCGGAAAAAAAGAACCATTTGGATACATAGGTATGGTCTGAGCTAT  3538

Query  6734  GATATCAATTGGCTTCCTAGGGTTTATCGTGTGAGCACACCATATATTTACAGTAGGAAT  6793
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3539  GATATCAATTGGCTTCCTAGGGTTTATCGTGTGAGCACACCATATATTTACAGTAGGAAT  3598

Query  6794  AGACGTAGACACACGAGCATATTTCACCTCCGCTACCATAATCATCGCTATCCCCACCGG  6853
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3599  AGACGTAGACACACGAGCATATTTCACCTCCGCTACCATAATCATCGCTATCCCCACCGG  3658

Query  6854  CGTCAAAGTATTTAGCTGACTCGCCACACTCCACGGAAGCAATATGAAATGATCTGCTGC  6913
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3659  CGTCAAAGTATTTAGCTGACTCGCCACACTCCACGGAAGCAATATGAAATGATCTGCTGC  3718

Query  6914  AGTGCTCTGAGCCCTAGGATTCATCTTTCTTTTCACCGTAGGTGGCCTGACTGGCATTGT  6973
             ||||||||||||||||||||| || |||||||||||||||||||||||||||||||||||
Sbjct  3719  AGTGCTCTGAGCCCTAGGATTTATTTTTCTTTTCACCGTAGGTGGCCTGACTGGCATTGT  3778
```

```
Query  6974  ATTAGCAAACTCATCACTAGACATCGTACTACACGACACGTACTACGTTGTAGCCCACTT  7033
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3779  ATTAGCAAACTCATCACTAGACATCGTACTACACGACACGTACTACGTTGTAGCCCACTT  3838

Query  7034  CCACTATGTCCTATCAATAGGAGCTGTATTTGCCATCATAGGAGGCTTCATTCACTGATT  7093
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3839  CCACTATGTCCTATCAATAGGAGCTGTATTTGCCATCATAGGAGGCTTCATTCACTGATT  3898

Query  7094  TCCCCTATTCTCAGGCTACACCCTAGACCAAACCTACGCCAAAATCCATTTCACTATCAT  7153
             |||||||||||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct  3899  TCCCCTATTCTCAGGCTACACCCTAGACCAAACCTACGCCAAAATCCATTTCGCTATCAT  3958
Query  7154  ATTCATCGGCGTAAATCTAACTTTCTTCCCACAACACTTTCTCGGCCTATCCGGAATGCC  7213
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3959  ATTCATCGGCGTAAATCTAACTTTCTTCCCACAACACTTTCTCGGCCTATCCGGAATGCC  4018

Query  7214  CCGACGTTACTCGGACTACCCCGATGCATACACCACATGAAACATCCTATCATCTGTAGG  7273
             |||||||||||||||||| |||||||||||||||||||||||| |||||||||||||||||
Sbjct  4019  CCGACGTTACTCGGACTATCCCGATGCATACACCACATGAAATATCCTATCATCTGTAGG  4078

Query  7274  CTCATTCATTTCTCTAACAGCAGTAATATTAATAATTTTCATGATTTGAGAAGCCTTCGC  7333
             ||||||||||||||||||||||||||||||||||||||||||| |||||||||||||||||
Sbjct  4079  CTCATTCATTTCTCTAACAGCAGTAATATTAATAATTTTCATAATTTGAGAAGCCTTCGC  4138

Query  7334  TTCGAAGCGAAAAGTCCTAATAGTAGAAGAACCCTCCATAAACCTGGAGTGACTATATGG  7393
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4139  TTCGAAGCGAAAAGTCCTAATAGTAGAAGAACCCTCCATAAACCTGGAGTGACTATATGG  4198
```

```
Query  7394  ATGCCCCCACCCTACCACACATTCGAAGAACCCGTATACATAAAATCTAGACAAAAAAG  7453
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4199  ATGCCCCCACCCTACCACACATTCGAAGAACCCGTATACATAAAATCTAGACAAAAAAG  4258

Query  7454  GAAGGAATCGAACCCCCCAAAGCTGGTTTCAAGCCAACCCCATGGCCTCCATGACTTTTT  7513
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4259  GAAGGAATCGAACCCCCCAAAGCTGGTTTCAAGCCAACCCCATGGCCTCCATGACTTTTT  4318

Query  7514  CAAAAAGGTATTAGAAAAACCATTTCATAACTTTGTCAAAGTTAAATTATAGGCTAAATC  7573
             ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4319  CAAAAAGATATTAGAAAAACCATTTCATAACTTTGTCAAAGTTAAATTATAGGCTAAATC  4378
Query  7574  CTATATATCTTAATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATC  7633
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4379  CTATATATCTTAATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTACTTCCCCTATC  4438

Query  7634  ATAGAAGAGCTTATCACCTTTCATGATCACGCCCTCATAATCATTTTCCTTATCTGCTTC  7693
             |||||||||||||||| |||||||||||||||||||||||||||||||||||||||||||
Sbjct  4439  ATAGAAGAGCTTATCATCTTTCATGATCACGCCCTCATAATCATTTTCCTTATCTGCTTC  4498

Query  7694  CTAGTCCTGTATGCCCTTTTCCTAACACTCACAACAAAACTAACTAATACTAACATCTCA  7753
             ||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4499  CTAGTCCTGTACGCCCTTTTCCTAACACTCACAACAAAACTAACTAATACTAACATCTCA  4558

Query  7754  GACGCTCAGGAAATAGAAACCGTCTGAACTATCCTGCCCGCCATCATCCTAGTCCTCATC  7813
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |||
Sbjct  4559  GACGCTCAGGAAATAGAAACCGTCTGAACTATCCTGCCCGCCATCATCCTAGTCCTTATC  4618
```

```
Query  7814  GCCCTCCCATCCCTACGCATCCTTTACATAACAGACGAGGTCAACGATCCCTCCCTTACC  7873
             |||||||||||||||||||||||||||||||||||||||||||||||||||||| |||||
Sbjct  4619  GCCCTCCCATCCCTACGCATCCTTTACATAACAGACGAGGTCAACGATCCCTCCTTTACC  4678

Query  7874  ATCAAATCAATTGGCCACCAATGGTACTGAACCTACGAGTACACCGACTACGGCGGACTA  7933
             ||||||||||||||||| |||||||||||||||||||| |||||||||||||||||||||
Sbjct  4679  ATCAAATCAATTGGCCATCAATGGTACTGAACCTACGAATACACCGACTACGGCGGACTA  4738

Query  7934  ATCTTCAACTCCTACATACTTCCCCCATTATTCCTAGAACCAGGCGACCTGCGACTCCTT  7993
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4739  ATCTTCAACTCCTACATACTTCCCCCATTATTCCTAGAACCAGGCGACCTGCGACTCCTT  4798
Query  7994  GACGTTGACAATCGAGTAGTACTCCCGATTGAAGCCCCCATTCGTATAATAATTACATCA  8053
             ||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||||
Sbjct  4799  GACGTTGACAATCGAGTAGTACTCCCGGTTGAAGCCCCCATTCGTATAATAATTACATCA  4858

Query  8054  CAAGACGTCTTGCACTCATGAGCTGTCCCCACATTAGGCTTAAAAACAGATGCAATTCCC  8113
             ||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4859  CAAGACGTCTTACACTCATGAGCTGTCCCCACATTAGGCTTAAAAACAGATGCAATTCCC  4918

Query  8114  GGACGTCTAAACCAAACCACTTTCACCGCTACACGACCGGGGGTATACTACGGTCAATGC  8173
             |||||||||||||||||||||||||| ||||||||||| |||||||||||||| ||||||
Sbjct  4919  GGACGTCTAAACCAAACCACTTTCACTGCTACACGACCAGGGGTATACTACGGCCAATGC  4978

Query  8174  TCTGAAATCTGTGGAGCAAACCACAGTTTCATGCCCATCGTCCTAGAATTAATTCCCCTA  8233
             ||||||||||||||||||||||  |||| |||||||||||||||||||||||||||||||
Sbjct  4979  TCTGAAATCTGTGGAGCAAACCA--GTTTTATGCCCATCGTCCTAGAATTAATTCCCCTA  5036
```

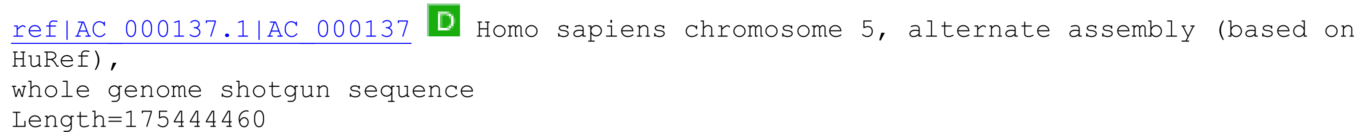
```
Query  8234  AAAATCTTTGAAATAGGGCCCGTATTTACCCTATAGCACCCCCTCTACCCCCTCTAGAGC  8293
             |||||||||||||||||||| |||||||||||||||||||||||||||||||||||||||
Sbjct  5037  AAAATCTTTGAAATAGGGCCTGTATTTACCCTATAGCACCCCCTCTACCCCCTCTAGAGC  5096

Query  8294  CCACTGTAAAGCTAACTTAGCATTAACCTTTTAAGTTAAAGATTAAGAGAACCAACACCT  8353
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5097  CCACTGTAAAGCTAACTTAGCATTAACCTTTTAAGTTAAAGATTAAGAGAACCAACACCT  5156

Query  8354  CTTTACAGTGAAATGCCCCAACTAAATACTACCGTATGGCCCACCATAATTACCCCCATA  8413
             |||||||||||||||||||||||||||||||||||||| |||||||||||||||||||||
Sbjct  5157  CTTTACAGTGAAATGCCCCAACTAAATACTACCGTATGACCCACCATAATTACCCCCATA  5216
Query  8414  CTCCTTACACTATTCCTCATCACCCAACTAAAAATATTAAACACAAACTACCACCTACCT  8473
             ||||||||||||||||||||||||||||||||||||||||| ||||| ||||||||||||
Sbjct  5217  CTCCTTACACTATTCCTCATCACCCAACTAAAAATATTAAATACAAATTACCACCTACCT  5276

Query  8474  CCCTCACCAAAGCCCATaaaaataaaaaaTTATAACAAACCCTGAGAACCAAAATGAACG  8533
             ||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct  5277  CCCTCACCAAAGCCCATAAAAATAAAAAACTATAACAAACCCTGAGAACCAAAATGAACG  5336

Query  8534  AAAATCTGTTCGCTTCATTCATTGCCCCCACAATCCTAGGCCTACCCGCCGCAGTACTGA  8593
             ||||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5337  AAAATCTGTTCACTTCATTCATTGCCCCCACAATCCTAGGCCTACCCGCCGCAGTACTGA  5396

Query  8594  TCATTCTATTTCCCCCTCTATTGATCCCCACCTCCAAATATCTCATCAACAACCGACTAA  8653
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5397  TCATTCTATTTCCCCCTCTATTGATCCCCACCTCCAAATATCTCATCAACAACCGACTAA  5456
```

```
Query  8654  TCACCACCCAACAATGACTAATCAAACTAACCTCAAAACAAATGATAACCATACACAACA  8713
             | ||||||||||||||||||||| ||||||||||||||||||||||| ||||||||||||
Sbjct  5457  TTACCACCCAACAATGACTAATCCAACTAACCTCAAAACAAATGATAGCCATACACAACA  5516

Query  8714  CTAAAGGACGAACCTGATCTCTTATACTAGTATCCTTAATCATTTTTATTGCCACAACTA  8773
             |||| |||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5517  CTAAGGGACGAACCTGATCTCTTATACTAGTATCCTTAATCATTTTTATTGCCACAACTA  5576

Query  8774  ACCTCCTCGGACTCCTGCCTCACTCATTTACACCAACCACCCAACTATCTATAAACCTAG  8833
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5577  ACCTCCTCGGACTCCTGCCTCACTCATTTACACCAACCACCCAACTATCTATAAACCTAG  5636
Query  8834  CCATGGCCATCCCCTTATGAGCGGGCACAGTGATTATAGGCTTTCGCTCTAAGATTAAAA  8893
             |||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
Sbjct  5637  CCATGGCCATCCCCTTATGAGCGGGCGCAGTGATTATAGGCTTTCGCTCTAAGATTAAAA  5696

Query  8894  ATGCCCTAGCCCACTTCTTACCACAAGGCACACCTACACCCCTTATCCCCATACTAGTTA  8953
             ||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  5697  ATGCCCTAGCCCACTTCTTACCACAAGGCACACCTACACCCCTTATCCCTATACTAGTTA  5756

Query  8954  TTATCGAAACCATCAGCCTACTCATTCAACCAATAGCCCTGGCCGTACGCCTAACCGCTA  9013
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5757  TTATCGAAACCATCAGCCTACTCATTCAACCAATAGCCCTGGCCGTACGCCTAACCGCTA  5816

Query  9014  ACATTACTGCAGGCCACCTACTCATGCACCTAATTGGAAGCGCCACCCTAGCAATATCAA  9073
             |||||||||||| ||||||||||||||||||||||||||||||||| |||||||||||||
Sbjct  5817  ACATTACTGCAG-CCACCTACTCATGCACCTAATTGGAAGCGCCACACTAGCAATATCAA  5875
```

```
Query  9074  C  9074
             |
Sbjct  5876  C  5876
```

2) Blast result for 9582-14479 mtDNA region.

Max score: 1.082e+04
Total score: 3.019e+04
Query coverage: 79%
E value 0.00
Max ident 100%



ref|AC_000137.1|AC_000137 D Homo sapiens chromosome 5, alternate assembly (based on HuRef),
whole genome shotgun sequence
Length=175444460

```
 Score = 5954 bits (3224), Expect = 0.0
 Identities = 4360/4918 (88%), Gaps = 39/4918 (0%)
 Strand=Plus/Minus

Query  9582      CCCCTAGAAGTCCCACTCCTAAACACATCCGTATTACTCGCATCAGGAGTATCAATCACC  9641
                 |||| |||||||||||| || |||||||||||| |||||||||||||||| ||||||||
Sbjct  94573924  CCCCCAGAAGTCCCACTTCTGAACACATCCGTACTACTCGCATCAGGAGTCTCAATCACT  94573865

Query  9642      TGAGCTCACCATAGTCTAATAGAAAACAACCGAAACCAAATAATTCAAGCACTGCTTATT  9701
                 ||||| ||||| || ||||||||||| || ||||||||||||||||||||  | ||||||
Sbjct  94573864  TGAGCCCACCACAGCCTAATAGAAAATAATCGAAACCAAATAATTCAAGCGTTACTTATT  94573805
```

```
Query  9702      ACAATTTTACTGGGTCTCTATTTTACCCTCCTACAAGCCTCAGAGTACTTCGAGTCTCCC  9761
                 ||||||||||| ||| | || |||||||||||||||||||||||||| |||||  | |||
Sbjct  94573804  ACAATTTTACTAGGTGTTTACTTTACCCTCCTACAAGCCTCAGAGTATTTCGAAACGCCC  94573745

Query  9762      TTCACCAT-T-T-CCGACGGCATCTACGGCTCAACATTTTTTGTAGCCACAGGCTTCCAC  9818
                 || ||||| | |  |||||||||||| ||||||||||||||||||||||||||||| |||
Sbjct  94573744  TTTACCATCTCTGACGACGGCATCTATGGCTCAACATTTTTTGTAGCCACAGGCTTTCAC  94573685

Query  9819      GGACTTCACGTCATTATTGGCTCAACTTTCCTCACTATCTGCTTCATCCGCCAACTAATA  9878
                 ||||| |||||||| ||||| ||||||||||||||||| ||| ||||||||||||||| |
Sbjct  94573684  GGACTCCACGTCATCATTGGATCAACTTTCCTCACTATTTGCCTCATCCGCCAACTAACA  94573625

Query  9879      TTTCACTTTACATCCAAACATCACTTTGGCTTCGAAGCCGCCGCCTGATACTGGCATTTT  9938
                 |||||||||||||| ||||||||||| ||||| |||||||||||||| ||||| || ||
Sbjct  94573624  TTTCACTTTACATCTAAACATCACTTCGGCTTTGAAGCCGCCGCCTGGTACTGACACTTC  94573565

Query  9939      GTAGATGTGGTTTGACTATTTCTGTATGTCTCCATCTATTGATGAGGGTCTTACTCTTTT  9998
                  ||||||| || ||||||||||| || ||||| ||||| |||||||| || |||||||||
Sbjct  94573564  ATAGATGTAGTCTGACTATTTCTATACGTCTCAATCTACTGATGAGGATCCTACTCTTTT  94573505

Query  9999      AGTATAAATAGTACCGTTAACTTCCAATTAACTAGTTTTGACAACATTCAAAAAAGAGTA  10058
                 |||||||| ||||||||||||||||||||||||||||||||  | |||||||||||||||
Sbjct  94573504  AGTATAAACAGTACCGTTAACTTCCAATTAACTAGTTTTGATGATATTCAAAAAAGAGTA  94573445

Query  10059     ATAAACTTCGCCTTAATTTTAATAATCAACACCCTCCTAGCCTTACTACTAATAATTATT  10118
                 |||||||||||| ||||| ||||| |||| |||||||||||| ||||| ||||| |||||
Sbjct  94573444  ATAAACTTCGCCCTAATTCTAATAGTCAATACCCTCCTAGCCCTACTATTAATAGTTATT  94573385
```

```
Query  10119     ACATTTTGACTACCACAACTCAACGGCTACATAGAAAAATCCACCCCTTACGAGTGCGGC  10178
                 ||||| ||| |||||||||||||| | |||||||||||||||| ||||||||| || |||
Sbjct  94573384  ACATTCTGATTACCACAACTCAACAGTTACATAGAAAAATCCAACCCTTACGAATGTGGC  94573325

Query  10179     TTCGACCCTATATCCCCCGCCCGCGTCCCTTTCTCCATAAAATTCTTCTTAGTAGCTATT  10238
                 ||||||||  ||| |||| ||||| | ||||||||||| ||||||||||||||||| ||
Sbjct  94573324  TTCGACCCCCTATTCCCCACCCGCATTCCTTTCTCCATGAAATTCTTCTTAGTAGCCATC  94573265

Query  10239     ACC-TTCTTATTATTTGATCTAGAAATTGCCCTCCTTTTACCCCTACCATGAGCCCTACA  10297
                 ||| ||| |||||||||| ||||||||||| |||||  | ||| ||||||||||||||||
Sbjct  94573264  ACCTTTC-TATTATTTGACCTAGAAATTGCTCTCCTACTGCCCTTACCATGAGCCCTACA  94573206

Query  10298     AACAACTAACCTGCCACTAATAGTTATGTCATCCCTCTTATTAATCATCATCCTAGCCCT  10357
                 |||||| ||| | ||||||| ||| || |||||||||||||||||||| |||||| ||||
Sbjct  94573205  AACAACCAACTTACCACTAACAGTCATATCATCCCTCTTATTAATCATTATCCTAACCCT  94573146

Query  10358     AAGTCTGGCCTATGAGTGACT-ACAAAAAGGATTAGACTGAACCGAATTGGTATATAGTT  10416
                 ||| || ||||| || ||| | |||||| |||||||||||| ||||||||||| ||||||
Sbjct  94573145  AAGCCTAGCCTACGAATGA-TCACAAAAGGGATTAGACTGAGCCGAATTGGTACATAGTT  94573087

Query  10417     TAAACAAAACGAATGATTTCGACTCATTAAATTATGATAATCATATTTACCAAATGCCCC  10476
                 |||| ||||||||||||||||||| |||||||||||||| ||||||||||||||||||||
Sbjct  94573086  TAAATAAAACGAATGATTTCGACTGATTAAATTATGATAGTCATATTTACCAAATGCCCC  94573027

Query  10477     TCATTTACATAAATATTATACTAGCATTTACCATCTCACTTCTAGGAATACTAGTATATC  10536
                 | ||||| ||||||||||||||||||||||| |||||||||||||||||| ||||||| |
Sbjct  94573026  TTATTTATATAAATATTATACTAGCATTTACTATCTCACTTCTAGGAATATTAGTATACC  94572967
```

```
Query  10537     GCTCACACCTCATATCCTCCCTACTATGCCTAGAAGGAATAATACTATCGCTGTTCATTA  10596
                 ||| |||||| ||||||||||||||||| ||||||||||||||| | ||  | || ||||
Sbjct  94572966  GCTTACACCTGATATCCTCCCTACTATGTCTAGAAGGAATAATATTGTCATTATTTATTA  94572907

Query  10597     TAGCTACTCTCATAACCCTCAACACCCACTCCCTCTTAGCCAATATTGTGCCTATTGCCA  10656
                 | ||||||||||||||||| ||||||||||||||||||||||| || ||||||||  |||
Sbjct  94572906  TGGCTACTCTCATAACCCTTAACACCCACTCCCTCTTAGCCAACATCGTGCCTATCACCA  94572847

Query  10657     TACTAGTCTTTGCCGCCTGCGAAGCAGCGGTGGGCCTAGCCCTACTAGTCTCAATCTCCA  10716
                 || |||||||||| |||||||| ||||| || ||||||||||||||||| ||||||||||
Sbjct  94572846  TATTAGTCTTTGCTGCCTGCGAGGCAGCAGTAGGCCTAGCCCTACTAGTTTCAATCTCCA  94572787

Query  10717     ACACATATGGCCTAGACTACGTACATAACCTAAACCTACTCCAATGCTAAAACTAATCGT  10776
                 ||||||| ||| | |||||||  ||||||||||||||||||||||||||||||||||  |
Sbjct  94572786  ACACATACGGCTTGGACTACGCCCATAACCTAAACCTACTCCAATGCTAAAACTAATTAT  94572727

Query  10777     CCCAACAATTATATTACTACCACTGACATGACTT-TCCAAAAAACACATAATTTGAATCA  10835
                  |||||||| |||||||||||||| |||||| || ||||||||||| |||||||||||||
Sbjct  94572726  TCCAACAATCATATTACTACCACTAACATGA-TTCTCCAAAAAACATATAATTTGAATCA  94572668

Query  10836     ACACAACCACCCACAGCCTAATTATTAGCATCATCCCTCTACTATTTTTTAACCAAATCA  10895
                 ||||||||||||||||||||||||| ||||| ||||| ||||||||||| ||||| ||||
Sbjct  94572667  ACACAACCACCCACAGCCTAATTATCAGCATTATCCCCCTACTATTTTTCAACCAGATCA  94572608

Query  10896     ACAACAACCTATTTAGCTGTTCCCCAACCTTTTCCTCCGACCCCCTAACAACCCCCCTCC  10955
                 |||||||||||| ||||| ||| | | |||| ||||||||||||||||| |||||||| |
Sbjct  94572607  ACAACAACCTATATAGCTATTCTCTATCCTTCTCCTCCGACCCCCTAACGACCCCCCTTC  94572548
```

```
Query  10956     TAATACTAACTACCTGACTCCTACCCCTCACAATCATGGCAAGCCAACGCCACTTATCCA  11015
                 ||||||  || ||||||||||||||||||| |||||| ||||||||||||||| ||| ||
Sbjct  94572547  TAATACGGACAACCTGACTCCTACCCCTCATAATCATAGCAAGCCAACGCCACCTATTCA  94572488

Query  11016     GTGAACCACTATCACGaaaaaaaCTCTACCTCTCTATACTAATCTCCCTACAAATCTCCT  11075
                   ||||| ||||||||||||||||||||| ||||||||||||||| ||| |||| |||||
Sbjct  94572487  ACGAACCCCTATCACGAAAAAAACTCTACGTCTCTATACTAATCTTCCTCCAAAACTCCT  94572428

Query  11076     TAATTATAACATTCACAGCCACAGAACTAATCATATTTTATATCTTCTTCGAAACCACAC  11135
                 ||||||||||||||||||||||||||||||| ||||| || |||||||||||| ||||||
Sbjct  94572427  TAATTATAACATTCACAGCCACAGAACTAATTATATTCTACATCTTCTTCGAAGCCACAC  94572368

Query  11136     TTATCCCCACCTTGGCTATCATCACCCGATGAGGCAACCAGCCAGAACGCCTGAACGCAG  11195
                 ||||||||||| | ||||| ||||||||||||||||||||  | ||||| || |||||||
Sbjct  94572367  TTATCCCCACCCTAGCTATTATCACCCGATGAGGCAACCAAACGGAACGTCTAAACGCAG  94572308

Query  11196     GCACATACTTCCTATTCTACACCCTAGTAGGCTCCCTTCCCCTACTCATCGCACTAATTT  11255
                 ||||||| |||||||| || |||||||| |||||||| ||||||||||||||| ||||
Sbjct  94572307  GCACATATTTCCTATTTTATACCCTAGTGGGCTCCCTCCCCCTACTCATCGCATTAATCC  94572248

Query  11256     ACACTCACAACACCCTAGGCTCACTAAACATTCTACTACTCACTCTCACTGCCCAAGAAC  11315
                 |||| ||||||||||| ||||||||||| ||| || ||||||| || ||||||||||| |
Sbjct  94572247  ACACACACAACACCCTGGGCTCACTAAATATTTTATTACTCACCCTTACTGCCCAAGAGC  94572188

Query  11316     TATCAAACTCCTGAGCCAACAACTTAATATGACTAGCTTACACAATAGCTTTTATAGTAA  11375
                 |||||||||||||||| |  ||||||||||||||||| |||||||||||||| |||||||
Sbjct  94572187  TATCAAACTCCTGAGCTAGTAACTTAATATGACTAGCATACACAATAGCTTTCATAGTAA  94572128
```

```
Query  11376      AGATACCTCTTTACGGACTCCACTTATGACTCCCTAAAGCCCATGTCGAAGCCCCCATCG  11435
                  | || || || ||| |||||||| |||||||||||||||||||||| ||||||||||| |
Sbjct  94572127   AAATGCCCCTCTACAGACTCCACCTATGACTCCCTAAAGCCCATGTTGAAGCCCCCATTG  94572068

Query  11436      CTGGGTCAATAGTACTTGCCGCAGTACTCTTAAAACTAGGCGGCTATGGTATAATACGCC  11495
                  |||| ||||||||||| || ||||| ||||||||| |||| |||||||| ||||| ||||
Sbjct  94572067   CTGGCTCAATAGTACTCGCTGCAGTGCTCTTAAAATTAGGTGGCTATGGCATAATGCGCC  94572008

Query  11496      TCACACTCATTCTCAACCCCCTGACAAAACACATAGCCTACCCCTTCCTTGTACTATCCC  11555
                  |||| |||||||||||||||||||||||||||||||| || ||||||||  ||||||| |
Sbjct  94572007   TCACCCTCATTCTCAACCCCCTGACAAAACACATAGCTTATCCCTTCCTCATACTATCTC  94571948

Query  11556      TATGAGGCATAATTATAACAAGCTCCATCTGCCTACGACAAACAGACCTAAAATCGCTCA  11615
                  ||||||||||||| ||||||||||||||||||||||||||||||||| || |||| ||||
Sbjct  94571947   TATGAGGCATAATCATAACAAGCTCCATCTGCCTACGACAAACAGACTTAGAATCACTCA  94571888

Query  11616      TTGCATACTCTTCAATCAGCCACATAGCCCTCGTAGTAACAGCCATTCTCATCCAAACCC  11675
                  |||||| ||||||||| ||||||||||||||| |||||||||| ||||||||||||||||
Sbjct  94571887   TTGCATGCTCTTCAATTAGCCACATAGCCCTCATAGTAACAGCTATTCTCATCCAAACCC  94571828

Query  11676      CCTGAAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTAC  11735
                  |||||||||||||||||||||||||||||||||| |||||||| ||||||||||| |||
Sbjct  94571827   CCTGAAGCTTCACCGGCGCAGTCATTCTCATAATTGCCCACGGACTTACATCCTCCTTAT  94571768

Query  11736      TATTCTGCCTAGCAAACTCAAACTACGAACGCACTCACAGTCGCATCATAATCCTCTCTC  11795
                  | |||||||||||||||||||||||||| || || ||||| ||||||||||| |||||||
Sbjct  94571767   TTTTCTGCCTAGCAAACTCAAACTACGAGCGAACCCACAGCCGCATCATAATTCTCTCTC  94571708
```

```
Query  11796     AAGGACTTCAAACTCTACTCCCACTAATAGCTTTTTGATGACTTCTAGCAAGCCTCGCTA  11855
                 ||||||||||||| ||||||||||||||||| |||||||||||||| ||||||||||| |
Sbjct  94571707  AAGGACTTCAAACCCTACTCCCACTAATAGCCTTTTGATGACTTCTGGCAAGCCTCGCCA  94571648

Query  11856     ACCTCGCCTTACCCCCCACTATTAACCTACTGGGAGAACTC-TCTGTGCTAGTAACCACG  11914
                 ||||||| || |||||||| ||||||||||| ||||||||| || || ||||| ||| |
Sbjct  94571647  ACCTCGCTTTGCCCCCCACCATTAACCTACTAGGAGAACTCTTC-GTACTAGTGACCTCA  94571589

Query  11915     TTCTCCTGATCAAATATCACTCTCCTACTTACAGGACTCAACATACTAGTCACAGCCCTA  11974
                 |||||||||||||| ||||| |||||||| |||||||| ||||||||| |||||||||||
Sbjct  94571588  TTCTCCTGATCAAACATCACCCTCCTACTCACAGGACTTAACATACTAATCACAGCCCTA  94571529

Query  11975     TACTCCCTCTACATATTTACCACAACACAATGGGGCTCACTCACCCACCACATTAACAAC  12034
                 ||||||||||| |||||||||||||||||||| ||||| ||||| |||||||||||||||
Sbjct  94571528  TACTCCCTCTATATATTTACCACAACACAATGAGGCTCGCTCACACACCACATTAACAAC  94571469

Query  12035     ATAAAACCCTCA-TTCACACGAGAAAACACCCTCATGTTCATACACCTATCCCCCATTCT  12093
                 |||||||||||| || |||||||||||||| ||||| |||||||||||||||||||| ||
Sbjct  94571468  ATAAAACCCTCACTT-ACACGAGAAAACACTCTCATATTCATACACCTATCCCCCATCCT  94571410

Query  12094     CCTCCTATCCCTCAACCCCGACATCATTACCGGGTTTT-CCTCTTGTAAATATAGTTTAA  12152
                 |||||||||||| ||||| || ||||||||| |||||| |||| ||||||||||||||||
Sbjct  94571409  CCTCCTATCCCTTAACCCTGATATCATTACC-GGTTTTACCTCCTGTAAATATAGTTTAA  94571351

Query  12153     CCAAAACATCAGATTGTGAATCTGACAACAGAGGCTTACGA-CCCCTTATTTACCGAGAA  12211
                 ||||||||||||||||||||||||| |||||||||| || | ||||||||||||||||||
Sbjct  94571350  CCAAAACATCAGATTGTGAATCTGATAACAGAGGCTCAC-AGCCCCTTATTTACCGAGAA  94571292
```

```
Query  12212     AGCTCACAAGAACTGCTAACTCATGCCCCCATGTCTAACAACATGGCTTTCTCAACTTTT  12271
                 |||||| ||||||||||||||||| | |||||||||||||||||||||||||| ||||||
Sbjct  94571291  AGCTCATAAGAACTGCTAACTCATACTCCCATGTCTAACAACATGGCTTTCTCGACTTTT  94571232

Query  12272     AAAGGATAACAGCTATCCATTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAATAA  12331
                 ||||||||||||| |||| |||||||||||||||||||||||||||||||||||||||||
Sbjct  94571231  AAAGGATAACAGCCATCCGTTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAATAA  94571172

Query  12332     AAGTAATAACCATGCACACTACTATAACCACCCTAACCCTGACTTCCCTAATTCCCCCCA  12391
                 |||||||||||||| |  | || ||||||| |||| |||| |||||| |||||||||| |
Sbjct  94571171  AAGTAATAACCATGTATGCCACCATAACCATCCTAGCCCTAACTTCCTTAATTCCCCCGA  94571112

Query  12392     TCCTTACCACCCTCGTTAACCCTAACaaaaaaaaCTCATACCCCCATTATGTAAAATCCA  12451
                 || || ||||| || | |||||||||||||| |  ||||||||||| ||||||||||| |
Sbjct  94571111  TCATTGCCACCTTCATCAACCCTAACAAAAAGAGTTCATACCCCCACTATGTAAAATCAA  94571052

Query  12452     TTGTCGCATCCACCTTTATTATCAGTCTCTTCCCCACAACAATATTCATGTGCCTAGACC  12511
                 || | |||||| |||||||||| || ||| ||||||||||||||||||| ||||||||||
Sbjct  94571051  TTATTGCATCCGCCTTTATTATTAGCCTCCTCCCCACAACAATATTCATATGCCTAGACC  94570992

Query  12512     AAGAAGTTATTATCTCGAACTGACACTGAGCCACAACCCAAACAACCCAGCTCTCCCTAA  12571
                 |||||||||||||||| |||||||||||||| |||||||||||||  || ||||| ||||
Sbjct  94570991  AAGAAGTTATTATCTCAAACTGACACTGAGCAACAACCCAAACAATACAACTCTCGCTAA  94570932

Query  12572     GCTTCAAACTAGACTACTTCTCCATAATATTCATCCCTGTAGCATTGTTCGTTACATGGT  12631
                 |||| ||||||||||| |||||||||||||| |||||||||||| | ||||| ||||| |
Sbjct  94570931  GCTTTAAACTAGACTATTTCTCCATAATATTAATCCCTGTAGCACTATTCGTCACATGAT  94570872
```

```
Query  12632     CCATCATAGAATTCTCACTGTGATATATAAACTCAGACCCAAACATTAATCAGTTCTTCA  12691
                 ||||||||||||||||||| || ||||||| ||||||||||||||||||||| || ||||
Sbjct  94570871  CCATCATAGAATTCTCACTATGGTATATAAGCTCAGACCCAAACATTAATCAATTTTTCA  94570812

Query  12692     AATATCTACTCATCTTCCTAATTACCATACTAATCTTAGTTACCGCTAACAACCTATTCC  12751
                 | ||||| || || |||||||| |||||||||||| ||||||| ||||||||||||||||
Sbjct  94570811  AGTATCTGCTTATTTTCCTAATCACCATACTAATCCTAGTTACAGCTAACAACCTATTCC  94570752

Query  12752     AACTGTTCATCGGCTGAGAGGGCGTAGGAATTATATCCTTCTTGCTCATCAGTTGATGAT  12811
                 |||| ||||||||||||||||||||||| |||||||||||||| ||||| || || || |
Sbjct  94570751  AACTCTTCATCGGCTGAGAGGGCGTAGGGATTATATCCTTCTTACTCATTAGCTGGTGGT  94570692

Query  12812     ACGCCCGAGCAGATGCCAACACAGCAGCCATTCAAGCAATCCTATACAACCGTATCGGCG  12871
                 | || ||| |||||||||||||||||||| | ||||||||||||||||||| ||| || |
Sbjct  94570691  ATGCTCGAACAGATGCCAACACAGCAGCCGTCCAAGCAATCCTATACAACCATATTGGTG  94570632

Query  12872     ATATCGGTTTCATCCTCGCCTTAGCATGATTTATCCTACACTCCAACTCATGAGACCCAC  12931
                 | ||| | |||||||| ||| ||||||||||| |||||||||||||||||||||| ||||
Sbjct  94570631  ACATCAGCTTCATCCTAGCCCTAGCATGATTTTTCCTACACTCCAACTCATGAGAACCAC  94570572

Query  12932     AACAAATAGCCCTTC-TAAACGCTAA-TCCAAGCCTCACCCCACTACTAGGCCTCCTCCT  12989
                 ||||||||  || || |||  |||||  |||| ||| | ||||||||||||  |||||||
Sbjct  94570571  AACAAATATTCC-TCTTAAGTGCTAACCCCAA-CCTTATCCCACTACTAGGTTTCCTCCT  94570514

Query  12990     AGCAGCAGCAGGCAAATCAGCCCAATTAGGTCTCCACCCCTGACTCCCCTCAGCCATAGA  13049
                 |||||||||||| |||||||| ||| |||| || ||||||||||||||||||||||| ||
Sbjct  94570513  AGCAGCAGCAGGAAAATCAGCTCAACTAGGCCTTCACCCCTGACTCCCCTCAGCCATGGA  94570454
```

```
Query  13050     AGGCCCCACCCCAGTCTCAGCCCTACTCCACTCAAGCACTATAGTTGTAGCAGGAATC-T  13108
                 |||||| ||||| |||||||||||||||||||||||||| ||||||||||| ||  || |
Sbjct  94570453  AGGCCCTACCCCTGTCTCAGCCCTACTCCACTCAAGCACCATAGTTGTAGCTGGGGTCTT  94570394

Query  13109     TCTTACTCATCCGCTTCCACCCCCTAGCAGAAAATAGCCCACTAATCCAAACTCTAACAC  13168
                 || ||||||||||||||||||||||||||||||||||||| | || |||||| || ||||
Sbjct  94570393  TC-TACTCATCCGCTTCCACCCCCTAGCAGAAAATAGCCCGCCAACCCAAACCCTCACAC  94570335

Query  13169     TATGCTTAGGCGCTATCACCACTCTGTTCGCAGCAGTCTGCGCCCTTACACAAAATGACA  13228
                 |||| ||||||||||| || || ||  ||||||||||||||||||| ||||||||||| |
Sbjct  94570334  TATGTTTAGGCGCTATAACTACCCTACTCGCAGCAGTCTGCGCCCTCACACAAAATGATA  94570275

Query  13229     TCaaaaaaaTCGTAGCCTTCTCCACTTCAAGTCAACTAGGACTCATAATAGTTACAATCG  13288
                 ||||||||||||| |||||||| |||||||| ||| | |||||||||||||| |||||||
Sbjct  94570274  TCAAAAAAATCGTGGCCTTCTCTACTTCAAGCCAATTGGGACTCATAATAGTCACAATCG  94570215

Query  13289     GCATCAACCAACCACACCTAGCATTCCTGCACATCTGTACCCACGCCTTCTTCAAAGCCA  13348
                 |||||||||| ||| ||||||||||||| |||||||||||||||||||||||||||||||
Sbjct  94570214  GCATCAACCAGCCATACCTAGCATTCCTACACATCTGTACCCACGCCTTCTTCAAAGCCA  94570155

Query  13349     TACTATTTATGTGCTCCGGGTCCATCATCCACAACCTTAACAATGAACAAGATATTCGAA  13408
                 | |||||| | |||||||| ||||| ||||| ||||| |||||||||||||| |||||||
Sbjct  94570154  TGCTATTTGTATGCTCCGGATCCATTATCCATAACCTCAACAATGAACAAGACATTCGAA  94570095

Query  13409     AAATAGGAGGACTACTCAAAACCATACCTCTCACTTCAACCTCCCTCACCATTGGCAGCC  13468
                 |||||||||||||||||||||  ||||| |||||||||||||||||||||||||||||||
Sbjct  94570094  AAATAGGAGGACTACTCAAAATTATACCCCTCACTTCAACCTCCCTCACCATTGGCAGCC  94570035
```

```
Query  13469     TAGCA-TTAGCAGGAATACCTTTCCTCACAGGTTTCTACTCCAAAGACCACATCATCGAA  13527
                 | ||| || |||||||| || |||||||| || ||||| |||| ||||| ||||||| ||
Sbjct  94570034  TGGCACTT-GCAGGAATGCCCTTCCTCACGGGCTTCTATTCCATAGACCTCATCATCAAA  94569976

Query  13528     ACCGCAAACATATCATACACAAACGCCTGAGCCCTATCTATTACTCTCATCGCTACCTCC  13587
                 |||||||| ||||||||||| |||||||||||||||||||||||||||||||| || |||
Sbjct  94569975  ACCGCAAATATATCATACACCAACGCCTGAGCCCTATCTATTACTCTCATCGCCACTTCC  94569916

Query  13588     CTGACAAGCGCCTATAGCACTCGAATAATTCTTCTCACCCTAACAGGTCAACCTCGCTTC  13647
                 || ||||||||| ||||||||||||||||||| |||||||||||||| |||||||| |||
Sbjct  94569915  CTAACAAGCGCCAATAGCACTCGAATAATTCTCCTCACCCTAACAGGCCAACCTCGTTTC  94569856

Query  13648     CCCACCCTTACTAACATTAACGAAAATAACCCCACCCTACTAAACCCCATTAAACGCCTG  13707
                 || ||||| || ||||| |||||||| ||||| ||||| |||| |||||| ||||||||
Sbjct  94569855  CCAACCCTAACCAACATCAACGAAAACAACCCTACCCTGCTAAGCCCCATCAAACGCCTA  94569796

Query  13708     GCAGCCGGAAGCCTATTCGCAGGATTTCTCATTACTAACAACATTTCCCCCGCATCC--C  13765
                  ||  ||||||||||||||||||||||||||| || | |||||||| |||| |||||  |
Sbjct  94569795  ACAATCGGAAGCCTATTCGCAGGATTTCTCATCACCAGCAACATTTTCCCCACATCCATC  94569736

Query  13766     CCTTCCAAACAACAATCCCCCT-CTACCTAAAACTCACAGCCCTCG-CTGTCACTTTCCT  13823
                 ||  |||||  |||||||| || | || |||||||||||||||| | |  |||| |||||
Sbjct  94569735  CC--CCAAATGACAATCCCACTTC-ACTTAAAACTCACAGCCCTAGGCA-TCACCTTCCT  94569680

Query  13824     AGGACTTCTAACAGCCCTAGACCTCAACTACCTAACCAACAAACTTAAAATAAAATCCCC  13883
                 ||||||||| ||||||||||||||||||||| ||||||||||||| |||||||||  |||
Sbjct  94569679  AGGACTTCTGACAGCCCTAGACCTCAACTACTTAACCAACAAACTCAAAATAAAAAACCC  94569620
```

```
Query  13884     ACTATGCACATTTTA-TTTCTCCAACATACTCGGATTCTA-CCCTAGCATCACACACCGC  13941
                 ||| || |||||| | ||| |||||||||||||||||||| |||| | || |  ||||||
Sbjct  94569619  ACTGTGTACATTTCACTTT-TCCAACATACTCGGATTCTATCCCT-GTATTATTCACCGC  94569562

Query  13942     ACAATCCCCTATCTAGGCCTTCTTACGAGCCAAAACCTGCCCCTACTCCTCCTAGACCTA  14001
                 ||||||||| | ||| ||||||  |  ||||||||  |  |||||||  |  ||||||||
Sbjct  94569561  ACAATCCCCCACCTAAGCCTTCCCATAAGCCAAAATTTATCCCTACTTTTATTAGACCTA  94569502

Query  14002     ACCTGACTAGAAAAGCTATTACCTAAAACAATTTCACAGCACCAAATCTCCACCTCCATC  14061
                 || ||||||||||| ||| ||||||||||||  |||||| |||||| |||| ||||||||
Sbjct  94569501  ACTTGACTAGAAAAACTAATACCTAAAACAACCTCACAGTACCAAACCTCCGCCTCCATC  94569442

Query  14062     ATCACCTCAACCCAAAAAGGCATAATTAAACTTTACTTCCTCTCTTTCTTCTTCCCACTC  14121
                 ||||||||||| |||||||||||||| |||||||||||||||| || |||||||||||||
Sbjct  94569441  ATCACCTCAACTCAAAAAGGCATAATCAAACTTTACTTCCTCT-TT-CTTCTTCCCACTC  94569384

Query  14122     ATCCTAACCCTACTCCTAATCACATAACCTATTCCCCCGAGCAATCTCAATTACAATATA  14181
                  |||||||||| |||||||||||||||||| || ||||||||| ||||||||||||||||
Sbjct  94569383  CTCCTAACCCTCCTCCTAATCACATAACCTGTTACCCCGAGCACTCTCAATTACAATATA  94569324

Query  14182     TACACCAACAAACAATGTTCAACCAGTAACTACTACTAATCAACGCCCATAATCATACAA  14241
                 ||||||||||||||| ||||||||||||||||| || |||||| |||| ||||| || ||
Sbjct  94569323  TACACCAACAAACAAGGTTCAACCAGTAACTACCACCAATCAATGCCCGTAATCGTATAA  94569264

Query  14242     AGCCCCCGCACCAATAGGATCCTCCCGAATCAACCCTGACCCCTCTCCTTCATAAATTAT  14301
                 |||||||||||||||||||||||||||||||||  ||| |||||| ||||||||||||||
Sbjct  94569263  AGCCCCCGCACCAATAGGATCCTCCCGAATCAATGCTGGCCCCTCCCCTTCATAAATTAT  94569204
```

```
Query  14302     TCAGCTTCCTACACTATTAAAGTTTACCACAACCACCACCCCATCATACTCTTTCACCCA  14361
                 ||| ||||||||||||||||| || ||||||| ||||||||||||||| | ||| |||||
Sbjct  94569203  TCAACTTCCTACACTATTAAAATTCACCACAATCACCACCCCATCATATTTTTTACCCA  94569144

Query  14362     CAGCACCAATCCTACCTCCATCGCTAACCCCACTAAAACACTCACCAAGACCTCAACCCC  14421
                  | ||| || || || || || ||||| |||||||||||||| ||||||||||||| |||
Sbjct  94569143  TAACACTAACCCCACTTCTATTGCTAATCCCACTAAAACACTTACCAAGACCTCAATCCC  94569084

Query  14422     TGACCCCCATGCCTCAGGATACTCCTCAATAGCCATCGCTGTAGTATATCCAAAGACA  14479
                 ||||||||||||||||||||||||||||||||| || |||||||| || ||||| |||
Sbjct  94569083  TGACCCCCATGCCTCAGGATACTCCTCAATAGCTATTGCTGTAGTGTACCCAAAAACA  94569026
```

Table I.1.2 Detailed description of the updated NUMT.

Database information	Homo sapiens chromosome 1, GRCh37 primary reference assembly/ ref|NC_000001.10|/ NC_000001
mtDNA homology region	3911-9755
Chr. 1 region (NUMT)	564461-570304
% homology	98,52890866
Score	11122
Total length (mtDNA / NUMT alignment)	5846
nDNA length (Chr. 1)	5844
mtDNA length	5845
Total of conserved positions	5760
Total of identical positions	5760
Not conserved positions (NUMT)	564463, 564598, 564654, 564862, 564868, 565006, 565286, 565319, 565406, 565454, 565464, 565490, 565508, 565541, 565591, 565697, 565870, 565901, 565937, 565976, 566021, 566024, 566048, 566130, 566371, 566390, 566573, 566771, 566792, 566816, 566849, 566916, 566933, 566960, 567002, 567033, 567062, 567092, 567119, 567191, 567240, 567486, 567489, 567697, 567783, 567807, 567867, 568072, 568201, 568256, 568361, 568419, 568442, 568463, 568572, 568616, 568691, 568703, 568718, 568752, 568941, 569004, 569010, 569052, 569094, 569204, 569226, 569250, 569267, 569409, 569492, 569609, 569624, 569717, 569803, 569874, 569878, 569983, 570076, 570079, 570089, 570094, 570097, 570178.
Not conserved positions (mtDNA)	3913, 4048, 4104, 4312, 4318, 4456, 4736, 4769, 4856, 4904, 4914, 4940, 4958, 4991, 5041, 5147, 5320, 5351, 5387, 5426, 5471, 5474, 5498, 5580, 5821, 5840, 6023, 6221, 6242, 6266, 6299, 6366, 6383, 6410, 6452, 6483, 6512, 6542, 6569, 6641, 6935, 6938,

	7146, 7232, 7256, 7316, 7521, 7650, 7705, 7810, 7868, 7891, 7912, 8021, 8065, 8140, 8152, 8167, 8195, 8196, 8203, 8392, 8455, 8461, 8503, 8545, 8655, 8677, 8701, 8718, 8860, 8943, 9060, 9075, 9168, 9254, 9325, 9329, 9434, 9527, 9530, 9540, 9545, 9548, 9629.
Not identical positions (NUMT)	564463, 564598, 564654, 564862, 564868, 565006, 565286, 565319, 565406, 565454, 565464, 565490, 565508, 565541, 565591, 565697, 565870, 565901, 565937, 565976, 566021, 566024, 566048, 566130, 566371, 566390, 566573, 566771, 566792, 566816, 566849, 566916, 566933, 566960, 567002, 567033, 567062, 567092, 567119, 567191, 567240, 567486, 567489, 567697, 567783, 567807, 567867, 568072, 568201, 568256, 568361, 568419, 568442, 568463, 568572, 568616, 568691, 568703, 568718, 568752, 568941, 569004, 569010, 569052, 569094, 569204, 569226, 569250, 569267, 569409, 569492, 569609, 569624, 569717, 569803, 569874, 569878, 569983, 570076, 570079, 570089, 570094, 570097, 570178.
Not identical positions (mtDNA)	3913, 4048, 4104, 4312, 4318, 4456, 4736, 4769, 4856, 4904, 4914, 4940, 4958, 4991, 5041, 5147, 5320, 5351, 5387, 5426, 5471, 5474, 5498, 5580, 5821, 5840, 6023, 6221, 6242, 6266, 6299, 6366, 6383, 6410, 6452, 6483, 6512, 6542, 6569, 6641, 6935, 6938, 7146, 7232, 7256, 7316, 7521, 7650, 7705, 7810, 7868, 7891, 7912, 8021, 8065, 8140, 8152, 8167, 8195, 8196, 8203, 8392, 8455, 8461, 8503, 8545, 8655, 8677, 8701, 8718, 8860, 8943, 9060, 9075, 9168, 9254, 9325, 9329, 9434, 9527, 9530, 9540, 9545, 9548, 9629.
Identical positions (NUMT)	564461, 564462, 564464-564597, 564599-564653, 564655-564861,

	564863-564867, 564869-565005, 565007-565285, 565287-565318, 565320-565405, 565407-565453, 565455-565463, 565465-565489, 565491-565507, 565509-565540, 565542-565590, 565592-565696, 565698-565869, 565871-565900, 565902-565936, 565938-565975, 565977-566020, 566022, 566023, 566025-566047, 566049-566129, 566131-566370, 566372-566389, 566391-566572, 566574-566770, 566772-566791, 566793-566815, 566817-566848, 566850-566915, 566917-566932, 566934-566959, 566961-567001, 567003-567032, 567034-567061, 567063-567091, 567093-567118, 567120-567190, 567192-567239, 567241-567485, 567487, 567488, 567490-567696, 567698-567782, 567784-567806, 567808-567866, 567868-568071, 568073-568200, 568202-568255, 568257-568360, 568362-568418, 568420-568441, 568443-568462, 568464-568571, 568573-568615, 568617-568690, 568692-568702, 568704-568717, 568719-568751, 568753-568940, 568942-569003, 569005-569009, 569011-569051, 569053-569093, 569095-569203, 569205-569225, 569227-569249, 569251-569266, 569268-569408, 569410-569491, 569493-569608, 569610-569623, 569625-569716, 569718-569802, 569804-569873, 569875-569877, 569879-569982, 569984-570075, 570077, 570078, 570080-570088, 570090-570093, 570095, 570096, 570098-570177, 570179-570304
Identical positions (mtDNA)	3911, 3912, 3914-4047, 4049-4103, 4105-4311, 4313-4317, 4319-4455, 4457-4735, 4737-4768, 4770-4855, 4857-4903, 4905-4913, 4915-4939, 4941-4957, 4959-4990, 4992-5040,

	5042-5146, 5148-5319, 5321-5350, 5352-5386, 5388-5425, 5427-5470, 5472, 5473, 5475-5497, 5499-5579, 5581-5820, 5822-5839, 5841-6022, 6024-6220, 6222-6241, 6243-6265, 6267-6298, 6300-6365, 6367-6382, 6384-6409, 6411-6451, 6453-6482, 6484-6511, 6513-6541, 6543-6568, 6570-6640, 6642-6934, 6936, 6937, 6939-7145, 7147-7231, 7233-7255, 7257-7315, 7317-7520, 7522-7649, 7651-7704, 7706-7809, 7811-7867, 7869-7890, 7892-7911, 7913-8020, 8022-8064, 8066-8139, 8141-8151, 8153-8166, 8168-8194, 8197-8202, 8204-8391, 8393-8454, 8456-8460, 8462-8502, 8504-8544, 8546-8654, 8656-8676, 8678-8700, 8702-8717, 8719-8859, 8861-8942, 8944-9059, 9061-9074, 9076-9167, 9169-9253, 9255-9324, 9326-9328, 9330-9433, 9435-9526, 9528, 9529, 9531-9539, 9541-9544, 9546, 9547, 9549-9628, 9630-9755
Conserved positions (NUMT)	564461, 564462, 564464-564597, 564599-564653, 564655-564861, 564863-564867, 564869-565005, 565007-565285, 565287-565318, 565320-565405, 565407-565453, 565455-565463, 565465-565489, 565491-565507, 565509-565540, 565542-565590, 565592-565696, 565698-565869, 565871-565900, 565902-565936, 565938-565975, 565977-566020, 566022, 566023, 566025-566047, 566049-566129, 566131-566370, 566372-566389, 566391-566572, 566574-566770, 566772-566791, 566793-566815, 566817-566848, 566850-566915, 566917-566932, 566934-566959, 566961-567001, 567003-567032, 567034-567061, 567063-567091, 567093-567118, 567120-567190, 567192-567239, 567241-567485, 567487, 567488, 567490-567696,

	567698-567782, 567784-567806, 567808-567866, 567868-568071, 568073-568200, 568202-568255, 568257-568360, 568362-568418, 568420-568441, 568443-568462, 568464-568571, 568573-568615, 568617-568690, 568692-568702, 568704-568717, 568719-568751, 568753-568940, 568942-569003, 569005-569009, 569011-569051, 569053-569093, 569095-569203, 569205-569225, 569227-569249, 569251-569266, 569268-569408, 569410-569491, 569493-569608, 569610-569623, 569625-569716, 569718-569802, 569804-569873, 569875-569877, 569879-569982, 569984-570075, 570077, 570078, 570080-570088, 570090-570093, 570095, 570096, 570098-570177, 570179-570304
Conserved positions (mtDNA)	3911, 3912, 3914-4047, 4049-4103, 4105-4311, 4313-4317, 4319-4455, 4457-4735, 4737-4768, 4770-4855, 4857-4903, 4905-4913, 4915-4939, 4941-4957, 4959-4990, 4992-5040, 5042-5146, 5148-5319, 5321-5350, 5352-5386, 5388-5425, 5427-5470, 5472, 5473, 5475-5497, 5499-5579, 5581-5820, 5822-5839, 5841-6022, 6024-6220, 6222-6241, 6243-6265, 6267-6298, 6300-6365, 6367-6382, 6384-6409, 6411-6451, 6453-6482, 6484-6511, 6513-6541, 6543-6568, 6570-6640, 6642-6934, 6936, 6937, 6939-7145, 7147-7231, 7233-7255, 7257-7315, 7317-7520, 7522-7649, 7651-7704, 7706-7809, 7811-7867, 7869-7890, 7892-7911, 7913-8020, 8022-8064, 8066-8139, 8141-8151, 8153-8166, 8168-8194, 8197-8202, 8204-8391, 8393-8454, 8456-8460, 8462-8502, 8504-8544, 8546-8654, 8656-8676, 8678-8700, 8702-8717, 8719-8859, 8861-8942, 8944-9059, 9061-9074, 9076-9167, 9169-9253,

	9255-9324, 9326-9328, 9330-9433, 9435-9526, 9528, 9529, 9531-9539, 9541-9544, 9546, 9547, 9549-9628, 9630-9755

Table I.1.3 Alignment of Homo sapiens, five primates and the horse sequences.

BLAST (cutoff E-value of 1e-10) was used to search in GenBank for the 400bp sequences flanking the region Human Chr1:564461-570304 (accession number CM000663). We obtained a common set of sequences from five species, belonging to the same accession numbers in relative close range, no farther apart than 900bp (e.g. Human Chr1:564161-564561 best hit Chimp NC_006468.2 Chr 1: 223983932-223984210; Human Chr1:570204-570604 best hit Chimp NC_006468.2 Chr: 223984204-223984501). We retrieved the sequences corresponding to the combined hit intervals (e.g. Chimp NC_006468.2 Chr 1: 223983932-223984501) and we aligned the human region Chr1:564161-570604 to them using DIALIGN2 (default setting) (Morgenstern 1999). The alignment was further used to infer a maximum likelihood tree of the 5 primates and the horse sequences with PAML (Yang 2007) (baseml, GTR substitution model, 8 Gamma categories for site-specific rates).

```
Human564350-570450   GTCATCCATCCTTTGGCCCCAATACCTAAACTAAGGTCTATGAACAATAAGATGATTTTC
Pan_troglodytes      ............................................................
Pongo_abelii         .......G.................T..........C..........G.....A......
Callithrix_jacchus   ......T.............C.............A.........G....C..G.......
Macaca_mulatta       ......T....................C.C....A.........................
Equus_caballus       .....TTT.........t.tgc.-----.C....A..AG.........G....TG.....

Human564350-570450   TTCAGTGGGACTTTTTTGTTTAATATattagatttgaccttcagcaaggtcaaagggagt
Pan_troglodytes      ..........................----------------------------------
Pongo_abelii         ..........................----------------------------------
Callithrix_jacchus   ..G......................G----------------------------------
Macaca_mulatta       -----------...............----------------------------------
Equus_caballus       C.GGA........GAC.....C...C----------------------------------

Human564350-570450   ccgaactagtctcaggcttcaacatcgaatacgccgcaggccccttcgccctattcttca
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tagccgaatacacaaacattattataataaacaccctcaccactacaatcttcctaggaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   caacatataacgcactctcccctgaactctacacaacatattttgtcaccaagaccctac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ttctgacctccctgttcttatgaattcgaacagcatacccccgattccgctacgaccaac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tcatacacctcctatgaaaaaacttcctaccactcaccctagcattacttatatgatatg
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tctccatacccattacaatctccagcattcccccctcaaacctaagaaatatgtctgataa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aagagttactttgatagagtaaataataggagtttaaatccccttatttctaggactatg
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   agaatcgaacccatccctgagaatccaaaattctccgtgccacctatcacaccccatcct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aaagtaaggtcagctaaataagctatcgggcccataccccgaaaatgttggttatatcct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------
```

```
Human564350-570450   tcccgtactaattaatcccctggcccaacccgtcatctactctaccatctttgcaggcac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   actcatcacagcgctaagctcgcactgattttttacctgagtaggcctagaaataaacat
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   gctagcttttattccagttctaaccaaaaaaataaaccctcgttccacagaagctgccat
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   caagtatttcctcacgcaagcaaccgcatccataatccttctaatagctatcctcttcaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   caatatactctccggacaatgaaccataaccaataccaccaatcaatactcatcattaat
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aatcataatggctatagcaataaaactaggaatagccccctttcacttctgagtcccaga
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ggttacccaaggcacccctctgacatccggcctgctccttctcacatgacaaaaactagc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ccccatctcaatcatataccaaatttctccctcattaaacgtaagccttctcctcactct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ttcaatcttatccatcatggcaggcagttgaggtggattaaaccaaacccaactacgcaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aatcttagcatactcctcaattacccacataggatgaataacagcagttctaccgtacaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ccctaacataaccattcttaatttaactatttatattatcctaactactaccgcattcct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------
```

```
Human564350-570450   actactcaacttaaactccagcaccacaaccctactactatctcgcacctgaaacaagct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aacatgactaacacccttaattccatccaccctcctctccctaggaggcctgcccccgct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aaccggctttttgcccaaatgggccattatcgaagaattcacaaaaaacaatagcctcat
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   catccccaccatcatagccatcatcaccctccttaacctctacttctacctgcgcctaat
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ctactccacctcaatcacactactccctatatctaacaacgtaaaaataaaatgacagtt
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tgaacacacaaaacccaccccattcctccccacactcatcgcccttaccacactgctcct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   acctatctccccttttatgctaataatcttatagaaatttaggttaaatacagaccaaga
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   gccttcaaagccctcagtaagttgcaatacttaatttctgcaacagctaaggactgcaaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   accccactctgcatcaactgaacgcaaatcagccactttaattaagctaagcccttacta
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   gaccaatgggacttaaacccacaaacacttagttaacagctaagcaccctaatcaactgg
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cttcaatctacttctcccgccgccgggaaaaaaggcgggagaagccccggcaggtttgaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------
```

```
Human564350-570450   gctgcttcttcgaatttgcaattcaatatgaaaatcacctcagagctggtaaaaagaggc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ttaacccctgtctttagatttacagtccaatgcttcactcagccattttacctcaccccc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   actgatgttcgccgaccgttgactattctctacaaaccacaaagacattggaacactata
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cctattattcggcgcatgagctggagtcctaggcacagctctaagcctccttattcgagc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cgaactgggccagccaggcaaccttctaggtaacgaccacatctacaacgttatcgtcac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   agcccatgcatttgtaataatcttcttcatagtaatacccatcataatcggaggctttgg
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   caactgactagttcccctaataatcggtgcccccgatatggcgtttccccgcataaacaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cataagcttctgactcttacccccctctctcctactcctgcttgcatctgctatagtgga
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ggccggcgcaggaacaggttgaacagtctaccctcccttggcagggaactactcccaccc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tggagcctccgtagacctaaccatcttctccttacacctagcaggtatctcctctatctt
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aggagccatcaatttcatcacaacaattattaatataaaaccccctgccataacccaata
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------
```

```
Human564350-570450    ccaaacgccccttttcgtctgatccgtcctaatcacagcagtcttacttctcctatctct
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    cccagtcctagccgctggcatcactatactactaacagaccgtaacctcaacaccacctt
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    cttcgacccagccggaggaggagaccccattctataccaacacctattctgatttttcgg
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    tcaccctgaagtttatattctcatcctaccaggcttcggaataatctcccatattgtaac
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    ttactactccgggaaaaaaagaaccatttggatacataggtatggtctgagctatgatat
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    caattggcttcctagggtttatcgtgtgagcacaccatatatttacagtaggaatagacg
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    tagacacacgagcatatttcacctccgctaccataatcatcgctatccccaccggcgtca
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    aagtatttagctgactcgccacactccacggaagcaatatgaaatgatctgctgcagtgc
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    tctgagccctaggatttatttttcttttcaccgtaggtggcctgactggcattgtattag
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    caaactcatcactagacatcgtactacacgacacgtactacgttgtagcccacttccact
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    atgtcctatcaataggagctgtatttgccatcataggaggcttcattcactgatttcccc
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------
```

```
Human564350-570450    tattctcaggctacaccctagaccaaacctacgccaaaatccatttcgctatcatattca
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    tcggcgtaaatctaactttcttcccacaacactttctcggcctatccggaatgccccgac
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    gttactcggactatcccgatgcatacaccacatgaaatatcctatcatctgtaggctcat
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    tcatttctctaacagcagtaatattaataattttcataatttgagaagccttcgcttcga
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    agcgaaaagtcctaatagtagaagaaccctccataaacctggagtgactatatggatgcc
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    ccccaccctaccacacattcgaagaacccgtatacataaaatctagacaaaaaaggaagg
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    aatcgaacccccccaaagctggtttcaagccaaccccatggcctccatgactttttcaaaa
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    agatattagaaaaaccatttcataactttgtcaaagttaaattataggctaaatcctata
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    tatcttaatggcacatgcagcgcaagtaggtctacaagacgctacttcccctatcataga
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    agagcttatcatctttcatgatcacgccctcataatcattttccttatctgcttcctagt
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------

Human564350-570450    cctgtacgcccttttcctaacactcacaacaaaactaactaatactaacatctcagacgc
Pan_troglodytes       ------------------------------------------------------------
Pongo_abelii          ------------------------------------------------------------
Callithrix_jacchus    ------------------------------------------------------------
Macaca_mulatta        ------------------------------------------------------------
Equus_caballus        ------------------------------------------------------------
```

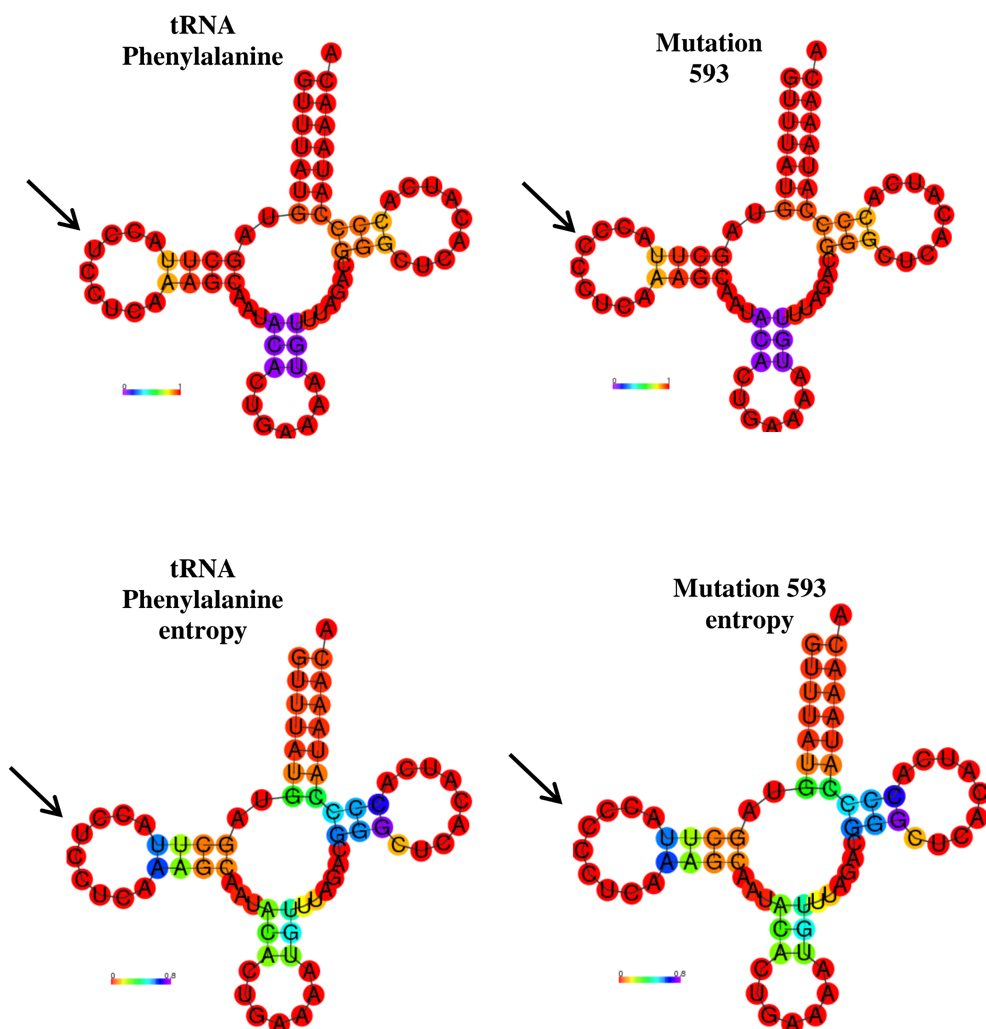
```
Human564350-570450   tcaggaaatagaaaccgtctgaactatcctgcccgccatcatcctagtccttatcgccct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cccatccctacgcatcctttacataacagacgaggtcaacgatccctcctttaccatcaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   atcaattggccatcaatggtactgaacctacgaatacaccgactacggcggactaatctt
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   caactcctacatacttcccccattattcctagaaccaggcgacctgcgactccttgacgt
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tgacaatcgagtagtactcccggttgaagcccccattcgtataataattacatcacaaga
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cgtcttacactcatgagctgtccccacattaggcttaaaaacagatgcaattcccggacg
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tctaaaccaaaccactttcactgctacacgaccaggggtatactacggccaatgctctga
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aatctgtggagcaaaccagttttatgcccatcgtcctagaattaattcccctaaaaatct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ttgaaatagggcccgtatttaccctatagcacccccctctacccccctctagagcccactgt
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aaagctaacttagcattaaccttttaagttaaagattaagagaaccaacacctctttaca
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   gtgaaatgccccaactaaatactaccgtatgacccaccataattacccccatactcctta
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------
```

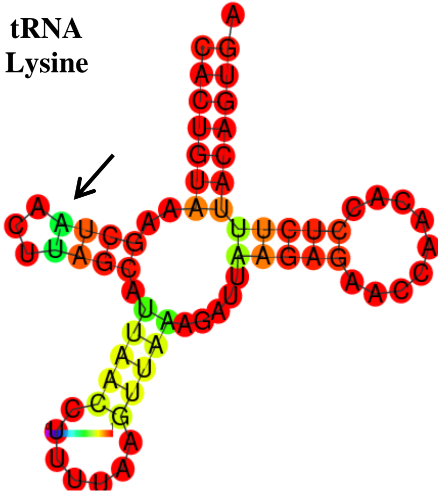
```
Human564350-570450   cactattcctcatcacccaactaaaaatattaaatacaaattaccacctacctccctcac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   caaagcccataaaaataaaaaactataacaaaccctgagaaccaaaatgaacgaaaatct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   gttcacttcattcattgcccccacaatcctaggcctacccgccgcagtactgatcattct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   atttccccctctattgatccccacctccaaatatctcatcaacaaccgactaattaccac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ccaacaatgactaatccaactaacctcaaaacaaatgatagccatacacaacactaaggg
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   acgaacctgatctcttatactagtatccttaatcatttttattgccacaactaacctcct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cggactcctgcctcactcatttacaccaaccacccaactatctataaacctagccatggc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   catccccttatgagcgggcgcagtgattataggctttcgctctaagattaaaaatgccct
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   agcccacttcttaccacaaggcacacctacaccccttatccctatactagttattatcga
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   aaccatcagcctactcattcaaccaatagccctggccgtacgcctaaccgctaacattac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tgcaggccacctactcatgcacctaattggaagcgccacactagcaatatcaactattaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------
```

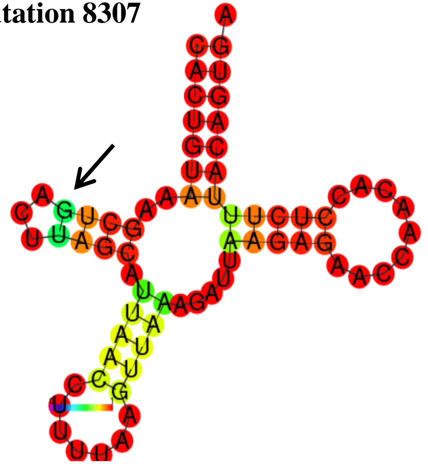
```
Human564350-570450   ccttccctctacacttatcatcttcacaattctaattctactgactatcctagaaatcgc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tgtcgccttaatccaagcctacgtttttacacttctagtaagcctctacctgcacgacaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cacataatgacccaccaatcacatgcctatcatatagtaaaacccagcccatggccccta
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   acaggggccctctcagccctcctaatgacctccggcctagccatgtgatttcacttccac
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tccacaaccctcctcatactaggcctactaaccaacacactaaccatataccaatgatgg
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cgcgatgtaacacgagaaagcacataccaaggccaccacacaccacctgtccagaaaggc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   cttcgatacgggataatcctatttattacctcagaagtttttttcttcgcaggatttttc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   tgagccttttaccactccagcctagctcccaccccccaactagggggacactggccccca
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   acaggcatcaccccgctaaatcccctagaagtcccactcctaaacacatccgtattactc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   gcatcaggggtatcaatcacctgagctcaccatagtctaatagaaaacaaccgaaaccaa
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------

Human564350-570450   ataattcaagcactgcttattacaattttactgggtctctattttaccctcctacaagcc
Pan_troglodytes      ------------------------------------------------------------
Pongo_abelii         ------------------------------------------------------------
Callithrix_jacchus   ------------------------------------------------------------
Macaca_mulatta       ------------------------------------------------------------
Equus_caballus       ------------------------------------------------------------
```

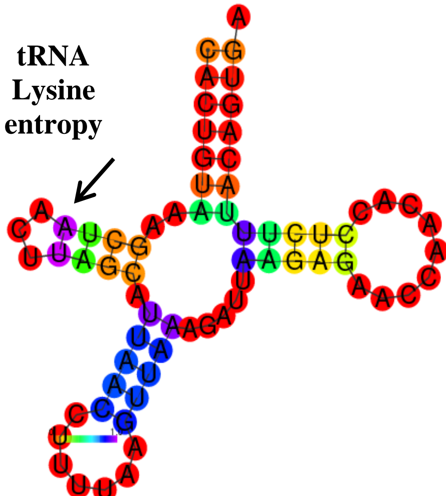
```
Human564350-570450   tcagagtacttcgaggttaaAATA-----TTAGATATTTCCCCTGATACAGGGCTCAATC
Pan_troglodytes      --------------------....-----..............CA...............
Pongo_abelii         --------------------....-----...............................
Callithrix_jacchus   --------------------....AGACA........C.......G......A......A
Macaca_mulatta       --------------------....AGACA....................G..........
Equus_caballus       --------------------.G..AGAtccc....................A.....G..

Human564350-570450   TTTTTCTTTTTAAAGCAATATTTCTCAAAGTAC-----------------TTTTCACAGA
Pan_troglodytes      .................................-----------------..........
Pongo_abelii         .................................-----------------......A...
Callithrix_jacchus   .................T...............a-TTTTTTTCAAAGTAC..........
Macaca_mulatta       .................................acTTTTTTTCAAAGTACA.........
Equus_caballus       ...--.........AT.G....CT..C..A...-----------------...G..TG..

Human564350-570450   ACTTAAGTTTCATTAAGCACTTCACTAAAAGAAAAGTCTGTGATCTAATAAATTTGGAAA
Pan_troglodytes      ............................................................
Pongo_abelii         ..........T....T...................................G.....G..
Callithrix_jacchus   ..........TC...G.T.............C...A........................
Macaca_mulatta       ............C....T..........C....G.A........................
Equus_caballus       ..C......CT..C.......C..Tc.........T.........C.G.G..........

Human564350-570450   ATA
Pan_troglodytes      ...
Pongo_abelii         ...
Callithrix_jacchus   ...
Macaca_mulatta       ...
Equus_caballus       ..t
```

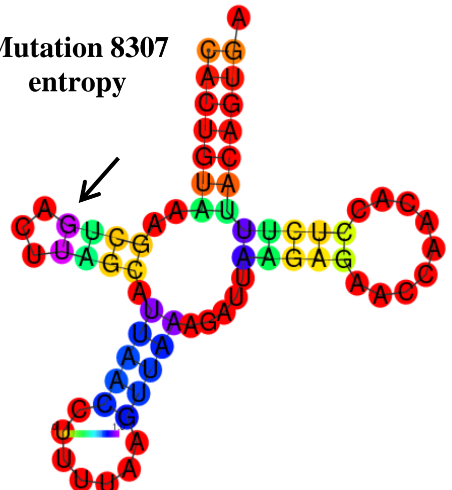
Figure I.3.1. Implication of mtDNA point heteroplasmies in the secondary structure of tRNAs, rRNAs and tertiary structure of COXI, COXII and CytB

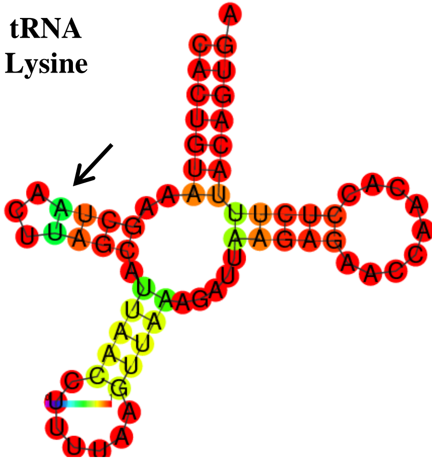


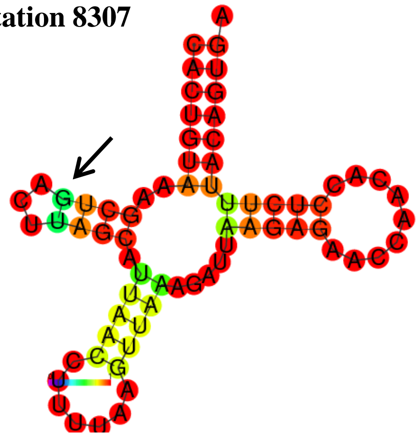

tRNA
Lysine

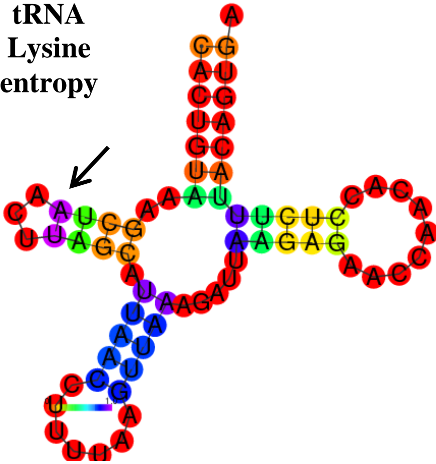

Mutation 8307

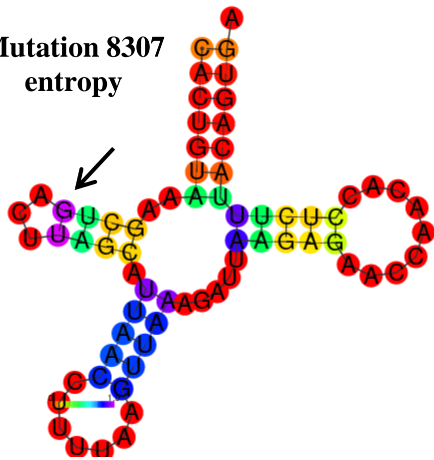

tRNA
Lysine
entropy


Mutation 8307
entropy

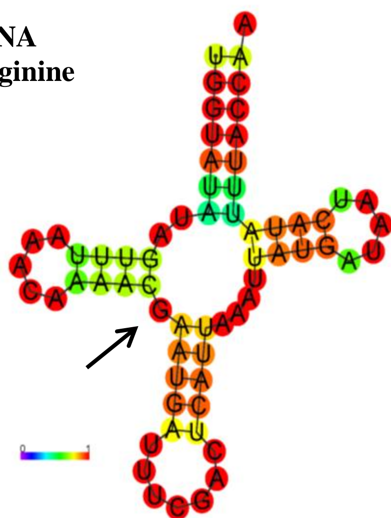

tRNA
Lysine


Mutation 8307

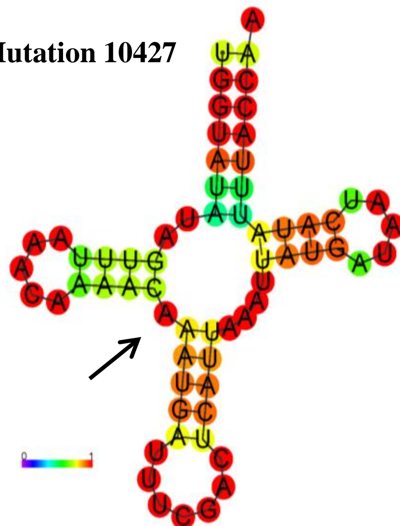

tRNA
Lysine
entropy


Mutation 8307
entropy

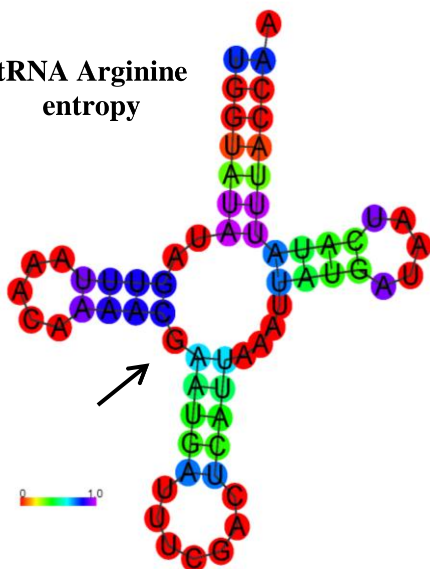
tRNA Arginine



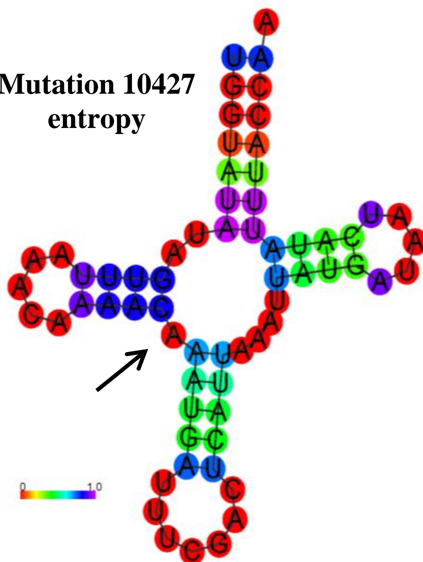
Mutation 10427

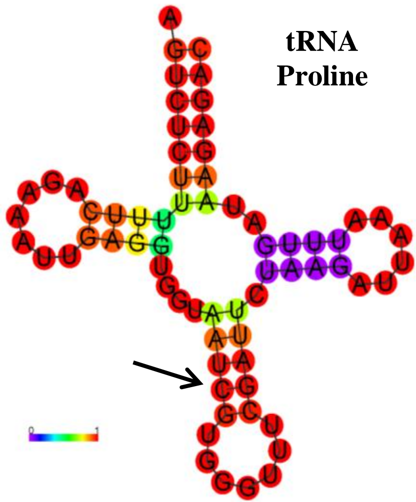


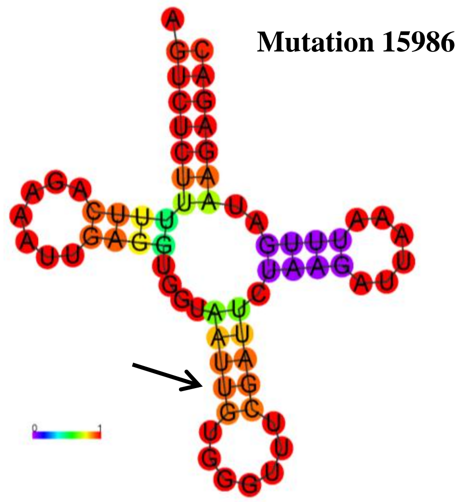
tRNA Arginine entropy

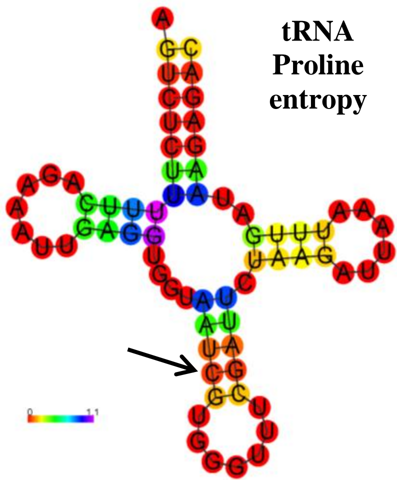


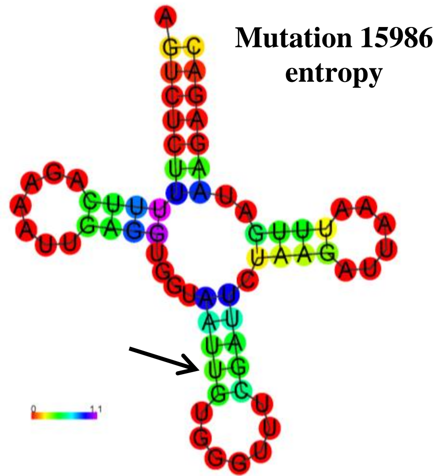
Mutation 10427 entropy

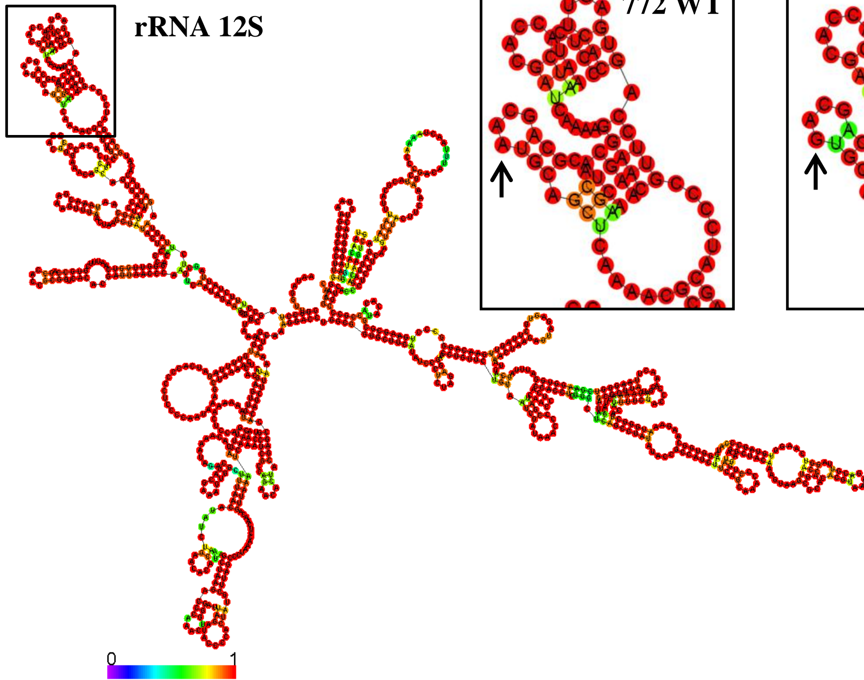



tRNA
Proline

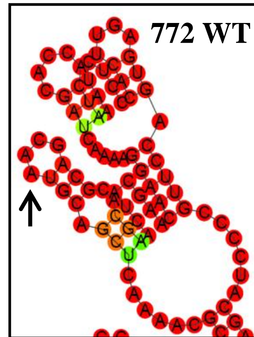

Mutation 15986


tRNA
Proline
entropy

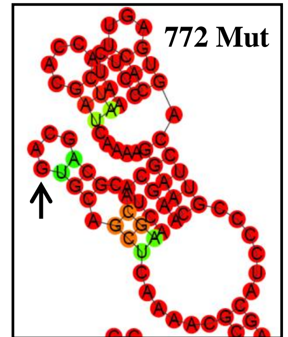

Mutation 15986
entropy



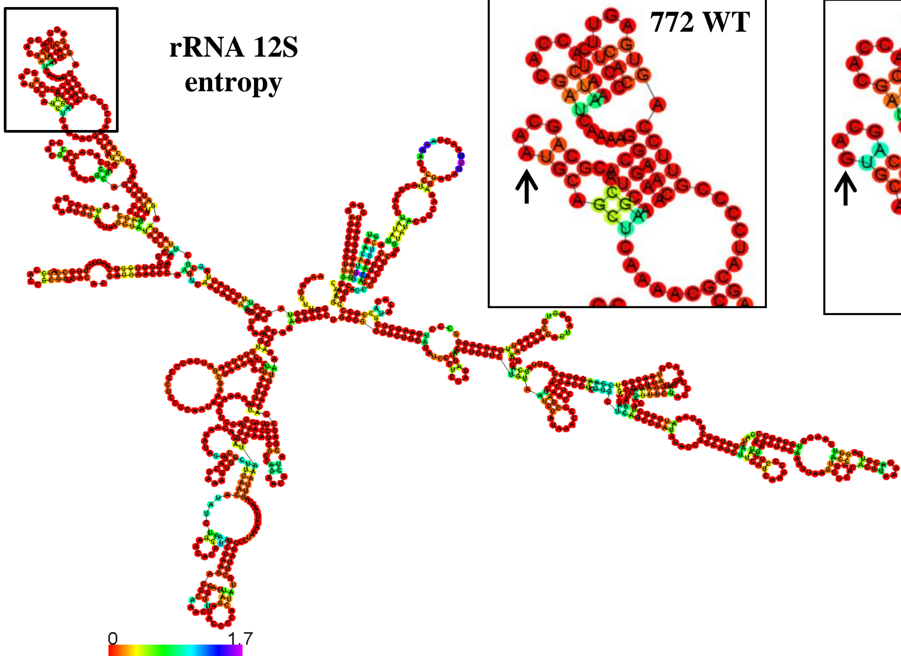
rRNA 12S
0
1



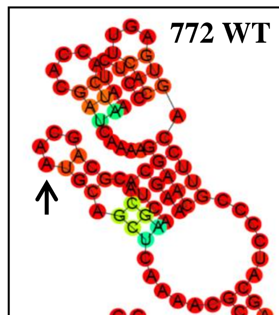
772 WT



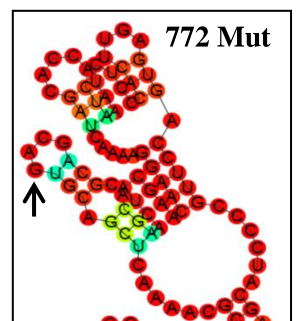
772 Mut



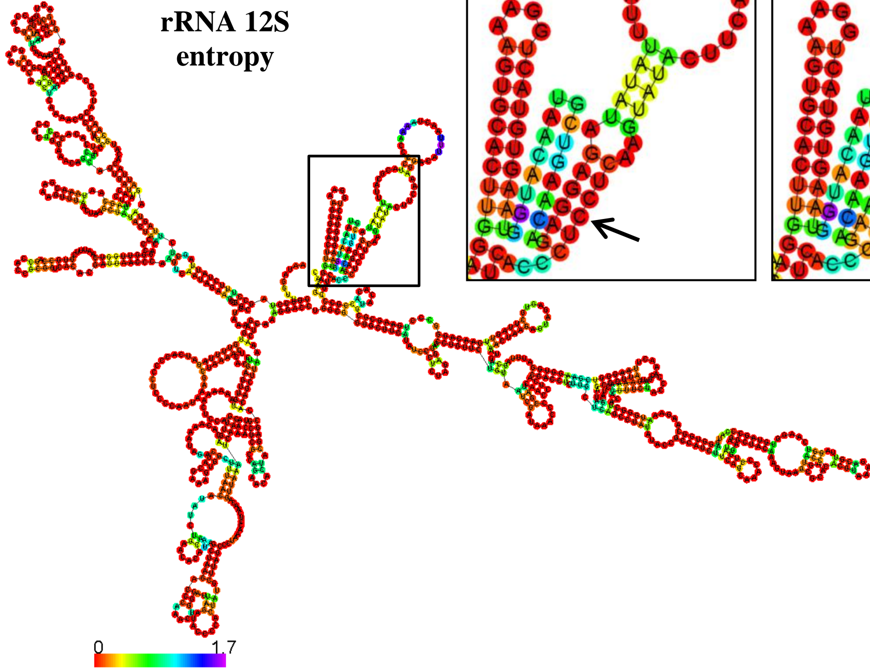
rRNA 12S
entropy
0
1.7

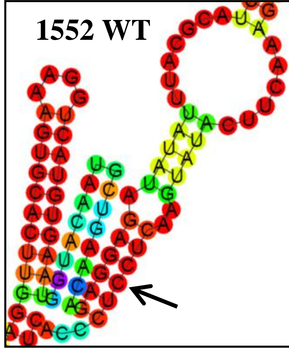


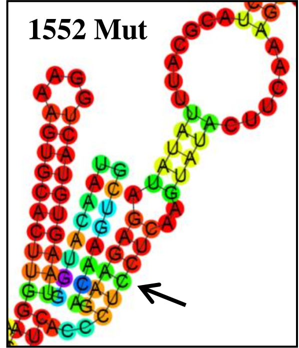
772 WT

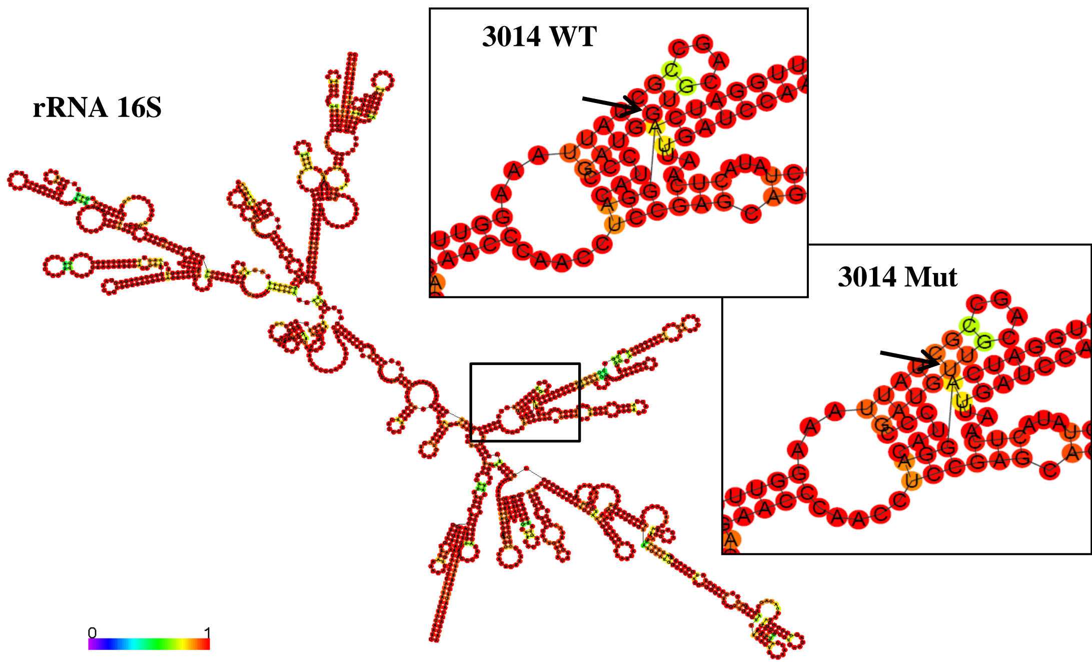


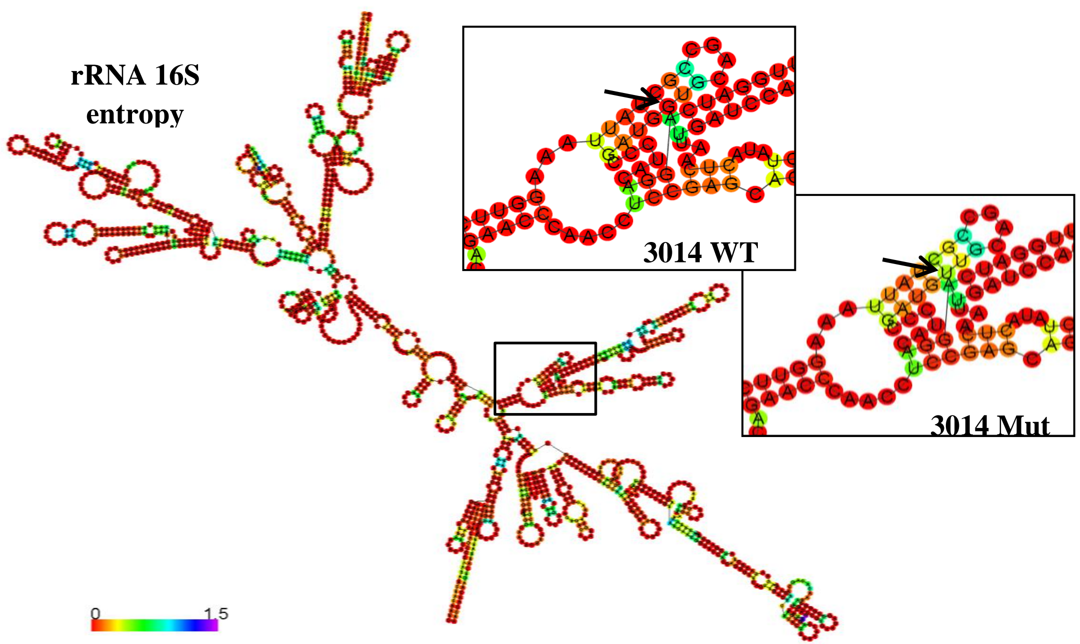
772 Mut

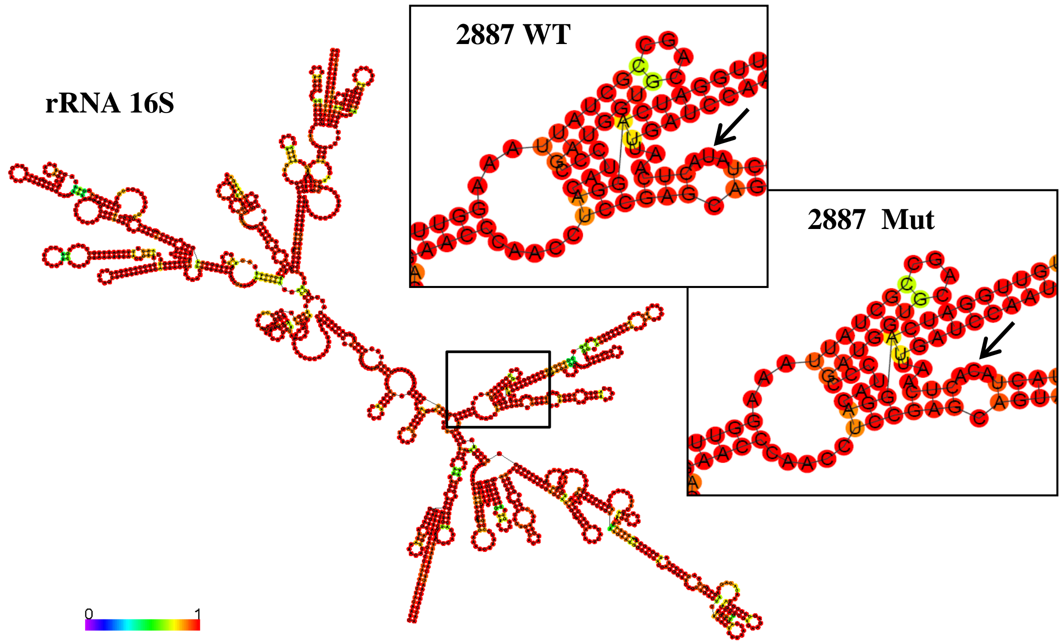

rRNA 12S
entropy
0
1.7

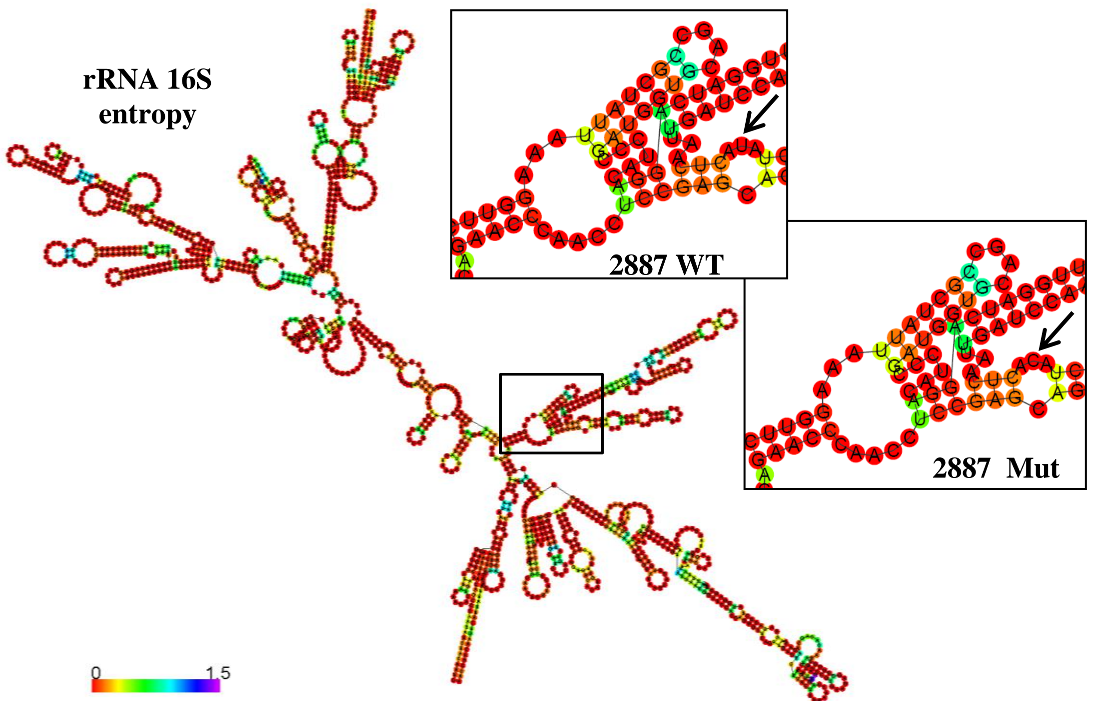

1552 WT

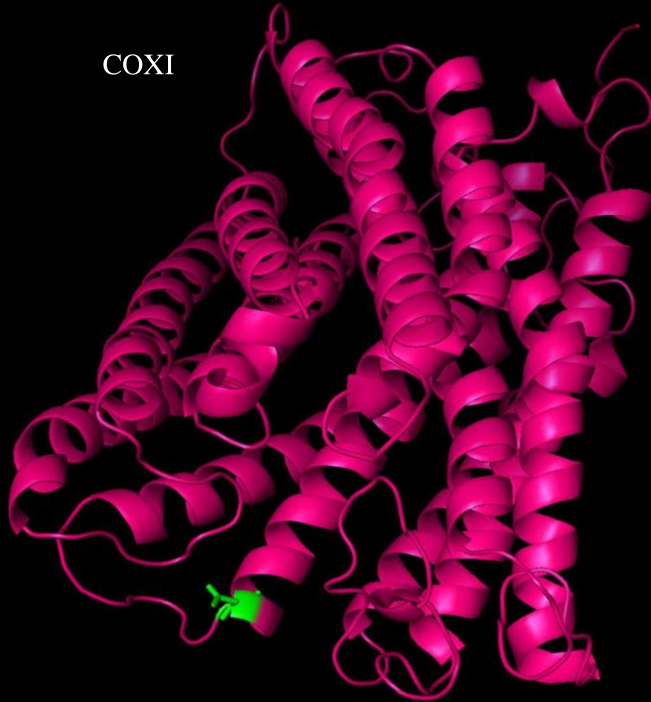

1552 Mut

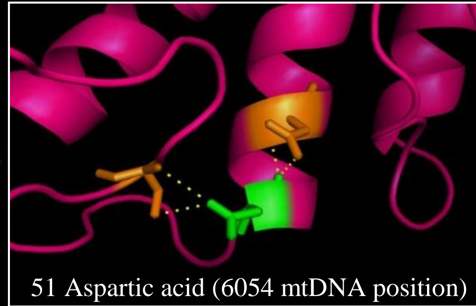

rRNA 16S
3014 WT
3014 Mut
0
1

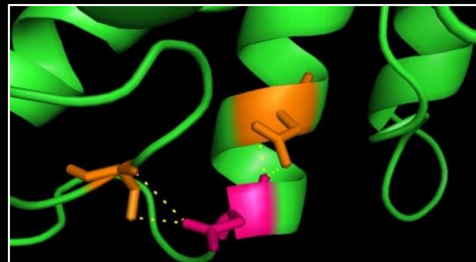

rRNA 16S
entropy
3014 WT
3014 Mut
0
1.5

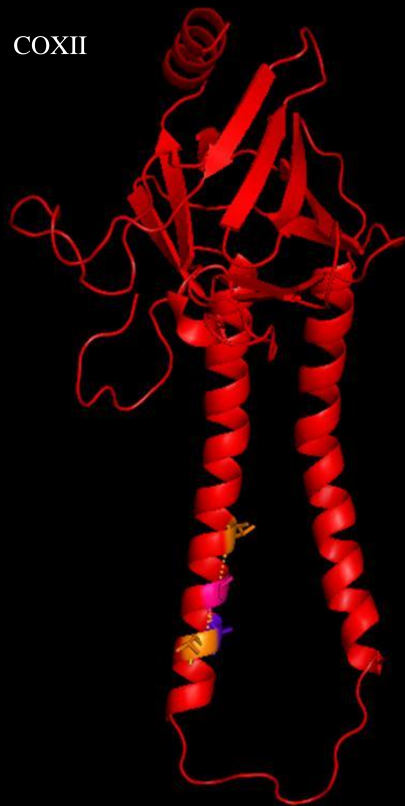

rRNA 16S
2887 WT
2887 Mut
0
1

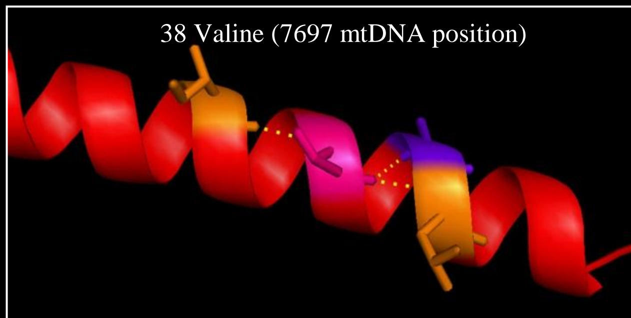

rRNA 16S
entropy
2887 WT
2887 Mut
0
1.5

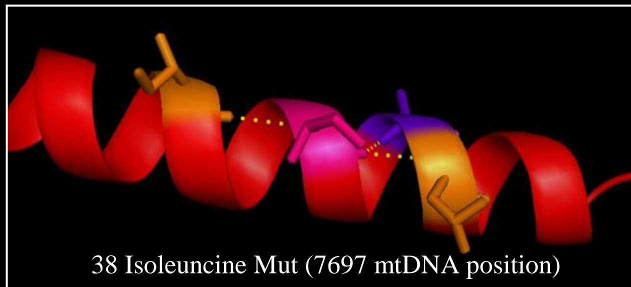

COXI

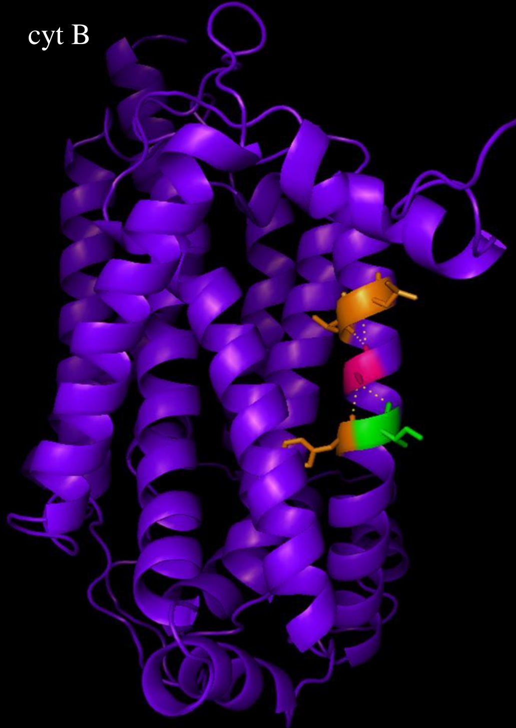

51 Aspartic acid (6054 mtDNA position)

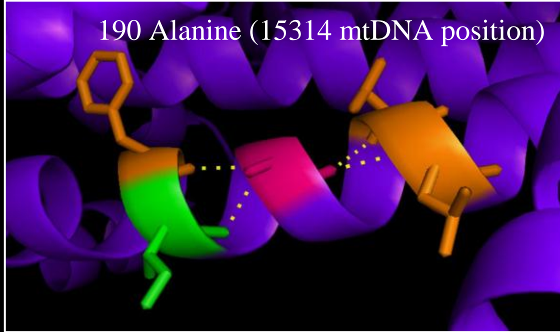

51 Asparagine Mut (6054 mtDNA position)


COXII


38 Valine (7697 mtDNA position)


38 Isoleuncine Mut (7697 mtDNA position)


cyt B


190 Alanine (15314 mtDNA position)

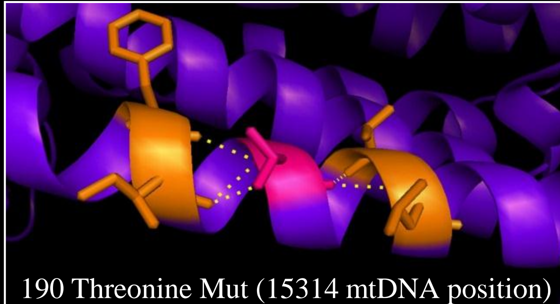

190 Threonine Mut (15314 mtDNA position)

Table I.3.2 Mutation report of complete mitochondrial DNA of 101 individuals. Point and length heteroplasmy are showed in bold. Sex and haplogroup is also reported.

Sample name	Sex	Haplogroup	Haplotype
Z19	male	J	73, 185, 228, 263, 295, 310, 310.1C, 462, 482, 489, ***Het Poly-C* 568-573**, 750, 1438, 2706, 3010, 3107d, 3394, 4216, 4769, 7028, 8860, 10398, 11251, 11719, 12612, 13708, 14766, 14798, 15326, 15452A, 16069, 16126
Z237	female	K	73, 263, ***Het Poly-C* 303-309**, 750, 761, 1438, 3107d, 4769, 6776, 8860, 15326, 16519
Snb85	male	H1	263, 315.1C, 750, 1438, 3010, 3107d, 4769, 6497, 7202, 8860, 15326, 16519
Z225	female	U	73, 94, 150, 263, 315.1C, 750, 1438, 1721, 2706, 3107d, 3197, 3221, 4639, 4769, **Het 5306Y**, 8860, 9097, 9477, 11467, 11653, 11719, 12308, 12372, 12634, 13617, 13630, 13637, 14182, 14766, 15326, 16148, 16261, 16390
Z153	female	pre*V	72, 195, 263, ***Het Poly-C* 303-309**, 750, 1438, 2706, 3107d, 4769, 7028, 8860, 15110, 15326, 16298, 16301
Z8	male	J1b	73, 263, 295, 315.1C, 462, 489, 750, 1438, 1733, 2706, 3010, 3107d, 4216, 4769, 6719, 7028, 8269, 8860, 10398, 11251, 11719, 12612, 13708, 14766, 14927, 15326, 15452A, 16069, 16126, 16145, 16222, 16261
Z326	male	H3	146, 263, 315.1C, 572.2C, 750, 1438, 3107d, 4769, 6776, 8860, 12957, 13359, 15326, 16519
Z349	male	U5	152, 195, 263, 315.1C, 750, 1438, 3010, 3107d, 4769, 4820, 8596, 8860, 9986, 11002, 12061, 15326, 16519

Sample name	Sex	Haplogroup	Haplotype
Z115	male	H*	263, 292, ***Het Poly-C* 303-309**, 750, 1438, 3107d, 4769, 8860, 15326, 16218, 16328A, 16362
Z218	female	T2	73, 207, 263, ***Het Poly-C* 303-309**, 709, 750, 930, 1438, 1888, 2706, 3107d, 3398, 4216, 4769, 4917, 5147, 7028, 8697, 8860, 9181, 10463, 11251, 11719, 11812, 13368, 14766, 14905, 15326, 15452A, 15607, 15928, 16126, 16294, 16296, 16304, 16519
Z265	female	J1a	73, 150, 152, 195, 215, 263, 295, 310.1C, 315.1C, 319, 489, 513, 750, 1438, 2706, 3107d, 4216, 4769, 5899.1C, 7028, 7476, 7768, 7789, 8860, 10398, 10499, 11251, 11377, 11719, 12612, 13708, 13722, 14133, 14766, 15257, 15326, 15452A, 15596, 16069, 16126, 16145, 16231, 16261
Z249	male	I	73, 199, 204, 250, 263, 315.1C, ***Het Poly-C* 568-573**, 750, 1438, 1719, 2706, 3107d, 4529T, 4769, 7028, 8251, 8519, 8860, 9090, 10034, 10238, 10398, 10819, 11719, 12501, 12705, 13780, 14766, 15043, 15326, 15924, 16129, 16223, 16304, 16391, 16519
Z269	female	J1b1	73, **Het 152Y**, 242, 263, 295, 315.1C, 462, 489, 524.1AC, 750, 1438, 2158, 2706, 3010, 3107d, 4216, 4769, 5460, 7028, 8269, 8557, 8860, 10398, 11251, 11719, 12007, 12612, 13708, 13879, 14766, 15326, 15452A, 16069, 16126, 16145, 16172, 16222, 16261
Z120	female	U5a	73, 150, 263, 315.1C, 750, 1438, 1721, 2706, 3107d, 3197, 4491, 4639, 4732, 4769, 6329, 6366, 7028, 7768, 8838, 8860, 9477, 11467, 11719, 12308, 12372, 12957, 13617, 13637, 13677, 14182, 14305, 14766, 15326, 16114A, 16192, 16270

Sample name	Sex	Haplogroup	Haplotype
Z223	female	H1	195, 263, ***Het Poly-C* 303-309**, 523-524d, 750, 1438, 3010, 3107d, 3745, 4769, 5584C, **Het 8603Y**, 8860, 15326, 16519
Z138	female	H*	146, 263, 315.1C, 750, 1438, 3107d, 4769, 4793, 7028, 7768, 8860, 15326, 15409, 16519
Z62	male	U2	73, 217, 228, 263, 315.1C, 340, 508, 750, 1438, 1811, 2706, 3107d, 3720, 4769, 5390, 5426, 6045, 6152, 7028, 8860, 10876, 11467, 11719, 12308, 12372, 13020, 13734, 14766, 15326, 15907, 16051, 16129C, 16183C, 16193.1C, 16291, 16362, 16519
Z144	female	K	73, 263, 310.1C, 315.1C, 497, ***Length Het Poly AC* 514-525**, 750, 1189, 1438, 1811, 2706, 3107d, 3480, 4769, 6260, 7028, 8860, 9055, 9698, 10398, 10550, 11467, 11719, 11840, 11944, 12308, 12372, 13740, 14167, 14766, 14798, 15326, 16224, 16311, 16519
Z217	female	U5a1	73, **Het 152Y**, 263, ***Het Poly-C* 303-309**, 750, 1438, 2706, 3107d, 3197, 4769, 7028, 8860, 9477, 9667, 11467, 11719, 12308, 12372, 13617, 14766, 14793, 15218, 15326, 16172, 16192, 16256, 16270, 16362, 16399
Z90	female	H1	185, 263, ***Het Poly-C* 303-309**, 750, 1438, **Het 2887Y**, 3010, 3107d, 4659, 4769, **Length het 8289-8290CCCCCTCTA**, 8860, 11215, 12810, 15326, 16209, 16519
Z63	female	H*	64, 263, 315.1C, 750, 1438, 3107d, 4769, 5237, 8860, 15326, 16145, 16519

Sample name	Sex	Haplogroup	Haplotype
Z248	male	U5a1a	73, 150, 263, 315.1C, 750, 1303, 1438, 3107d, 3192, 3197, 3591, 4592, 4769, **Het 6054R**, 7028, 8860, 9477, 11296, 11467, 11719, 11938, 12308, 12372, 12618, 13617, 14766, 14793, 15218, 15326, 16192, 16239, 16256, 16270, 16399
Z80	female	L1b	73, 152, 182, 185T, 189, 195, 247, 263, 315.1C, 357, 523d, 709, 710, 750, 769, 825A, 1018, 1738, 2352, 2706, 2758, 2768, 2885, 3107d, 3308, 3594, 3666, 3693, 4104, 4769, 5036, 5046, 5393, 5655, 6548, 6827, 6989, 7028, 7055, 7146, 7256, 7389, 7521, 7867, 8248, 8468, 8655, 8701, 8860, 9476, 9540, 10398, 10688, 10810, 10873, 11719, 12519, 12705, 13105, 13506, 13650, 13789, 13880A, 14178, 14203, 14560, 14766, 14769, 15115, 15326, 15479, 16126, 16187, 16189, 16223, 16264, 16270, 16278, 16293, 16311, 16362, 16400, 16519
Snb33	female	V	72, 263, ***Het Poly-C* 303-309**, 750, 1438, 2706, 3107d, 4580, 4769, 7028, 8860, 12810, 12840, 15326, 15431, 15904, 16240, 16298
Z213	male	N1b	73, 152, 204, 263, 310.1C, 315.1C, 750, 1438, 1598, 2639, 2706, 3107d, 3921A, 4769, 4904, 4960, 5471, 7028, 8251, 8472, 8836, 8860, 9335, 9947, 10238, 10948G, 11362, 11719, 12501, 12705, 12822, 14766, 15326, 16145, 16176G, 16223, 16390, 16519
Z272	male	J	73, 185, 228, 263, 295, ***Het Poly-C* 303-309**, 462, 482, 489, ***Het Poly-C* 568-573**, 750, 1438, 2706, 3010, 3107d, 3394, 4216, 4769, 7028, 8860, 10398, 11251, 11719, 12612, 13708, 14766, 14798, 15326, 15452A, 16069, 16126
Z64	female	H3	263, 293, ***Het Poly-C* 303-309**, 709, 750, 1438, 3107d, 4769, 6776, 8860, 9521, 12346, 15326, 15530, 16519

Sample name	Sex	Haplogroup	Haplotype
Z156	male	H1	263, 315.1C, 750, 1438, 3010, 3107d, 4769, 5100, 5460, 5746, 8860, 15326, 15928, 16166C, 16519
Z163	female	M1	73, 195, 263, ***Het Poly-C* 303-309**, 489, 524.1AC, 709, 750, 813, 1438, 2706, 3107d, 4769, 6446, 6671, 6680, 7028, 8701, 8860, 9540, 10398, 10400, 10873, 11719, 12403, 12414, 12705, 12950C, 13637, 14110, 14766, 14783, 15043, 15301, 15326, 16093, 16129, 16148, 16183C, ***Het Poly-C* 16184-16193**, 16249, 16311, 16519
Z172	female	pre*V	72, 195, 263, ***Het Poly-C* 303-309**, 750, 1438, 2706, 3107d, 3263, 4769, 5567, 7028, 8860, 10410, 12662, 15326, 16298, 16519
Z188	female	J	73, 185, 188, 228, 263, 295, 315.1C, 462, 489, 750, 1438, 1555, 2706, 2963, 3010, 3107d, 4216, 4769, 7028, 8860, 10398, 11251, 11719, 12612, 13708, 14766, 14798, 15326, 15452A, 16069, 16126, 16519
Z340	male	H3	263, 315.1C, 750, 1438, 3107d, 4769, 6776, 8860, **Het 9029R**, 10754, 12308, 12372, 15326, 16519
Z71	male	U5a1a, U5b	73, 195, 263, 310.1C, ***Het Poly-C* 568-573**, 750, 1438, 2706, 3027, 3107d, 3197, 3552, 4769, 4823, 5583, 7028, 8860, 9477, 11467, 11719, 12308, 12372, 13135, 13617, 14766, 14793, 15218, 15326, 15496, 16145, ***Het Poly-C* 16184-16193**, 16311, 16399
Z141	male	H*	73, **Het 189R**, 195, 198, 204, 207, 263, ***Het Poly-C* 303-309**, 709, 750, 1243, 1438, 2706, 3107d, 3505, 4659, 4769, 5046, 5460, 7028, 7864, 8251, 8860, 8994, 11674, 11719, 11947, 12414, 12705, 14766, 15326, **Het 15496R**, 15884C, 15930, 16223, 16292, 16346, 16519

Sample name	Sex	Haplogroup	Haplotype
Z66	female	T5	41, 73, 150, 263, 310.1C, 315.1C, ***Length Het* 514-525**, 709, 750, 1438, 1888, 2706, 3107d, 4216, 4769, 4917, 7028, 8697, 8860, 10463, 11251, 11719, 11812, 13368, 14178, 14203, 14766, 14905, 15326, 15452A, 15607, 15928, 16126, 16153, 16294, 16519
Z125	female	J1b1	73, 242, 263, 295, 315.1C, 462, 489, 524.1AC, 750, 1438, 2158, 2706, 3010, 3107d, 4216, 4769, 5460, 7028, 8269, 8557, 8860, 10398, 11251, 11719, 12007, 12612, 13708, 13879, 14766, 15326, 15452A, 16069, 16126, 16145, 16172, 16222, 16261
Z99	male	T1	73, 263, ***Het Poly-C* 303-309**, 709, 750, 1438, 1888, 2706, 3107d, 3867, 4216, 4769, 4917, 7028, 8697, 8860, 10376, 10463, 11251, 11719, 12633A, 13368, 14766, 14905, 15326, 15452A, 15607, 15928, 16126, 16163, 16186, 16189d, 16294, 16519
Snb9	male	K	73, 263, ***Het Poly-C* 303-309**, 497, 524.1AC, 750, 1189, 1438, 2706, 3107d, 3480, 4769, 6260, 7028, 8860, 9055, 9698, 10398, 10550, 11299, 11467, 11485, 11719, 11840, 11944, 12308, 12372, 13740, 14167, 14766, 14798, 15326, 16224, 16311, 16519
Z187	female	U5a	73, 150, 152, 263, 315.1C, ***Het Poly-C* 568-573**, 750, 1438, 2706, 3107d, 3197, 4769, 5004, 5656, 7028, 7768, 8860, 9477, 11467, 11719, 12308, 12372, 13617, 14182, 14766, 15326, 16192, 16270, 16519
Z253	female	H1	73, 263, 315.1C, 750, 1438, 3010, 3107d, 4769, 5039, 6365, 8860, 15326, 16162, **Het 16189Y**, 16209, 16519
Z201	female	H*	263, ***Het Poly-C* 303-309**, 750, 1438, 2259, 3107d, 4745, 4769, 7337, 8860, 13326, 13680, 14872, 15326, 16365

Sample name	Sex	Haplogroup	Haplotype
Z298	male	U5	73, 150, 263, 310.1C, 315.1C, 456, 517T, 750, 1438, 1721, 2706, 2755, 3107d, 3197, 4769, 6242A, 7028, 7768, 8860, 11467, 11653, 11719, 12308, 12372, 12634, 13630, 13637, 14182, 14766, 15068, 15326, 15905, 16224, 16270, 16311, 16519
Z273	female	K	73, 263, 310.1C, 315.1C, 497, 524.1AC, 750, 1189, 1438, 1811, 2706, 3107d, 3480, **Het 3729R**, 4769, 6260, 7028, 8860, 9055, 9698, 10398, 10463, 10550, 11299, 11467, 11485, 11719, 11840, 11944, 12308, 12372, 13740, 14167, 14766, 14798, 15326, 16224, 16311, 16519
Snb37	female	U6a1	263, ***Het Poly-C* 303-309**, 750, 1438, 3010, 3107d, 4769, 5665, 8860, 15326, 16519
Snb78	female	H3	263, ***Het Poly-C* 303-309**, 750, 1438, 3107d, 4769, 5498, 6293, 6776, 8860, **Het 13812Y**, 15326, 16519
Z167	male	(pre-HV)1	60.1T, 64, 263, 315.1C, 451, 750, 1438, 2442, 2706, 3107d, 3275A, 3847, 4769, 7028, 8860, 9620, 11969, 13188, 14668, 14766, 14905, 15326, 16126, 16362
Z48	male	U5a1a	73, 152, 182, 185T, 189, 195, 247, 263, 315.1C, 357, 471, 524.1AC, 709, 750, **Het 772R**, 930, 1438, 1700, 1811, 2706, 3107d, 4025, 4164, 4769, 6503, 7028, 8860, 8938, 10343, 11467, 11719, 11812C, 11893, 12127, 12308, 12372, 14766, 14926, 15326, 16051, 16183C, ***Het Poly-C* 16184-16193**, 16234, 16266, 16294, 16519, 16525

Sample name	Sex	Haplogroup	Haplotype
Z18	female	L2b	73, 146, 150, 152, 182, 183, 195, 198, 204, 263, ***Het Poly-C* 303-309**, 750, 769, 1018, 1438, 1442, 1706, 2332, 2358, 2416, 2706, 3107d, 3594, 4104, 4158, 4370, 4767, 4769, 5027, 5331A, 5814, 6614, 6713, 6806, 7028, 7256, 7521, 7624A, 7670C, 8080, 8206, 8387, 8503, 8701, 8860, 9221, 9540, 10115, 10398, 10873, 11719, 11944, 12236, 12373, 12705, 12948, 13590, 13650, 14016, 14059, 14766, 15110, 15217, 15326, 16114A, 16129, 16213, 16223, 16278, 16390
Z87	male	U5b	73, 150, 152, 195, 263, 310, 310.1C, 315.1C, 750, 1438, 2234, 2706, 3107d, 3197, 4769, 5656, 7028, 7768, 8860, 9477, 11467, 11719, 12308, 12372, 13617, 14182, 14766, 15122, 15326, 16172, 16183C, ***Het Poly-C* 16184-16193**, 16270, 16274, 16311, 16325, 16357, 16465, 16519
Snb22	male	pre*V	195, 263, 315.1C, 750, 1438, 2706, 3107d, 4769, 7028, 8860, 9180, 15326, 16153, **Het 16293R**, 16298
Z362	male	J	73, 185, 188, 228, 263, 295, 310.1C, 315.1C, 462, 489, 523d, 524d, 750, 1438, 2706, 3010, 3107d, 4216, 4769, 7028, 8860, 10398, 11251, 11719, 12612, 13708, 14766, 14798, 15326, 15452A, 16069, 16126, 16366, 16390, 16519
Z241	male	H3	**Het 93R**, 263, 315.1C, 750, 930, 1438, 2581, 3107d, 4769, 5471, 6776, 8860, 15326, 16129, 16519
Z176	male	H1	263, 310.1C, 315.1C, 750, 1438, 3010, 3107d, 4769, 4859, **Het 8307R**, 8860, 15326, **Het 15908Y**, 16188A, 16519

Sample name	Sex	Haplogroup	Haplotype
Z83	female	U2	73, 217, 228, 263, 315.1C, 340, 750, 1438, 1811, 2706, 3107d, 3720, 4769, 5390, 5426, 5510T, 6045, 6152, 7028, 8860, 10876, 11467, 11719, 12308, 12372, 13020, 13734, 14766, 15326, 15907, 16051, 16129C, 16183C, 16193.1C, 16291, 16362, 16519
Snb48	male	preHV	73, 146, 263, 315.1C, 750, 1438, 1811, 2217, 2706, 3107d, 3480, 4769, 5231, 7028, 8860, 9055, 9698, 9716, 10550, 11299, 11467, 11719, 11869A, 12308, 12372, 13135, **Het 13590R**, 14037, 14167, 14766, 14798, 15326, 16129, **Het 16189Y**, 16222, 16224, 16270, 16311, 16519
Z256	female	J2	73, 150, 152, 263, ***Het Poly-C* 303-309**, 489, 750, 1438, 2101A, 2706, 3107d, 4216, 4769, 5633, 6216, 6893, 7028, 7476, 7859, 8860, 10172, 10398, 11251, 11719, 13708, 14766, 15257, 15326, 15452A, 15812, 16069, 16126, 16193, 16218, 16278, 16290
Z38	male	H3	204, 263, ***Het Poly-C* 303-309**, **Het 593Y**, 750, 1438, 3107d, 4769, 6776, 8541, 8860, 15326, 16286, 16519
Z104	female	H1	150, 263, 315.1C, 750, 1438, 3010, 3107d, 4769, 5460, 7715, 8512, 8860, 14902, 15326, 16519
Z297	male	U5	263, 315.1C, 523d, 524d, 750, 1438, 3107d, 3992, 4024, 4769, 5004, 5514, 8860, 9011.1CA, 9123, 14365, 14582, 15326, 16311, 16319
Z131	male	T	263, 315.1C, 523d, 524d, 750, 1438, 3107d, 3992, 4024, 4769, 5004, 7581, 8860, 9011.1A, 9123, 12308, 12372, 12618, 14365, 14582, 15326, 15497, 15930

Sample name	Sex	Haplogroup	Haplotype
Z5	male	H*	152, 207, 263, 315.1C, 750, 1438, 2259, 3107d, 4745, 4769, 6221, 8860, 12501, 13680, 14872, 15326
Snb63	male	T3	73, 146, 152, 263, 279, ***Het Poly-C* 303-309**, 709, 750, 1438, 1888, 2706, 3107d, 4216, 4769, 4917, 5187, 6261, 7028, 7873, 8652, 8697, 8860, 10822, 10999, 11251, 11719, 11812, 12952, 13368, 14233, 14766, 14905, 15119, 15326, 15452A, 15607, 15928, 16111, 16126, 16292, 16294, 16519
Z195	male	K	73, 152, 263, 310.1C, 315.1C, ***Length Het Poly AC 514-525***, 750, 1189, 1438, 1811, 2706, 3107d, 3480, 4769, 5913, 7028, 8860, 9055, 9698, 9962, 10196G, 10289, 10398, 10550, 11299, 11467, 11719, 11923, 12308, 12372, 13967, 14167, 14766, 14798, 15257, 15326, 15946, 16093, 16224, 16311, 16319, 16463
Z3	female	U5a	73, 150, 263, 310.1C, 315.1C, 750, 1438, 2706, 3105, 3107d, 3197, 3498, 4769, 5656, 6674, 7028, 7768, 8860, 8944, 9477, 11467, 11719, 12308, 12372, 13617, 14182, 14766, 15326, 15777, 16167, 16192, 16270, 16311, 16356
Z30	female	H1	263, 315.1C, 750, 1438, 3010, 3107d, 3394, 4769, 8512, 8860, 14458, 14902, 15326, 16519
Z229	female	H*	95C, 195, 263, ***Het Poly-C* 303-309**, 750, 961G, 1438, 2626, 3107d, 4769, 8448, **Het 8577R**, 8860, 11179, 13759, 15326, 16311, 16320

Sample name	Sex	Haplogroup	Haplotype
Z59	male	U5	73, 150, 263, 310.1C, 315.1C, 750, 1438, 1721, 2231, 2706, 3107d, 3197, 3861, 4769, 7028, 7768, 8860, 9477, 11467, 11653, 11719, 12308, 12372, 12634, 13617, 13630, 13637, 14182, 14766, 15326, 15497, 15778, 16270, 16292, 16362
Snb6	female	K	73, 146, 263, ***Het Poly-C* 303-309**, 750, 1438, 1811, 2217, 2706, 3107d, 3480, 4769, 5231, 7028, 8860, 9055, 9698, 9716, 10550, 11299, 11467, 11719, 11869A, 13135, 14037, 14167, 14766, 14798, 15326, 16129, 16222, 16224, 16270, 16311, 16519
Z128	male	V	72, 263, ***Het Poly-C* 303-309**, 460, 750, 1438, 2706, 3107d, 4580, 4769, 7028, 8860, 15326, 15904, 16172, 16298, 16519
Z137	female	H3	152, 263, 315.1C, 750, 1438, 3107d, 4769, 6776, 8860, 9011.1A, 14350, 15326, 15885, 16299
Snb13	male	H1	263, ***Het Poly-C* 303-309**, 750, 1438, 3010, 3107d, 4769, 6497, 7202, 8860, 15326, 15452A, 16519
Z49	male	U5a, U5b	73, 150, 195, 263, 315.1C, 750, 1438, 1822, 2706, 3107d, 3197, 3316, 4769, 5656, 5788, 7028, 7385, 7768, 8860, 9477, 10301, 10927, 11467, 11719, 13818, 14182, 14470, 14766, 15326, 16129, 16223, 16304, 16391, 16519
Snb74	male	H1	73, 263, 315.1C, 523d, 524d, 750, 1438, 3010, 3107d, 4769, 6647, 8844, 8860, 9746, 11827, 11863, 12591, 15326, 16519

Sample name	Sex	Haplogroup	Haplotype
Snb61	female	U	73, 150, 263, ***Het Poly-C* 303-309**, 750, 896, 1438, 2706, 3107d, 3197, 4732, 4769, 7028, 7768, 8860, **Het 9449Y**, 9477, **Het 11016R**, 11467, 11719, 12308, 12372, 12972C, 13617, 13637, 14182, 14766, **Het 15244R**, 15326, 15511, 16192, 16311
Z12	male	pre*V	72, 195, 263, ***Het Poly-C* 303-309**, 750, 1438, 2706, 3107d, 4769, 7028, 8860, 12618, 14179, 14766, 15110, 15326, 16298, 16301
Z75	male	J1	73, 146, 263, 295, ***Het Poly-C* 303-309**, 462, 489, 750, 1438, 2706, 3010, 3107d, 4216, 4769, 5501, 7028, 8269, 8277, ***Lenght Het* 8272-8278**, 8990, 10398, 11251, 11719, 12519, 12612, 13708, 14766, 15326, 15452A, 16069, 16126, 16145, 16162, ***Het Poly-C* 16184-16193**, 16261, 16286
Z252	male	H3	263, ***Het Poly-C* 303-309**, 709, 750, 1438, 3107d, 4769, 5585, 6776, 8854, 8860, 9287, 10535, 15326, 16209, 16519
Z22	male	H1	263, 315.1C, 750, 1438, 3010, 3107d, 4769, 6497, 7202, 8860, 9986, 15326, 16468, 16519
Snb38	male	H*	63, 315.1C, 750, 1438, 3107d, 3915, 4769, 8860, 8895A, 12811, 15326, 16129, 16519
Z24	male	R1	146, 263, ***Het Poly-C* 303-309**, 750, 1438, 2706, 3107d, 4769, 5471, 7028, **Het 7912R**, 8380, 8860, 14560, 15326, 16256, 16278, 16311, 16519

Sample name	Sex	Haplogroup	Haplotype
Z247	female	J1	73, 146, 189, **Het 234R**, 242, 263, 295, 315.1C, 462, 489, 750, 1438, 2158, 2706, 3010, 3107d, 4216, 4769, 5460, 7028, 8269, 8557, 8860, 10398, 11251, 11719, 12007, 12612, 13708, 13879, 14470, 14766, 15326, 15452A, 16069, 16093, 16126, 16145, 16172, 16184, 16261
Snb3	female	HV	153, 263, 315.1C, 523d, 524d, 750, 1438, 2158, 2706, 3107d, 4231, 4745, 4769, 7903, 8860, 14470, 14872, 15326, 16519
Z108	female	H1	263, ***Het Poly-C* 303-309**, 750, 1438, 3010, 3107d, 3951, 4769, 8860, 12341, 15323, 15326, 16519
Z330	male	H*	263, ***Het Poly-C* 303-309**, 750, 1438, 3107d, 3915, 4769, 8860, 8895A, 15326, 15924, 16129, 16519
Z203	female	H1	185, 263, 315.1C, 750, 1438, 3010, 3107d, 4659, 4769, 8860, 11215, 12810, 15326, 16209, 16311, 16519
Snb27	female	H3	152, 263, 315.1C, 338, 383, 750, 1438, 3107d, 4769, 6776, 8860, 15326, 16259, 16274, 16292, 16294G, 16519
Z168	male	U	73, 146, 263, 310.1C, 315.1C, 750, 1438, 2706, 3107d, 3348, 4769, 5471, 7028, 7391, 7805, 8860, 11467, 11719, 11914, 12308, 12372, 12501, 12618, 13440, 15326, 15941, 16129, 16172, 16278

Sample name	Sex	Haplogroup	Haplotype
Z34	male	K	73, 195, 263, 315.1C, 497, 524.1AC, 750, 1189, 1438, 1811, 2706, 3107d, 3480, 4769, 5460, 5655, 7028, 7521, 8860, 9055, 9698, 10398, 10550, 11299, 11467, 11719, 11914, 12308, 12372, 14167, 14766, 14798, 15326, 16093, 16189, 16224, 16311, 16519
Z234	female	U5a1	73, 263, 310.1C, 315.1C, 750, 1438, 2706, 3107d, 3197, 4655, 4769, 7028, 8697, 8860, 9477, 11467, 11719, 12308, 12372, 13617, 13827, 13928C, 14766, 14793, 15326, 16114A, 16192, 16256, 16270, 16294, 16526
Z53	female	H1	263, ***Het Poly-C*** **303-309**, 750, 1438, 3010, 3107d, 4769, 6719, 8860, 11016, 15326, 16519
Z251	female	H*	152, 263, 315.1C, 750, 1438, 3107d, 4769, 4793, 5348, **Het 7697R**, 8860, 12702, 15326, 16519
Snb28	male	L1b	73, 152, 182, 185T, 189, 195, 247, 263, 315.1C, 357, 523d, 524d, 709, 710, 750, 769, 825A, 1018, 1738, 2706, 2758, 2768, 2885, 3107d, 3308, 3594, 3666, 3693, 4104, 4769, 5036, 5046, 5393, 5655, 6548, 6827, 6989, 7028, 7055, 7146, 7256, 7389, 7521, 7867, 8248, 8468, 8655, 8701, 8860, 9476, 9540, 10398, 10688, 10810, 10873, 11719, 12519, 12705, 13105, 13506, 13650, 13789, 13880A, 14178, 14203, 14560, 14766, 14769, 15115, 15326, 15479, 16126, 16187, 16189, 16223, 16264, 16270, 16278, 16293, 16311, 16362, 16400, 16519
Z287	male	V	72, 263, ***Het Poly-C*** **303-309**, 750, 1438, 2352, 2706, 3107d, 4580, 4769, 7028, 8860, 11151, 14988, 15326, 15904, 16298, 16399

Sample name	Sex	Haplogroup	Haplotype
Z161	female	H3	263, 293, 315.1C, 750, 1438, 3107d, 4769, 6776, 8860, 12346, 15326, 15530, 16519
Snb19	male	H1	263, 315.1C, 750, 1438, 3010, 3107d, 4769, 6497, 7202, 8764, 8860, 15326, **Het 15986R**, 16468, 16519
Z190	female	H1	152, 263, 315.1C, 750, ***Het Poly-C* 956-965**, 1438, 3010, 3107d, 4769, 6237A, 8860, 15326, 16519
Z224	female	H3	263, ***Het Poly-C* 303-309**, 750, 1438, 3107d, 4769, 6776, 8860, 15326, 16519
Snb5	male	J	73, 185, 188, 263, 295, 315.1C, 482, 489, 750, 1438, 2706, 3010, 3107d, 3394, 4216, 4769, 7028, 7184, 8860, 9180, 10398, 11251, 11719, 12612, 13708, 14766, 14798, 15326, 15452A, 16069, 16126, 16291
Z31	female	V	72, 263, ***Het Poly-C* 303-309**, 750, 1438, 2706, 3107d, **Het 4055Y**, 4580, 4769, 6773, 7028, 8860, 15326, 15904, 15927, 16254, 16298
Z226	female	W	73, 189, 194, 195, 204, 207, 263, ***Het Poly-C* 303-309**, 709, 750, 1243, 1406, 1438, 2706, 3107d, 3505, 4769, 5046, 5460, 7028, 8251, 8860, 8994, 11674, 11719, 11947, 12414, 12705, **Het 12948R**, 13263, 14766, 15326, 15784, 15883, 15884C, 16223, 16292, 16519
Z222	male	X	73, 150, 185, 188, 228, 263, 295, ***Het Poly-C* 303-309**, 462, 489, 523d, 524d, 750, 1438, 2706, 3010, 3107d, 4216, 4769, 7028, 8860, 10398, 10463, 11251, 11719, 12612, 13708, 14766, 14798, 15326, 15452A, 16069, 16126, 16366, 16519

Table I.3.2. Complete results of each heteroplasmic position analyzed (map locus, position, sample name, heteroplasmy type, heteroplasmy origin, mean proportion of height peaks, distribution in population database, number of hits in mtDNA phylogeny (PhyloTree.org) and by Soares et al. (2009), probability of mutation and nucleotide and amino acid Conservation Index).

Map Locus	Position	Sample name	Het	CRS	Origin	Mean proportion heigth peaks	Distribution in population database	No. Hits phylogeny (PhyloTree.org)	No. Hits Soares et al. (2009)
CR[a]	93	Z241	G/a	A	Germinal	84.51G 15.49A	GAP:0.1, A:96.3, G:3.6	11	19
CR[b]	64	Ir28	T/c	C	n.a.	87T 13C	C:96.8, T:3.1, GAP:0.1	10	22
CR[b]	146	Ir17	T/c	T	n.a.	90T 10C	T:81.4, C:18.4, A:0.2, GAP:0.1	66	109
CR[b]	146	G65	T/c	T	n.a.	80T 20C	T:81.4, C:18.4, A:0.2, GAP:0.1	66	109
CR[b]	150	Arm17	C/t	C	n.a.	88C 12T	C:88.2, T:11.7, G:0.1, GAP:0.1	41	63
CR[b]	152	Ir11	T/c	T	n.a.	89T 11C	T:70.3, C:29.6, GAP:0.1	117	157
CR[a]	152	Z217	T/c	T	Germinal	66.74T 33.26C	GAP:0.1, C:29.6, T:70.3	117	157
CR[a]	152	Z269	T/c	T	Germinal	87T 13C	GAP:0.1, C:29.6, T:70.3	117	157
CR[b]	195	G67	T/c	T	n.a.	90T 10C	GAP:0.1, T:73.7; C:25.8, A:0.4	71	98
CR[b]	203	Az5	A/g	G	n.a.	66A 34G	G:99.7, A:0.3, GAP:0.1	2	6
CR[b]	204	Arm17	T/c	T	n.a.	89T 11C	T:93.4, C:6.5, A:0.1, GAP:0.1	23	43
CR[a]	234	Z247	G/a	A	Somatic	83.45G 16.55A	GAP:0.1, A:99.4, G:0.5	8	43
CR[a]	189	Z141	A/g	A	Somatic	85.145T 14.855C	GAP:0.1, A:91, C:1.8, G:7.1	19	31

Map Locus	Position	Sample name	Het	CRS	Origin	Mean proportion heigth peaks	Distribution in population database	No. Hits phylogeny (PhyloTree.org)	No. Hits Soares et al. (2009)
tRNA phe[a]	593	Z38	T/c	T	Germinal	91.77T 8.23C	C:0.6, T:99.4	4	11
12S[a]	772	Z48	A/g	A	Germinal	85.83A 14.17G	A:100	0	0
12S[b]	1552	Ir54	G/a	G	n.a.	87G 13A	G:100	0	0
16S[a]	2887	Z90	T/c	T	Germinal	91.95T 8.05C	T: 99.8, C:0.2	2	2
16S[b]	3014	Az10	G/t	G	n.a.	86G 14T	G:100	0	0
ND1[b]	3492	Arm25	A/c	A	n.a.	65A 35C	A:100	0	1
ND1[b]	3492	G20	A/c	A	n.a.	78A 22C	A:100	0	1
ND1[b]	3532	Az14	A/g	A	n.a.	82A 18G	A:100	0	0
ND1[a]	3729	Z273	A/g	A	Germinal	88.68A 11.32G	A:99.9, G:0.1	1	1
ND1[a]	4055	Z31	T/c	T	Somatic	90.38T 9.62C	T:100	0	0
ND2[b]	4991	Az4	G/a	G	n.a.	65G 35A	G:99.3, A:0.7	4	6
ND2[a]	5306	Z225	C/t	C	Germinal	54.42C 45.58T	C:100	0	0
ND2[b]	5460	Az7	A/g	G	n.a.	68A 32G	G:92.8, A:7.2	22	36
COX1[a]	6054	Z248	G/a	G	Germinal	78.75G 21.25A	G:100	0	0
COX2[a]	7697	Z251	A/g	G	Germinal	75.55A 24.45G	G:99.8, A:0.2	2	1

Map Locus	Position	Sample name	Het	CRS	Origin	Mean proportion heigth peaks	Distribution in population database	No. Hits phylogeny (PhyloTree.org)	No. Hits Soares et al. (2009)
COX2[b]	7754	T186	A/g	G	n.a.	89A 11G	G:100	0	1
COX2[a]	7912	Z24	A/g	G	Germinal	75.64A 24.36G	G:99.9, A:0.1	3	4
ATP6/ATP8[b]	8551	Ir30	C/t	T	n.a.	54C 46T	T:99.9, C:0.1	0	4
ATP6[a]	8603	Z223	T/c	T	Germinal	87.95T 12.05C	T:99.9, C:0.1	1	1
ATP6[a]	8577	Z229	G/a	A	Germinal	59.51G 40.49A	A:99.8, G:0.2	0	3
ATP6[b]	8743	Ir29	G/a	G	n.a.	68G 32A	G:99.9, A:0.1	0	1
ATP6[a]	9029	Z340	G/a	A	n.a.	84.6G 15.4A	A:100	0	0
COX3[a]	9449	Snb61	T/c	C	Germinal	16.26T 83.73C	C:98.6, T:1.4	4	6
ND3[b]	10208	Ir10	C/t	T	n.a.	89C 11T	T:99.9, C:0.1	1	2
tRNA arg[b]	10427	Az39	G/a	G	n.a.	85G 15A	G:99.8, A:0.2	1	2
ND4[a]	11016	Snb61	G/a	G	Germinal	82.82G 17.18A	G:99.3, A:0.7	8	9
ND4[b]	11253	G1	C/t	T	n.a.	76C 24T	T:99.6, C:0.4	2	5

Map Locus	Position	Sample name	Het	CRS	Origin	Mean proportion heigth peaks	Distribution in population database	No. Hits phylogeny (PhyloTree.org)	No. Hits Soares et al. (2009)
ND4[b]	11692	G38	A/c	C	n.a.	76A 24C	C:100	0	1
ND4[b]	11809	G82	T/c	T	n.a.	78T 22C	T:99.8, C:0.2	1	0
ND5[b]	12654	Arm37	A/g	A	n.a.	86A 14G	A:99.7, G:0.3	4	3
ND5[b]	13368	Arm19	G/a	G	n.a.	54G 45A	G:96.5, A:3.5	8	8
ND5[a]	13590	Snb48	G/a	G	Germinal	62.11G 37.89A	G:93.6, A:6.4	6	6
ND5[b]	13604	Ir43	G/a	G	n.a.	90G 10A	G:100	0	0
ND5[a]	13812	Snb78	T/c	T	Germinal	76.63T 23.37C	T:99.9, C:0.1	1	2
ND6[b]	14527	G25	A/g	A	n.a.	89A 11G	A:99.9, G:0.1	1	3
ND6[b]	14561	T9	A/g	A	n.a.	78A 22G	A:100	0	0
CytB[b]	14770	Az14	T/c	C	n.a.	55T 45C	C:99.9, T:0.1	1	1
CytB[b]	15046	T186	G/a	A	n.a.	74G 26A	A:99.9, G:0.1	0	1
CytB[a]	15244	Snb61	G/a	A	Germinal	71.74G 28.26A	A:98.5, G:1.5	7	16

Map Locus	Position	Sample name	Het	CRS	Origin	Mean proportion heigth peaks	Distribution in population database	No. Hits phylogeny (PhyloTree.org)	No. Hits Soares et al. (2009)
CytB[b]	15314	Arm20	G/a	G	n.a.	84G 14A	G:99.3, A:0.7	6	12
CytB[a]	15496	Z141	G/a	A	Somatic	86.21G 17.79A	A:99.9, G:0.1	1	2
tRNA thr[a]	15908	Z176	T/c	T	Germinal	73T 27C	T:99.9, C:0.1	1	2
tRNA pro[a]	15986	Snb19	G/a	G	Germinal	58.5G 41.5A	G:100	0	0
CR[b]	16093	Az49	C/t	T	n.a.	91C 8T	T:98.8, C:1.2	36	79
CR[a]	16189	Snb48	T/c	T	Germinal	91.31T 8.69C	GAP:0.1, A:0.1, C:59.6, T:40, G:0.2	66	90
CR[a]	16189	Z253	T/c	T	Germinal	88.75T 11.25C	GAP:0.1, A:0.1, C:59.6, T:40, G:0.2	66	90
CR[b]	16217	T30	C/t	T	n.a.	91C 9T	T:96.5, C:3.4, GAP:0.1	1	6
CR[b]	16223	G73	T/c	C	n.a.	89T 11C	T:51.9, C:48, GAP:0.1	22*	37
CR[b]	16223	Arm20	T/c	C	n.a.	89T 11C	T:51.9, C:48, GAP:0.1	22	37
CR[a]	16293	Snb22	A/g	A	Germinal	79A 21G	A:96.5, C:0.2, T:3, G:0.3	8	17
CR[b]	16362	G67	T/c	T	n.a.	87T 13C	T:77.6, C:22.3, A:0.1	48	67

[a] Heteroplasmic positions detected in present study

[b] Heteroplasmic positions reported by Li et al. (2010)

Position	Probability of mutation	Nucleotide Conservation Index (%)	Amino acid change	Aa Conservation Index (%)
93	1,78E-03	**A:34.04**, C:6.38, G:29.79, T:29.79	-	-
64	2,06E-03	**C:12.77**, G:8.51, T:78.72	-	-
146	1,02E-02	A:27.66, C:42.55, **T:29.79**	-	-
146	1,02E-02	A:27.66, C:42.55, **T:29.79**	-	-
150	5,90E-03	A:19.15, **C:42.55**, G:27.66, T:10.64	-	-
152	1,47E-02	A:17.03, C:48.93, **T:34.04**	-	-
152	1,47E-02	A:17.03, C:48.93, **T:34.04**	-	-
152	1,47E-02	A:17.03, C:48.93, **T:34.04**	-	-
195	9,17E-03	A:31.91, C:38.30, G:8.51, **T:21.28**	-	-
203	5,62E-04	A:53.19, C:2.13, **G:34.04**, T:10.64	-	-
204	4,03E-03	A:8.51, C:17.02, G:53.19, **T:21.28**	-	-
234	4,03E-03	GAP:8.51, **A:76.60**, C:2.13, G:6.38, T:6.38	-	-
189	2,90E-03	GAP:2.13, **A:57.44**, C:10.64, G:29.79	-	-

Position	Probability of mutation	Nucleotide Conservation Index (%)	Amino acid change	Aa Conservation Index (%)
593	1,03E-03	GAP:85.57, A:4.43, C:3.29, G:0.87, **T:5.84**	-	-
772	0,00E+00	GAP:58.46, **A:17.79**, C:9.8, G:4.83, T:9.13	-	-
1552	0,00E+00	GAP:0.47, A:11.88, C:3.09, **G:73.29**, T:11.28	-	-
3492	9,36E-05	**A:44.77**, C:40.13, G:5.97, T:9.13	NS (Lys->Asn)	A:0.1, F:0.3, G:0.1, H:1.2, I:0.1, **K:6.2**, L:1.5, P:0.1, Q:2.1, R:85.2, S:0.3, W:0.3, Y:2.7
3532	0,00E+00	**A:85.23**, C:1.01, G:13.42, T:0.34	NS (Thr->Ala)	A:5.1, F:0.1, G:0.1, I:31.5, L:1.1, M:33, S:0.1, **T:20.5**, V:8.5
3729	9,36E-05	**A:47.99**, C:33.76, G:4.23, T:14.03	S (Ser)	**S:100**
4055	0,00E+00	GAP:1.07, A:27.52, C:42.42, G:2.82, **T:26.17**	NS (Leu->Pro)	GAP:0.3, A:7.7, C:0.5, D:0.8, E:0.3, F:2.7, G:1.1, H:7.5, I:2.7, K:0.2, **L:16.4**, M:4.4, N:9.5, P:5.1, Q:1.8, R:0.1, S:11.5, T:17.9, V:0.6, W:0.1, Y:8.8
4991	5,62E-04	A:78.86, C:1.74, **G:18.86**, T:0.54	S (Gln)	E:0.1, H:2.5, L:0.2, N:0.1, **Q:97.2**
5306	0,00E+00	A:31.61, **C:51.68**, G:1.54, T:15.17	S (Pro)	A:79, F:0.1, G:0.4, I:0.2, L:0.2, M:0.1, **P:15.9**, S:2.2, T:1.8, V:0.3
5460	3,37E-03	A:62.08, C:3.62, **G:21.88**, T:12.42	NS (Ala->Thr)	**A:5.5**, C:0.3, F:0.1, I:22.8, L:1.8, M:11, N:0.2, P:2.7, S:10.8, T:27.1, V:17.6, Y:0.2
6054	0,00E+00	GAP:0.2, A:0.07, **G:99.66**, T:0.07	NS (Asp->Asn)	**D:99.8**, G:0.1, N:0.1, Y:0.1
7697	9,36E-05	GAP:0.07, A:5.3, C:1.81, **G:92.48**, T:0.34	NS (Val->Ile)	GAP:0.1, G:0.1, I:4.3, **V:95.6**

Position	Probability of mutation	Nucleotide Conservation Index (%)	Amino acid change	Aa Conservation Index (%)
7754	9,36E-05	GAP:0.07, A:5.97, C:0.13, **G:93.83**	NS(Asp->Asn)	GAP:0.1, **D:92.9**, E:2.8, G:0.1, N:3.9, Q:0.1, S:0.1
7912	3,74E-04	GAP:0.07, A:65.91, C:1.34, **G:32.08**, T:0.6	S (Glu)	GAP:0.1, D:0.1, **E:99.7**, K:0.1, Q:0.1, V:0.1
8152	4,68E-04	GAP:0.07, A:52.15, C:20.27, **G:4.83**, T:22.68	S (Pro)	GAP:0.1, A:0.4, H:0.1, L:0.1, N:0.2, **P:95.1**, Q:0.1, S:0.9, T:1.8, V:1.3
8307	0,00E+00	GAP:22.89, **A:63.29**, C:0.47, G:1.95, T:11.41	-	-
8551	3,74E-04	GAP:0.13, A:2.95, C:35.84, G:0.2, **T:60.87**/A:2.01, C:84.77, G:0.54, **T:12.68**	S (His)	A:0.1, F:0.1, **H:21.3**, I:15.7, L:53.6, M:4.3, N:0.1, Q:0.1, S:0.2, T:0.4, V:4.2/GAP:77.P418, **H:4.2**, L:15.9, N:0.8, P:0.7, S:0.5, Y:0.2
8603	9,36E-05	A:0.4, C:3.89, G:0.2, **T:95.5**	NS (Phe->Ser)	GAP:0.1, A:1.6, C:0.1, D:0.1, **F:28.7**, I:5.0, L:55.8, M:4.2, P:0.1, S:0.5, T:1.6, V:1.6, W:0.1, Y:0.4
8577	2,81E-04	GAP:0.4, **A:11.28**, C:44.97, G:3.69, T:39.66	S (Leu)	GAP:0.4, A:0.3, F:0.8, I:64.4, **L:17.4**, M:2.8, N:0.1, Q:0.6, T:1.0, V:12.2
8743	9,36E-05	A:67.05, C:1.74, **G:27.72**, T:3.49	NS (Val->Met)	GAP:0.1, A:21.5, C:0.6, F:0.1, I:8.7, L:2.8, M:16.1, S:2.3, T:43, **V:4.9**
9029	0,00E+00	**A:100**	NS (His->Arg)	D:0.1, **H:99.8**, Q:0.1, Y:0.1
9449	5,62E-04	A:0.2, **C:56.85**, T:42.95	S (Tyr)	C:0.2, F:1.8, H:0.1, I:0.1, L:0.3, M:0.1, W:0.5, **Y:97**
10208	1,87E-04	A:39.06, C:34.03, G:2.48, **T:24.43**	S (Pro)	I:0.1, **P:99.7**, Q:0.1, T:0.1
10427	1,87E-04	A:59.87, C:26.04, **G:1.95**, T:12.15	-	-
11016	8,42E-04	A:47.79, C:40.2, **G:0.74**, T:11.28	NS (Ser->Asn)	A:1.8, C:0.1, D:0.3, E:1.1, F:0.2, G:0.3, H:10.6, I:0.2, K:20.2, L:8.3, M:1.7, N:7.8, P:19.5, Q:6.5, R:0.3, **S:10.6**, T:8.8, V:0.6, Y:1
11253	4,68E-04	GAP:0.07, C:5.03, G:0.07, **T:94.83**	NS(Ile->Thr))	A:0.1, F:0.2, **I:7.8**, L:80.4, M:3, S:2.1, T:3, V:3.4

Position	Probability of mutation	Nucleotide Conservation Index (%)	Amino acid change	Aa Conservation Index (%)
11692	9,36E-05	GAP:0.07, A:48.46, **C:16.91**, G:23.56, T:11.01	S (Gly)	A:0.3, **G:99.5**, S:0.3
11809	0,00E+00	GAP:0.07, A:48.46, C:21.81, G:14.16, **T:15.5**	S (Thr)	A:3.9, C:0.1, F:0.1, G:0.2, I:8.8, K:0.3, L:6.9, M:36.8, N:2.6, P:9.1, S:2, **T:21.8**, V:7.4, W:0.1
12654	2,81E-04	**A:80.87**, C:0.54, G:18.39, T:0.2	S (Trp)	A:0.1, H:0.1, K:0.1, L:0.1, Q:0.8, S:0.6, **W:97.9**, Y:0.4
12948	9,36E-05	GAP:0.81, **A:17.38**, C:44.56, G:3.22, T:34.03	S(Leu)	GAP:1.6, A:9.4, C:0.8, D:0.1, E:1.1, F:0.4, G:0.1, H:1.3, I:4.6, **L:35.5**, M:2.5, N:3, P:6.8, Q:3.5, S:9.4, T:17.9, V:1.4, Y:0.6
13368	7,49E-04	A:30.27, C:42.42, **G:17.65**, T:9.66	S (Gly)	A:0.1, **G:99.9**
13590	5,62E-04	A:16.24, C:52.95, **G:1.14**, T:29.66	S (Leu)	A:0.1, C:0.2, F:73.6, I:0.1, **L:17.1**, M:8.9, T:0.1, V:0.1
13604	0,00E+00	A:0.54, C:6.04, **G:93.42**	NS (Ser->Asn)	G:0.1, N:0.5, **S:95.3**, T:4.1
13812	1,87E-04	A:44.03, C:30.4, G:6.58, **T:18.99**	S (Ala)	**A:19**, C:0.7, F:1.4, G:0.8, I:22.8, L:23.8, M:10.3, N:0.1, Q:0.1, S:2.2, T:13.8, V:4.7, Y:0.3
14527	2,81E-04	**A:65.5**, C:13.15, G:10, T:11.34	S (Gly)	A:3.1, C:0.6, F:0.2, **G:2.1**, H:0.4, I:0.3, L:0.4, M:0.1, N:0.1, P:9.5, Q:0.1, S:72.7, T:8.7, V:1.3, Y:0.3
14561	0,00E+00	**A:77.25**, C:8.79, G:13.76, T:0.2	NS (Asp->Gly)	A:10.2, C:3.3, **D:0.1**, F:8.2, G:4.1, I:3.9, L:12.3, M:11.6, S:3.9, T:0.7, V:41.6
14770	9,36E-05	A:0.27, **C:83.42**, G:2.95, T:13.36	S (Asn)	GAP:0.2, E:0.1, H:93.2, I:0.1, K:0.1, L:0.1, M:0.3, **N:3.2**, P:0.1, Q:2.1, R:0.1, S:0.1, T:0.1, V:0.1, X:0.1, Y:0.2
15046	9,36E-05	**A:77.11**, C:8.12, G:8.05, T:6.71	S (Arg)	A:0.1, L:0.1, **R:99.7**, W:0.1
15244	1,50E-03	**A:27.72**, C:24.77, G:31.74, T:15.77	S (Gly)	**G:99.9**, R:0.1

Position	Probability of mutation	Nucleotide Conservation Index (%)	Amino acid change	Aa Conservation Index (%)
15314	1,12E-03	A:18.66, C:4.56, **G:71.07**, T:5.7	NS (Ala->Thr)	**A:61.4**, F:0.2, I:5.3, L:7.7, M:4, S:3.7, T:9.9, V:7.9
15496	1,87E-04	**A:66.98**, C:10, G:9.66, T:13.36	S (Leu)	GAP:0.1, F:2.6, **L:96.7**, M:0.4, P:0.1, V:0.1, Y:0.1
15908	1,87E-04	GAP:1.81, A:80.94, C:0.2, G:0.27, **T:16.78**	-	-
15986	0,00E+00	GAP:0.13, A:50.34, C:0.13, **G:49.4**	-	-
16093	7,39E-03	GAP:4.3, A:21.3, C:4.3, G:6.4, **T:63.8**	-	-
16189	8,42E-03	GAP:2.13, A:19.15, C:40.42, G:8.51, **T:29.79**	-	-
16189	8,42E-03	GAP:2.13, A:19.15, C:40.42, G:8.51, **T:29.79**	-	-
16217	5,62E-04	GAP:14.89, A:10.63, C:25.53, G:6.38, **T:42.55**	-	-
16223	3,46E-03	GAP:8.51, A:23.40, **C:44.68**, T:23.41	-	-
16223	3,46E-03	GAP:8.51, A:23.40, **C:44.68**, T:23.41	-	-
16293	1,59E-03	**A:51.06**, C:19.15, G:4.26, T:25.53	-	-
16362	6,27E-03	A:25.53, C:19.15, **T:55.32**	-	-

7.2 Annex 2

Part del resultats presentats en aquesta Tesi, han quedat recollits en les següents publicacions científiques:

Ramos, A; Santos, C; Alvarez, L; Nogués, R; Aluja, MP. ***Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification****. Electrophoresis. 2009 May; 30(9):1587-1593*.

Ramos, A; Santos, C; Barbena, E; Mateiu, L; Alvarez, L; Nogues, R; Aluja MP. ***Validated primer set that prevents nuclear DNA sequences of mitochondrial origin coamplification: A revision based on the New Human Genome Reference (GRCh37).*** *Electrophoresis. 2011 Mar;32(6-7):782-3.*

Ramos, A; Barbena, E; Matieu, L; González, MM; Mairal, Q; Lima, M; Montiel, R; Aluja, MP; Santos, C. ***Nuclear insertions of mitochondrial origin: data base updating and usefullness in cancer studies****. Mitochondrion. 2011 Nov;11(6):946-53.*

Amanda Ramos
Cristina Santos
Luis Alvarez
Ramon Nogués
Maria Pilar Aluja

Unitat d'Antropologia Biològica, Departament BABVE, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

Received September 15, 2008
Revised November 11, 2008
Accepted November 13, 2008

Research Article

Human mitochondrial DNA complete amplification and sequencing: A new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification

To date, there are no published primers to amplify the entire mitochondrial DNA (mtDNA) that completely prevent the amplification of nuclear DNA (nDNA) sequences of mitochondrial origin. The main goal of this work was to design, validate and describe a set of primers, to specifically amplify and sequence the complete human mtDNA, allowing the correct interpretation of mtDNA heteroplasmy in healthy and pathological samples. Validation was performed using two different approaches: (i) Basic Local Alignment Search Tool and (ii) amplification using isolated nDNA obtained from sperm cells by differential lyses. During the validation process, two mtDNA regions, with high similarity with nDNA, represent the major problematic areas for primer design. One of these could represent a non-published nuclear DNA sequence of mitochondrial origin. For two of the initially designed fragments, the amplification results reveal PCR artifacts that can be attributed to the poor quality of the DNA. After the validation, nine overlapping primer pairs to perform mtDNA amplification and 22 additional internal primers for mtDNA sequencing were obtained. These primers could be a useful tool in future projects that deal with mtDNA complete sequencing and heteroplasmy detection, since they represent a set of primers that have been tested for the non-amplification of nDNA.

Keywords:
Mitochondrial DNA / Nuclear DNA sequences of mitochondrial origin / Primer design
DOI 10.1002/elps.200800601

1 Introduction

It is now accepted that mitochondria descend from prokaryotic endosymbionts and that inter-organelle DNA transfer has been rearranging genetic material between the nucleus and mitochondria. Moreover, it seems that the transfer of DNA from mitochondria to the nucleus has contributed significantly to the evolution and function of eukaryotic genomes (for a review see Leister [1]). Almost all the present-day mitochondrial DNA (mtDNA) transfers to the nucleus give rise to non-coding sequences, the so-called nuclear DNA sequences of mitochondrial origin (NUMTs). Human NUMT loci are evenly distributed within and among chromosomes [2–4], and up to 9857 bp of the human mtDNA can be found at a single nuclear locus [5]. However, only about a third of all NUMTs present in the human nuclear genome are due to insertions of mitochondrial sequences, the rest being originated as duplications of pre-existing NUMTs [6]. Moreover, a subset of human NUMTs is highly rearranged, comprising sequences derived from different regions of the organellar chromosome that have undergone inversions, deletions and duplications [3–5, 7].

A comparison of published compilations of NUMTs clearly shows significant discrepancies among data [8]. For example, in relatively recent studies, Ricchetti *et al.* [9], Mishmar *et al.* [10] and Hazkani-Covo and Dan Graur [11] identify 211, 247 and 452 NUMTs, respectively. According to Lascaro *et al.* [8] these discrepancies are due both to an unwise

Correspondence: Amanda Ramos, Unitat d'Antropologia Biològica, Departament Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain
E-mail: amanda.ramosr@campus.uab.cat
Fax: +34-935811321

Abbreviations: ABC, ammonium bicarbonate; **BLAST,** Basic Local Alignment Search Tool; **mtDNA,** mitochondrial DNA; **nDNA,** nuclear DNA; **NUMT,** nuclear DNA sequences of mitochondrial origin

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

application of Bioinformatic methods and to the use of a not yet correctly assembled nuclear genome. Thus, to optimize the quantification and location of NUMTs, the authors produced a consensus compilation of Human NUMTs – Reference Human Numt Sequences (RHNumtS)– by applying various bioinformatics and experimental approaches. This allows the identification of 190 human NUMTs that could constitute a valuable tool in the implementation of experimental designs of mtDNA amplification and sequencing.

The human mtDNA has been widely used in population genetics, phylogeographic and phylogenetic studies, because it presents a high copy number *per* cell, a compact organization, is maternally transmitted and provides an easy access to an orthologous set of genes with little or no recombination, with rapid evolution and that are selectively neutral [12, 13]. However, this traditional view has been challenged. Many works have been published pointing out: (i) the possibility of recombination of the mtDNA [14, 15]; (ii) the hypothesis that mtDNA frequency variation is due to natural selection [16–18] and (iii) that mitochondrial DNA heteroplasmy is not an exceptional condition related to mitochondrial disease [19–27].

The majority of population studies based on the mtDNA sequencing have focused on the control region, which constitutes less than 7% of the mitochondrial genome. However, the routinary study of the coding region has been performed for a long time to study the implication of mtDNA in diseases such as cancer [28]. Moreover, in the last years, the rapid development of automated DNA sequencing technology has permitted to study the complete mtDNA genome in a large number of healthy individuals [29–35]. Many of these studies, particularly those that deal with the study of mtDNA in pathological samples, have been severely criticized and a large number of manuscripts addressing errors related to the interpretation of mtDNA results (particularly mtDNA heteroplasmy) have been published [21, 36–42]. One of the critiques reported is that NUMTs are a potential source of contamination when PCR is used to study mtDNA [42]. Although this problem has previously been considered to be muted because of the high copy number of mtDNA over the corresponding nuclear loci, caution is warranted [42–44] since amplification of overlapping NUMTs paralogous to the mitochondrial genome indicates that co-amplification of nuclear mitochondrial pseudogenes is a real problem for accurate sequence interpretation, and particularly for the interpretation of mtDNA heteroplasmy [42, 44]. Parr *et al.* [44], in a study performed with ρ0 cells, suggested that among the factors that determine whether an NUMT will or not co-amplify with mtDNA, there are the region of the mtDNA targeted by the PCR and the number of copies of the NUMT. Moreover, samples of ancient DNA or a tissue with a reduced quantity of mtDNA copy number, in both physiological (sperm) and pathological states, also seem to be important factors that could determine the amplification of NUMTs [44, 45]. Although, Goios and co-workers [44, 45] stated that, in standard sequencing of samples used in a population characterization [29, 30], the amplification of NUMTs does not constitute a real problem (since the mtDNA content of samples is much higher than the content of nuclear DNA (nDNA) and the detection of mtDNA heteroplasmy is not a priority), Yao *et al.* demonstrate that it can occur even when standard samples are used [42].

To date, primers to amplify the entire mtDNA have already been published [31, 34, 46], however, none of the published sets completely prevents the amplification of NUMTs. The main goal of this work was to design, validate and describe a set of primers to specifically amplify and sequence the complete human mtDNA, allowing the correct interpretation of mtDNA heteroplasmy in healthy and pathological samples.

2 Materials and methods

2.1 Primer design and validation

In order to analyze the whole mitochondrial genome, primers previously published [34] were tested and new ones were designed. The primer design was performed using the Lasergene 7.2 software, from the DNASTAR package, following the general criteria of primers design [47]. Particular attention was devoted to adjust, as much as possible, the melting temperatures (T_m) of all the primer pairs.

To discard co-amplifications of nDNA and mtDNA, PCR primers were submitted to the Basic Local Alignment Search Tool (BLAST) available in the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST/) [48]. BLAST finds regions of local similarity between sequences comparing nucleotide sequences with sequence databases and calculating the statistical significance of matches. In this work, the specific tool Basic BLAST, 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). A primer pair was discarded if both primers showed homology (expected number of chance matches in a random model – E-value – lower than 1) inside the same chromosome region.

After BLAST validation, primers were further tested in PCR experiments employing total DNA standard samples (see Section 2.2) and nDNA obtained from sperm cells. The nDNA isolation was performed by sperm differential lyses using the QIAmp DNA Investigator Kit (Qiagen) according to the manufacturer's specifications. To guarantee the complete mtDNA elimination, the initial washing step mentioned in QIAmp DNA Investigator Kit manual was repeated at least four times. Moreover, to ensure the integrity of nDNA, Y-chromosome STRs were amplified using AmpFlSTR® Y-filer™ PCR kit (Applied Biosystems), under conditions recommended by the manufacturer and all the markers were correctly amplified.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Validated primers to amplify the complete mtDNA in nine overlapping fragments. Melting temperatures (T_m) and annealing temperature (T_a) for each pair of primers are also presented

Fragment	Fragment length	Name	Sequence (5′–3′)	Primers pair length (bp)	T_a	T_m
1	1822	14898for	tagccatgcactactcaccaga	22	60	60.3
		151rev	ggatgaggcaggaatcaaagac			
2	1758	16488for[a)]	ctgtatccgacatctggttcct	22	60	60.3
		1677rev[a)]	gtttagctcagagcggtcaagt			
3	2543	1404for[a)]	acttaagggtcgaaggtggatt	22	57	58.4
		3947rev	tcgatgttgaagcctgagacta			
4	3005	3734for	aagtcaccctagccatcattcta	23	61	58.9
		6739rev	gatatcatagctcagaccatacc			
5	2709	6511for	ctgctggcatcactatactacta	23	58	58.9
		9220rev	gattggtgggtcattatgtgttg			
6	1738	8910for[a)]	cttaccacaaggcacacctaca	22	61	60.3
		10648rev	ggcacaatattggctaagaggg			
7	1866	10360for	gtctggcctatgagtgactaca	22	61	60.3
		12226rev	cagttcttgtgagctttctcgg			
8	1853	11977for	ctccctctacatatttaccacaac	24	63	59.3
		13830rev	aagtcctaggaaagtgacagcga	23		60.6
9	1872	13477for[a)]	gcaggaatacctttcctcacag	22	63	60.3
		15349rev[a)]	gtgcaagaataggaggtggagt			

a) Oligonucleotides are from Torroni *et al.*, [34].

2.2 Total DNA extraction, amplification and sequencing of the entire mtDNA

Total DNA from blood samples was extracted using JETQUICK Blood DNA Spin Kit (Genomed) and isoamilic phenol-chloroform [49].

Optimized primers and PCR conditions for the mtDNA amplification are reported in Table 1. The PCR mix for each sample consisted of 50 pmol of each primer, 200 μM of each dNTP, 2 mM of $MgCl_2$, 10 × NH_4-based reaction buffer, 1 U of *Taq* DNA polymerase, and 30 ng of DNA in a final volume of 50 μL. The PCRs were performed in a G-Storm GS1 thermocycler and the amplification program consisted of an initial denaturation step of 5 min, followed by 35 cycles of PCR (1 min at 94°C, 40 s at annealing temperature (T_a) and 2.5 min at 72°C), and a final extension step of 5 min at 72°C. The PCR amplification results were visualized by electrophoresis in agarose gels (2%).

PCR products were purified using the JETQUICK PCR Purification Spin Kit (Genomed) and the mtDNA was fully sequenced in both strands using 62 primers (Table 2). Sequence reactions were carried out using the sequencing kit BigDye Terminator v.3 (Applied Biosystems) according to the manufacturer's specifications and were run in an ABI 3130XL sequencer (Servei de Genòmica, Universitat Autònoma de Barcelona).

2.3 Cloning

PCR products that, in the optimization process, show consistently more than a one size product, were cloned into the pCR®4-TOPO® vector, using the TOPO TA Cloning® Kit for Sequencing (Invitrogen). Ten clones were sequenced using the previously explained methodology.

3 Results

3.1 Primer design and validation

From the primers reported in the literature for the complete mtDNA amplification, we evaluated the 11 primer pairs published by Torroni *et al.* [34] since, compared with other works, the authors define a small number of fragments (see Table 2 in Torroni *et al.* [34]). These primers were tested using BLAST. After this analysis, four pairs of primers that show high similarity with nuclear regions were rejected.

The BLAST analysis allows detecting two major problematic regions for primer design. The first one, located between 3914 and 9074 mtDNA positions, presents a 98% similarity with a region within chromosome 1 (see BLAST results as online Supporting Information data); the second one encompasses the region located between 9582 and 14479 mtDNA positions and it shows a 88% similarity with a region within chromosome 5 (see BLAST results as online Supporting Information data). For these regions new primers were designed and verified until the selective amplification of mtDNA was obtained. The 3914–9074 region (Fig. 1) was split into two fragments (fragments 4 and 5 in Fig. 1); For each primer pair, one of the primers was designed outside the 3914–9074 region, preventing the amplification of the homologue region of chromosome 1. For the 9582–14479 region (Fig. 1), minimum of four fragments with ~2000 bp were necessary to completely amplify it. Thus, the previous strategy could only be applied

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2. Primers designed to sequencing the whole mtDNA after PCR amplification in nine overlapping fragments

Fragment	Name/sequence (5′–3′)				
1	14898for	15416for	15966for		
	tagccatgcactactcaccaga	tacacaatcaaagacgccctc	agtctttaactccaccattag		
	151rev	16281rev	15825rev		
	ggatgaggcaggaatcaaagac	gttggtatcctagtgggtgag	gtgaagtatagtacggatgct		
2	16488for	411for	909for		
	ctgtatccgacatctggttcct	cggtatgcacttttaacagtc	gattaacccaagtcaatagaa		
	1677rev	1159rev	638rev		
	gtttagctcagagcggtcaagt	taagctgtggctcgtagtgt	ggtgatgtgagcccgtctaaa		
3	1404for	2028for	2646for	3239for	
	acttaagggtcgaaggtggatt	gatagaatcttagttcaactt	ggttcagctgtctcttacttt	gcagagcccggtaatcgcata	
	3947rev	3382rev	2801rev	2263rev	
	tcgatgttgaagcctgagacta	ttcgttcggtaagcattagga	taatgcaggtttggtagttta	gggtgtgaggagttcagttat	
4	3734for	4346for	4896for	5468for	5995for
	aagtcaccctagccatcattcta	gaacccatccctgagaatcca	taccaaatctctccctcacta	cacgctactcctacctatctc	cagctctaagcctccttattc
	6739rev	6154rev	5571rev	5017rev	4421rev
	gatatcatagctcagaccatacc	ggaactagtcagttgccaaag	aagtattgcaacttactgagg	gaggagtatgctaagattttg	cttatttagctgaccttactt
5	6511for	7111for	7713for	8311for	
	ctgctggcatcactatactacta	acaccctagaccaaacctacg	tcctaacactcacaacaaaac	tagcattaaccttttaagtta	
	9220rev	8600rev	8000rev	7359rev	
	gattggtgggtcattatgtgttg	agaatgatcagtactgcggcg	caacgtcaaggagtcgcaggt	ctactattaggacttttcgct	
6	8910for	9393for	9874for		
	cttaccacaaggcacacctaca	cgagaaagcacataccaaggc	taatatttcactttacatcca		
	10648rev	10154rev	9647rev		
	ggcacaatattggctaagaggg	ttctatgtagccgttgagttg	agctcaggtgattgatactcc		
7	10360for	10892for	11461for		
	gtctggcctatgagtgactaca	atcaacaacaacctatttagc	actcttaaaactaggcggcta		
	12226rev	11673rev	11163rev		
	cagttcttgtgagctttctcgg	gtttggatgagaatggctgtt	cgggtgatgatagccaaggtg		
8	11977for	12500for	12988for		
	ctccctctacatatttaccacaac	tgtgcctagaccaagaagtta	ctagcagcagcaggcaaatca		
	13830rev	13297rev	12763rev		
	aagtcctaggaaagtgacagcga	ggttgatgccgattgtaacta	cgatgaacagttggaataggt		
9	13477for	13950for	14440for		
	gcaggaatacctttcctcacag	ctatctaggccttcttacgag	atactcctcaatagccatcgc		
	15349rev	14838rev	14325rev		
	gtgcaagaataggaggtggagt	catcatgcggagatgttggat	aactttaatagtgtaggaagc		

for flanking fragments (fragments 6 and 9 in Fig. 1), and the primers used to amplify the two fragments that were placed completely within the region (fragments 7 and 8 in Fig. 1) were designed in low-sequence similarity regions.

After BLAST validation, PCR optimization was performed using three samples of total DNA extracted using JETQUICK Blood DNA Spin Kit (Genomed). Surprisingly, the outcome of the amplifications of fragments between positions 12012–13828 and 3798–6739 resulted in more than one PCR product. However, a new amplification with a different sample isolated with isoamilic phenol-chloroform [49] was carried out and only one PCR fragment (with the expected size) was obtained (Fig. 2). These results allow to hypothesize that: (i) the observed additional bands correspond to deletions in mtDNA; (ii) the amplification of an NUMT not detected by BLAST analysis; (iii) the extraction method influences the DNA quality and by consequence the PCR performance and specificity, resulting in unspecific amplification of nuclear or mtDNA. To test these hypotheses the fragment 12012–13828 was cloned into the pCR®4-TOPO® vector, using the TOPO TA Cloning® Kit for Sequencing (Invitrogen) and ten clones were sequenced. The resulting sequences were tested with BLAST showing a complete homology with mtDNA. A careful analysis reveals that the PCR forward primer hybridized with two close regions of mtDNA: the target region and another one that present several differences with the primer. Therefore, the amplification result is an artifact that can be attributed to the poor quality of DNA. For this reason, the forward primer was redesigned and tested again as previously described.

Finally, to discard co-amplifications of nDNA and mtDNA, which could not be deduced by BLAST analysis, PCR amplifications were performed using isolated nDNA obtained from sperm cells. The PCR primers were tested and no positive amplifications were observed.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

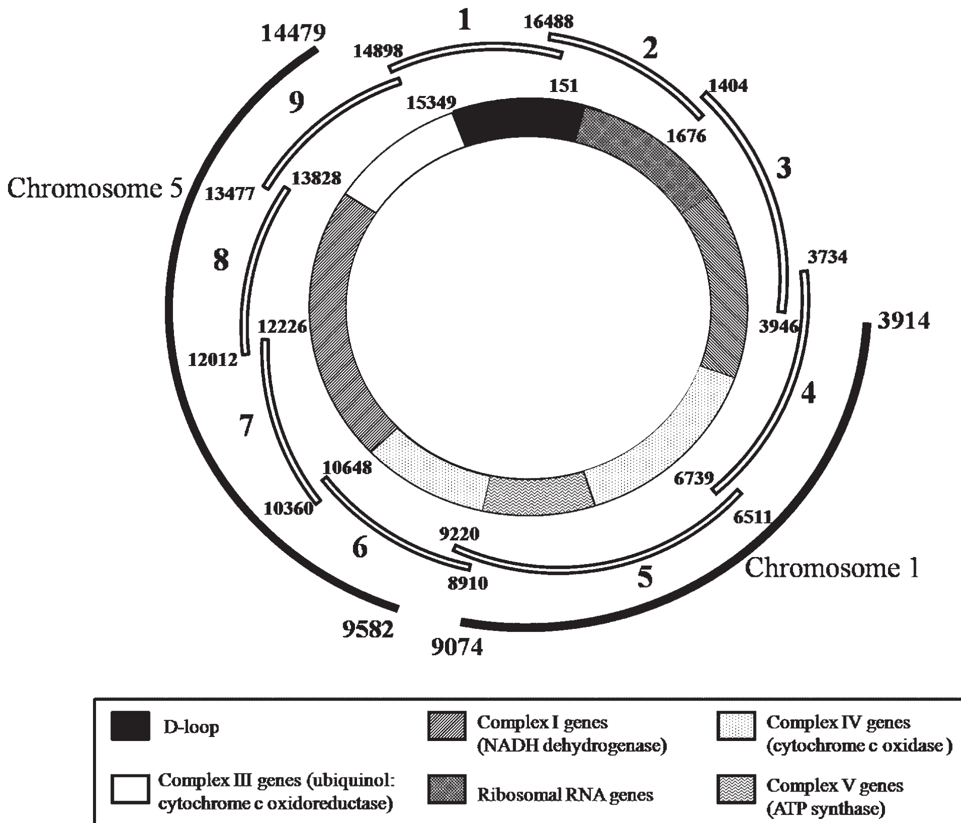


Figure 1. Schematic representation of mtDNA. The nine overlapping fragments defined to PCR amplify the complete mtDNA genome are represented as well as the two nuclear regions with high homology with mtDNA.

3.2 PCR and sequencing primers to mtDNA analysis

For these primers, validated as previously explained, melting temperatures were adjusted as much as possible and the annealing temperatures were obtained by a PCR gradient being selected the highest temperature that shows positive amplification. In Table 1 PCR primers and annealing and melting temperatures to amplify the nine overlapping fragments that cover the entire mtDNA (Fig. 1) are reported. Some of the primers were previously published by Torroni *et al.* [34] and the remaining ones were designed for this study.

Since the size of fragments is of $\sim$2000 bp, internal primers were designed to fully sequence both strands of each fragment. In Table 2 the sequence of the primers to sequence each fragment in both stands are shown. All the primers were tested to sequence three samples and a good efficiency of sequencing was obtained.

4 Discussion

The PCR amplification and sequencing of mtDNA is routinely used in many research fields; however, the interpretation of results, particularly of mtDNA heteroplasmy, could be complicated by the co-amplification of NUMTs [42, 44]. Therefore, an effort to design PCR primers that selectively amplify mtDNA must be performed. In this work we report a set of primers to amplify the whole mtDNA that were validated using the BLAST analysis and by the performance of PCR amplifications using nDNA isolated from sperm cells.

The BLAST search allowed the identification of two problematic regions in primer design (region 3914–9074 of the mtDNA present in chromosome 1 and region 9582–14479 present in chromosome 5). Region 3914–9074 was previously reported [8] as presenting a high sequence similarity with a region within chromosome 1, a result

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

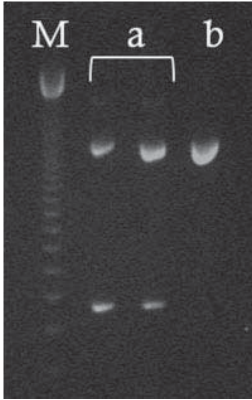


Figure 2. Gel electrophoresis showing amplification results between positions 12012 and 13828 using blood samples extracted with different methods. (a) JETQUICK Blood DNA Spin Kit (Genomed); (b) isoamilic phenol-chloroform.

that is in accordance with our results. However, the presence of an NUMT in chromosome 5, which shows a high identity with the 9582–14479 region of the mtDNA, was not reported in previous studies; accordingly, this region could represent a non-reported NUMT and this deserves further investigation.

According to Goios *et al.* [45] the amplification of NUMTs can only be obtained when mtDNA is almost completely removed from the samples, such as those resulting from preferential semen lyses. Thus, to test the proposed set of primers for the selective amplification of mtDNA, we used a sperm sample that was submitted to DNA extraction using four steps of preferential lyses and for which Y-STRs were successfully amplified and no signs of amplification with the designed primers were detected. This result indicates that the proposed primers specifically amplify mtDNA and we can ensure that no NUMTs will be amplified if this set of nine pair primers is used.

The primers proposed amplification regions of about 2000 bp of mtDNA. As a consequence, in some samples (such as ancient and forensic ones) it would be almost impossible to obtain PCR amplifications with this set of primers. To overcome this problem, we suggest that additional primers within each fragment could be designed using a similar strategy to that applied for fragment 9582–14479, that is, locating primers in regions that show low homology with nDNA.

5 Concluding remarks

In conclusion, we report a set of primers that permit the selective amplification and sequencing of the whole mtDNA in nine overlapping fragments. These primers could be a useful tool in future projects that deal with mtDNA complete sequencing since they represent the first published set of primers tested for the non-amplification of nDNA and therefore suitable for the screening of mtDNA heteroplasmy in both standard and pathological samples.

This work was supported by MEC (project: CGL2006 – 07374) and by the Instituto de estudios zamoranos Florián de Ocampo. We are indebted to all the sample donors and with the personal of the Provincial Hospital of Zamora "Rodríguez Chamorro".

The authors have declared no conflict of interest.

6 References

[1] Leister, D., *Trends Genet.* 2005, *21*, 655–663.

[2] Bensasson, D., Feldman, M. W., Petrov, D. A., *J. Mol. Evol.* 2003, *57*, 343–354.

[3] Moraes, C. T., Srivastava, S., Kirkinezos, I., Oca-Cossio, J., van Waveren, C., Woischnick, M., Diaz, F., *Int. Rev. Neurobiol.* 2002, *53*, 3–23.

[4] Woischnik, M., Moraes, C. T., *Genome Res.* 2002, *12*, 885–893.

[5] Tourmen, Y., Baris, O., Dessen, P., Jacques, C., Malthiery, Y., Reynier, P., *Genomics* 2002, *80*, 71–77.

[6] Hazkani-Covo, E., Sorek, R., Graur, D., *J. Mol. Evol.* 2003, *56*, 169–174.

[7] Blanchard, J. L., Schmidt, G. W., *Mol. Biol. Evol.* 1996, *13*, 537–548.

[8] Lascaro, D., Castellana, S., Gasparre, G., Romeo, G., Saccone, C., Attimonelli, M., *BMC Genomics* 2008, *9*, 267–280.

[9] Ricchetti, M., Tekaia, F., Dujon, B., *PLoS Biol.* 2004, *2*, 1313–1325.

[10] Mishmar, D., Ruiz-Pesini, E., Brandon, M., Wallace, D. C., *Hum. Mutat.* 2004, *23*, 125–133.

[11] Hazkani-Covo, E., Graur, D., *Mol. Biol. Evol.* 2007, *24*, 13–18.

[12] Ballard, J. W. O., Rand, D. M., *Annu. Rev. Ecol. Evol. Syst.* 2005, *36*, 621–642.

[13] Rand, D. M., *Sci. Aging Knowledge Environ.* 2005, re5–13.

[14] Kraytsberg, Y., Schwartz, M., Brown, T. A., Ebralidse, K., Kunz, W. S., Clayton, D. A., Vissing, J. *et al.*, *Science* 2004, *304*, 981.

[15] Zsurka, G., Kraytsberg, Y., Kudina, T., Kornblum, C., Elger, C. E., Khrapko, K., Kunz, W. S., *Nat. Genet.* 2005, *37*, 873–877.

[16] Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., Brandon, M. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 171–176.

[17] Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V., Wallace, D. C., *Science* 2004, *303*, 223–226.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[18] Wallace, D. C., Ruiz-Pesini, E., Mishmar, D., *Cold Spring Harb. Symp. Quant. Biol.* 2003, *68*, 479–486.

[19] Bendall, K. E., Macaulay, V. A., Baker, J. R., Sykes, B. C., *Am. J. Hum. Genet.* 1996, *59*, 1276–1287.

[20] Bendall, K. E., Macaulay, V. A., Sykes, B. C., *Am. J. Hum. Genet.* 1997, *61*, 1303–1308.

[21] Brandstatter, A., Sanger, T., Lutz-Bonengel, S., Parson, W., Beraud-Colomb, E., Wen, B., Kong, Q. P. *et al.*, *Electrophoresis* 2005, *26*, 3414–3429.

[22] Grzybowski, T., Malyarchuk, B. A., Czarny, J., Miscicka-Sliwka, D., Kotzbach, R., *Electrophoresis* 2003, *24*, 1159–1165.

[23] Monnat, Jr. R. J., Loeb, L. A., *Proc. Natl. Acad. Sci. USA* 1985, *82*, 2895–2899.

[24] Santos, C., Montiel, R., Arruda, A., Alvarez, L., Aluja, M. P., Lima, M., *BMC Evol. Biol.* 2008, *8*, 167–179.

[25] Santos, C., Montiel, R., Sierra, B., Bettencourt, C., Fernandez, E., Alvarez, L., Lima, M. *et al.*, *Mol. Biol. Evol.* 2005, *22*, 1490–1505.

[26] Santos, C., Sierra, B., Alvarez, L., Ramos, A., Fernandez, E., Nogues, R., Aluja, M. P., *J. Mol. Evol.* 2008, *67*, 191–200.

[27] Wilson, M. R., Polanskey, D., Replogle, J., DiZinno, J. A., Budowle, B., *Hum. Genet.* 1997, *100*, 167–171.

[28] Santos, C., Martínez, M., Lima, M., Hao, Y.-J., Simoes, N., Montiel, R., *Curr. Top Med., Chem*. 2008, *8*, 1351–1366.

[29] Abu-Amero, K. K., Larruga, J. M., Cabrera, V. M., Gonzalez, A. M., *BMC Evol. Biol.* 2008, *8*, 45–60.

[30] Derenko, M., Malyarchuk, B., Grzybowski, T., Denisova, G., Dambueva, I., Perkova, M., Dorzhu, C. *et al.*, *Am. J. Hum. Genet.* 2007, *81*, 1025–1041.

[31] Maca-Meyer, N., Gonzalez, A. M., Larruga, J. M., Flores, C., Cabrera, V. M., *BMC Genet.* 2001, *2*, 13–21.

[32] Rajkumar, R., Banerjee, J., Gunturi, H. B., Trivedi, R., Kashyap, V. K., *BMC Evol. Biol.* 2005, *5*, 26–34.

[33] Tanaka, M., Cabrera, V. M., Gonzalez, A. M., Larruga, J. M., Takeyasu, T., Fuku, N., Guo, L. J. *et al.*, *Genome Res.* 2004, *14*, 1832–1850.

[34] Torroni, A., Rengo, C., Guida, V., Cruciani, F., Sellitto, D., Coppa, A., Calderon, F. L., Sellitto, D., Coppa, A., Calderon, F. L. *et al.*, *Am. J. Hum. Genet.* 2001, *69*, 1348–1356.

[35] Volodko, N. V., Starikovskaya, E. B., Mazunin, I. O., Eltsov, N. P., Naidenko, P. V., Wallace, D. C., Sukernik, R. I. *Am. J. Hum. Genet.* 2008, *82*, 1084–1100.

[36] Bandelt, H. J., Salas, A., Lutz-Bonengel, S., *Int. J. Legal Med.* 2004, *118*, 267–273.

[37] Davis, J. I., Simmons, M. P., Stevenson, D. W., Wendel, J. F., *Syst. Biol.* 1998, *47*, 282–310.

[38] Hirano, M., Shtilbans, A., Mayeux, R., Davidson, M. M., DiMauro, S., Knowles, J. A., Schon, E. A. *Proc. Natl. Acad. Sci. USA* 1997, *94*, 14894–14899.

[39] Salas, A., Carracedo, A., Macaulay, V., Richards, M., Bandelt, H. J., *Biochem. Biophys. Res. Commun.* 2005, *335*, 891–899.

[40] Salas, A., Quintans, B., Alvarez-Iglesias, V., *Methods Mol. Biol.* 2005, *297*, 197–208.

[41] Yao, Y. G., Bravi, C. M., Bandelt, H. J., *Forensic Sci. Int.* 2004, *141*, 1–6.

[42] Yao, Y. G., Kong, Q. P., Salas, A., Bandelt, H. J., *J. Med. Genet.* 2008, *45*, 769–772.

[43] Parfait, B., Rustin, P., Munnich, A., Rotig, A., *Biochem. Biophys. Res. Commun.* 1998, *247*, 57–59.

[44] Parr, R. L., Maki, J., Reguly, B., Dakubo, G. D., Aguirre, A., Wittock, R., Robinson, K. *et al.*, *BMC Genomics* 2006, *7*, 185–198.

[45] Goios, A., Prieto, L., Amorim, A., Pereira, L., *Int. J. Legal Med.* 2008, *122*, 341–345.

[46] Wong, L. J., Liang, M. H., Kwon, H., Park, J., Bai, R. K., Tan, D. J., *Clin. Chem.* 2002, *48*, 1901–1912.

[47] Roche Diagnostics GmbH, *PCR Applications Manual*, Mannheim 2006.

[48] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., *J. Mol. Biol.* 1990, *215*, 403–410.

[49] Sambrook, J., *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor 2001.

© 2009 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Electrophoresis 2011, *32*, 782–783

Amanda Ramos[1]
Cristina Santos[1]
Elena Barbena[1]
Ligia Mateiu[2]
Luis Alvarez[1,3]
Ramon Nogués[1]
Maria Pilar Aluja[1]

[1]Unitat d'Antropologia Biològica, Departament BABVE, Universitat Autònoma de Barcelona, Barcelona, Spain
[2]Department of Biology, University of Antwerp, Groenenborgerlaan, Antwerpen, Belgium
[3]IPATIMUP, Instituto de Patologia e Imunologia da Universidade do Porto, R. Dr. Roberto Frias, Porto, Portugal

Received November 1, 2010
Revised November 26, 2010
Accepted November 29, 2010

Short Communication

Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: A revision based on the New Human Genome Reference Sequence (GRCh37)

A new human genome reference sequence – GRCh37 – was recently generated and made available by the Genome Reference Consortium. Since the prior disposable human reference sequence – hg18 – was previously used for the mitochondrial DNA primer BLAST validation, a revision of those previously published primer pairs is required. Thus, the aim of this Short Communication is to perform an in silico BLAST test of the published disposable nine primer pairs using the new human reference sequence and to report the pertinent modifications. The new analysis showed that one of the tested primer pairs requires a revision. Therefore, a new validated primer pair, which specifically amplifies the mitochondrial region located between positions 6520 and 9184, is presented.

Keywords:
Mitochondrial DNA / Nuclear sequences of mitochondrial origin / Primer design
DOI 10.1002/elps.201000583

Recently, a paper that deals with the complete amplification of Human mitochondrial DNA (mtDNA) was published in this journal [1]. The main goal of the mentioned work was to design, validate and describe a set of primers, to specifically amplify and sequence the complete human mtDNA, allowing the correct interpretation of mtDNA heteroplasmy in healthy and pathological samples. According to the authors, the interpretation of mtDNA data, particularly the presence of heteroplasmy, could be complicated due to the co-amplification of mtDNA and nuclear sequences of mitochondrial origin (NUMTs) [2, 3]. Therefore, an effort to design PCR primers that selectively amplify mtDNA was necessary. In accordance, Ramos et al. [1] proposed a set of nine primer pairs to amplify the whole mtDNA. Primers were validated, through the Basic Local Alignment Search Tool (BLAST) available in NCBI (National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/BLAST/) [4], and subsequently tested using a nuclear DNA (nDNA) sample.

In Ramos et al. [1], the human reference sequence hg18 was used in the in silico primer validation. However, a new human reference sequence – GRCh37 – was recently (February 2009) generated by the Genome Reference Consortium. Thus, the aim of this work is to test again with BLAST the previously published primers using the new human reference sequence and to report the pertinent modifications.

Accordingly to the latest human reference sequence update, one of the most problematic NUMTs for the primer design appears to be larger than earlier reported. This nuclear sequence is located in chromosome 1 and presents a 98% of homology with one-third of the mtDNA. This redefined NUMT (detailed and described as online Supporting Information S1) lays between positions 564 461 and 570 304 (in relation to the *Homo sapiens* chromosome 1, GRCh37 primary reference assembly – NC_000001), and presents a 98% homology with the region 3911–9755 of human mtDNA [5].

The new in silico analysis revealed that one of the nine previously described primer pairs, specifically the one designed to amplify the 2709-bp region of the mtDNA, located between positions 6511 and 9220 (fragment 5 in Table 1 of Ramos et al. [1]), lies on the redefined homologous region. Thus, an accidental co-amplification of mtDNA and nDNA could occur. The described NUMT

Correspondence: Amanda Ramos, Unitat d'Antropologia Biològica, Departament Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Barcelona, Spain
E-mail: amanda.ramos@uab.cat
Fax: +34-935811321

Abbreviations: BLAST, Basic Local Alignment Search Tool; **mtDNA,** mitochondrial DNA; **nDNA,** nuclear DNA; **NUMT,** nuclear sequences of mitochondrial origin

© 2011 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. New validated primer pair to amplify the mtDNA region affected by the new updated NUMT is reported. Melting temperatures (T_m) are also presented

Fragment	Fragment length	Name	Sequence (5′–3′)	Primers pair length (bp)	T_m
5	2664	6520for	tcactatactactaacagaccg**c**	23	58.9
		9184rev	gtagaggcttactagaagtgt**g**	22	58.4

The primer base located in position 3′ that differs from the sequence of chromosome 1 is shown in bold.

appears to be Human specific since it is present in *Homo sapiens* but it is not present in other primates (see alignment results as online Supporting Information S2). To date, its variability (presence/absence) in Human population is unknown, and we cannot discard the possibility of nDNA co-amplification at least in some individuals. For this reason, a new design, description and in silico validation (according to the methods presented in [1]) of the fragment 5 primer pair is presented.

The size of the redefined chromosome 1 NUMT does not allow locating a primer out of this region following a normal PCR amplification strategy. Consequently, a new approach to primer design was undertaken: (i) primers were relocated to areas with low degree of homology with chromosome 1 and (ii) the 3′-end of each primer were located in mismatch positions relatively to chromosome 1 sequence. Therefore, the possibility of co-amplification of nDNA and mtDNA is reduced. Optimized primer pair and PCR conditions for this concrete region of mtDNA are reported in Table 1. Internal sequence primers previously reported [1] are not affected and can still be used to sequence this region.

This work was supported by MEC (project: CGL2006–07374) and by Generalitat de Catalunya (Ref. 2009 SGR 566).

The authors have declared no conflict of interest.

References

[1] Ramos, A., Santos, C., Alvarez, L., Nogues, R., Aluja, M. P., *Electrophoresis* 2009, *30*, 1587–1593.

[2] Yao, Y. G., Kong, Q. P., Salas, A., Bandelt, H. J., *J. Med. Genet.* 2008, *45*, 769–772.

[3] Parr, R. L., Maki, J., Reguly, B., Dakubo, G. D., Aguirre, A., Wittock, R., Robinson, K., Jakupciak, J. P., Thayer, R. E., *BMC Genomics* 2006, *7*, 185.

[4] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., *J. Mol. Biol.* 1990, *215*, 403–410.

[5] Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., Howell, N., *Nat. Genet.* 1999, *23*, 147.

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 mitochondrial origin. 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.

© 2011 Elsevier B.V. and Mitochondria Research Society. All rights reserved.

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

```
Score =  149 bits (164),  Expect = 6e-33
 Identities = 223/311 (72%), Gaps = 17/311 (5%)
 Strand=Plus/Minus

Query  12032     AACATAAAACCCTCATTCACACGAGAAAACACCCTCATGTTCATACACCTATCCCCCATT  12091
                 ||||| ||||| ||||||||||||||||       || |||  | || |||       ||
Sbjct  94387832  AACATTAAACCATCATTCACACGAGAAAA-------ATATTCTAATACTTATATAT--TT  94387782

Query  12092     CTCCTCCTATCCCTCAACCCCGACATCATTACCGGGTTTTCCTCT--TGTAAATATAGTT  12149
                 || |||||||| |  |||||| | |  |||  |  ||   | | |  | ||||||| |||
Sbjct  94387781  CTTCTCCTATCAC--AACCCCCAAACTATT--CTAGTACCCATATATTATAAATATTGTT  94387726

Query  12150     TAACCAAAACATCAGATTGTGAATCTGACAACAGAGGCTTACGACCCCTTATTTACCGAG  12209
                  ||| ||||||| | ||||||| |    |||||||| || ||  || |||||||||| |||
Sbjct  94387725  AAAC-AAAACATTATATTGTGACTAAAACAACAGAAGCCTAAAACTCCTTATTTACTGAG  94387667

Query  12210     AAAGCTCACAAGAACTGCTAACTCATGCCCCCATGTCTAACAACATGGCTTTCTCAACTT  12269
                 ||||     ||| |||| |||||||||||| ||||||  || || || | ||| |||||||
Sbjct  94387666  AAAGTAAGCAAAAACTACTAACTCATGCCTCCATGTAGAATAATATTG-TTTTTCAACTT  94387608

Query  12270     TTAAAGGATAACAGCTATCCATTGGTCTTAGGCCCCAAAAATTTTGGTGCAACTCCAAAT  12329
                 |||||| |||  |   ||||  |||  |||||  | |||||  |   ||| ||||||||||
Sbjct  94387607  TTAAAGCATAGAAATAATCCGCTGGAGTTAGGAGCTAAAAAAATGAGTGTAACTCCAAAT  94387548

Query  12330     AAAAGTAATAA  12340
                 ||||| |||||
Sbjct  94387547  AAAAGCAATAA  94387537
```

Fig. 1. Alignment between mtDNA (region 12,032–12,340) and a NUMT located in chromosome 1 (region 94,387,537–94,387,832). Examples of identical positions (box _ _ _ _), non-identical position (box ____) and gap region (box ……) are shown. Length in nDNA: 296 bp; length in mtDNA: 309 bp; total length of match: 311 bp.

genome (7.86%), appears as an outlier. It encompasses 106 insertions (representing 14.04% of total number of insertions), nearly double the number of insertions that would be expected (Fig. 3a).

The correlation between the size of insertions and their representation on each chromosome is also positive and significant (Spearman correlation: 0.587 $p = 0.003$). However, two outliers, chromosomes 2 and 3, were identified (Fig. 3b). As before, chromosome 2 is an outlier containing the highest number of insertions. But the length of insertions is quite short, covering only 0.96% of the total length of insertions (Table 1). On the contrary, chromosome 3, which is also one of the largest (6.40% in relation to the entire genome) with a relatively high frequency of insertion (representing 5.83% of total number of insertion), presents the highest total length of insertions (17.78%). In both analyses, chromosome 2 appears as an outlier. This would suggest a differential evolution of this chromosome, a hypothesis that deserves further investigation (Fig. 3b).

The mean percentage of identity between nuclear insertions and mtDNA sequence is 79.2%, ranging from 63.5% to 100%. When the analysis of identity is performed for each chromosome (Table 2), the similarity between inserted mitochondrial sequences and the original mitochondrial sequences presents almost the same pattern in all chromosomes, with a mean percentage of identity of ~80% along the chromosomes.

Table 1
Number and length of insertions and their percentages for each chromosome. Length of each chromosome and their percentage are also reported.

Chromosome	Length chromosome (% in relation to the entire genome)	Number of insertions (% in relation to the total number of insertions)	Length of insertions (% in relation to the total length of insertions)
1	249,250,621 (8.05%)	51 (6.75%)	44,798 (8.17%)
2	243,199,373 (7.86%)	106 (14.04%)	5268 (0.96%)
3	198,022,430 (6.40%)	44 (5.83%)	97,489 (17.78%)
4	191,154,276 (6.17%)	41 (5.43%)	35,103 (6.40%)
5	180,915,260 (5.84%)	35 (4.64%)	33,053 (6.03%)
6	171,115,067 (5.53%)	31 (4.11%)	11,738 (2.14%)
7	159,138,663 (5.14%)	43 (5.70%)	36,328 (6.63%)
8	146,364,022 (4.73%)	47 (6.23%)	36,917 (6.73%)
9	141,213,431 (4.56%)	34 (4.50%)	28,498 (5.20%)
10	135,534,747 (4.38%)	35 (4.64%)	20,926 (3.82%)
11	135,006,516 (4.36%)	31 (4.10%)	26,640 (4.86%)
12	133,851,895 (4.32%)	42 (5.56%)	8264 (1.51%)
13	115,169,878 (3.72%)	22 (2.91%)	10,691 (1.95%)
14	107,349,540 (3.47%)	9 (1.19%)	8899 (1.62%)
15	102,531,392 (3.31%)	14 (1.85%)	9395 (1.71%)
16	90,354,753 (2.92%)	15 (1.99%)	15,965 (2.91%)
17	81,195,210 (2.62%)	23 (3.05%)	21,990 (4.01%)
18	78,077,248 (2.52%)	6 (0.79%)	1334 (0.24%)
19	59,128,983 (1.91%)	20 (2.65%)	9318 (1.70%)
20	63,025,520 (2.04%)	9 (1.19%)	7471 (1.36%)
21	48,129,895 (1.55%)	18 (2.38%)	6753 (1.23%)
22	51,304,566 (1.66%)	17 (2.25%)	23,544 (4.29%)
X	155,270,560 (5.02%)	50 (6.62%)	38,843 (7.08%)
Y	59,373,566 (1.92%)	12 (1.59%)	9025 (1.65%)
Total	3,095,677,412 (100%)	755 (100%)	548,250 (100%)

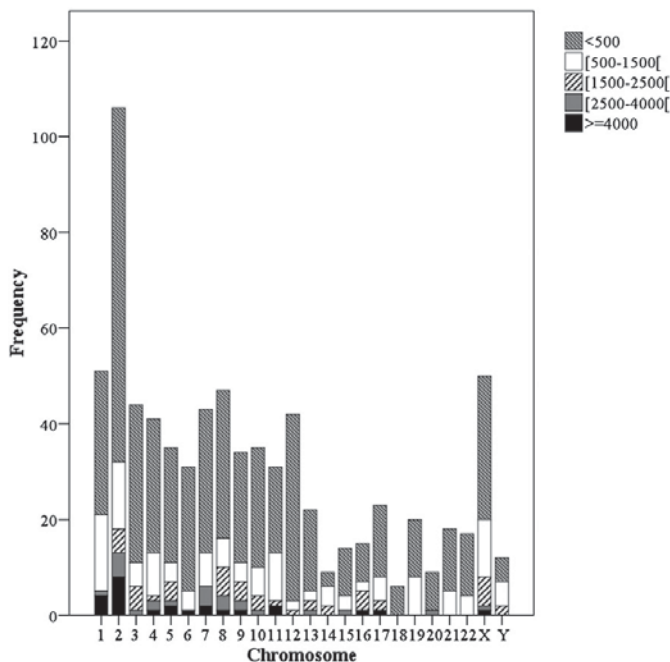


Fig. 2. Frequency and size (bp) of insertions for each chromosome.

The relationship between the size of the insertion and the percentage identity is shown in Table 3. Concerning the size of insertions, it appears that most refer to small insertions (Table 3 and Fig. 2) of less than 500 bp. Only 12.85% of the insertions have more than 1500 bp. The smallest insertions found are 39 bp long and are located on chromosomes 1 and 3. On the other hand, the largest insertion (14,836 bp) is located on chromosome 4 (Table I, Online Supplementary Material I).

About half of the small insertions have a percentage identity higher than 80% of the mitochondrial sequence (Table 3). As the size of the insertion increases, however, the proportion of insertions with a percentage identity equal to or higher than 80% decreases. There are 33 paralogous sequences with more than 500 bp, which present a similarity equal to or higher than 80% (Table 4). These sequences would represent the most problematic, in terms of mtDNA/nDNA co-amplification.

3.2. Applications to cancer

From the 271 mtDNA mutations previously described in cancer samples, and classified as having zero hits in the mtDNA phylogeny by Santos et al. (2008), 15 have one or more hits in the current phylogeny by van Oven and Kayser (2009) (mit. Tree build 8). From the 256 positions with zero hits in the current mtDNA phylogeny (Table II, Online Supplementary Material I), 220 could represent changes in one or more nuclear insertions of mtDNA (Table III of Online Supplementary Material I includes the 220 positions and their matches in non-identical positions in the insertions and the paralogous mtDNA sequences), since they are non-identical between mtDNA and nDNA. Accordingly at that point, only 36 positions can be considered as authentic mtDNA mutations, since they were not found as non-identical in any of the nuclear insertions reported in our study (positions signaled with * in Table II of Online Supplementary Material I).

The 220 positions that can represent false mtDNA mutations correspond to a total of 592 mutations described in the cancer studies, revised in Santos et al. (2008). Of these, 186 mutations were once observed in different types of cancer and/or studies (Table II of Online Supplementary Material I and references therein). The mutations that occur more than once are mostly reported in the study of ovarian cancer by Aikhionbare et al. (2007). In this work, the D-loop and part of the coding region of mtDNA were characterized in 102 samples. The PCR reactions were performed with two primer pairs; after primer BLAST validation, the primer pair that amplifies the region located between positions 16,453–16,569 and 1–1693 reveals no risk of co-amplification. The *in silico* validation of the other primer pair, located between positions 7392 and 8921, reveals a risk of co-amplification between mtDNA and the region of chromosome 1 located between positions 46,578 and 48,102. However, in the NUMTs that can potentially be amplified by this primer set, the mtDNA mutations described by Aikhionbare et al. (2007) were not reported as non-identical positions in our NUMT database (Table III of Online Supplementary Material I).

With regard the remaining studies, in five we were unable to assess their primers (Liu et al., 2001; Lorenc et al., 2003; Parr et al., 2006a; Petros et al., 2005; Zhou et al., 2007), since PCR primers were not reported in the original manuscripts, and after requesting them from the authors no response was obtained.

In the study by Allegra et al. (2006), the primer pair which amplifies the 10,688–11,500 mtDNA region also co-amplifies nDNA; specifically, it amplifies the region 99,385,306–99,386,137 on chromosome 5. Moreover, position 11,203, reported as a mitochondrial mutation by Allegra et al. (2006), was reported as a non-identical position between mtDNA and the NUMT located on chromosome 5 in the region 99,381,642–99,390,749 (position signaled with † in Table III of Online Supplementary Material I).

In ten different studies (Aikhionbare et al., 2007; Allegra et al., 2006; Fliss et al., 2000; Hervouet and Godinot, 2006; Jin et al., 2007; Lievre et al., 2005; Mithani et al., 2007; Polyak et al., 1998; Tan et al., 2006) the primer *in silico* validation also revealed a risk of co-amplification

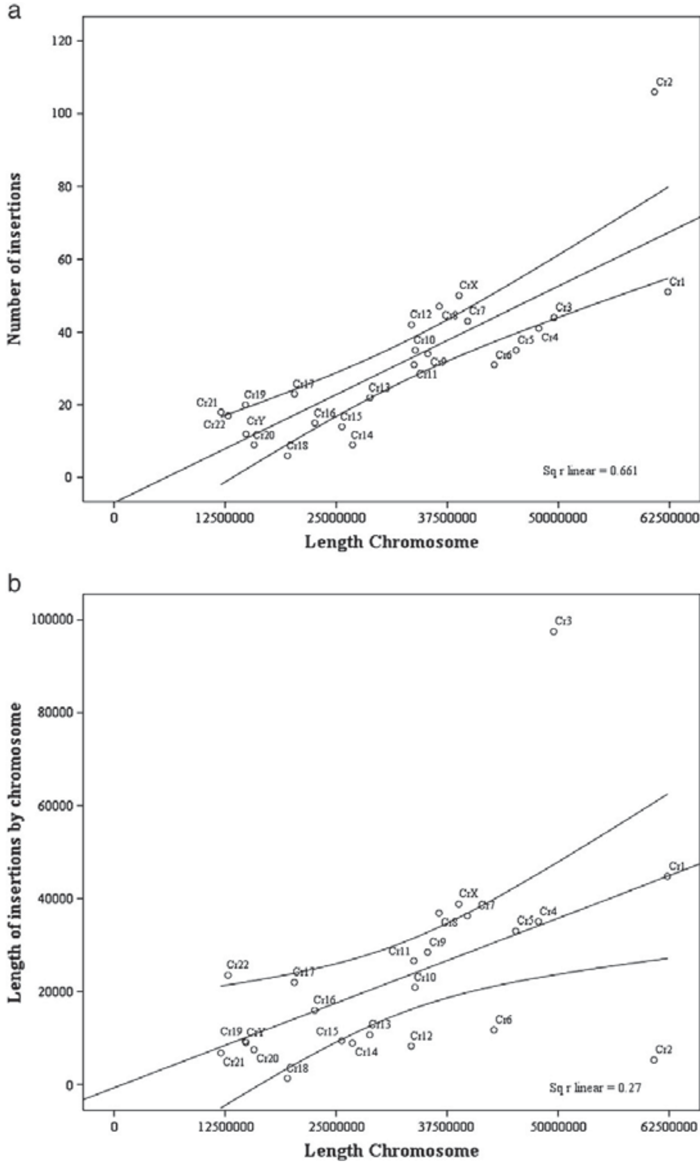


Fig. 3. Correlation between chromosome length and number of insertions (a), and correlation between chromosome and insertion length (b). Coefficient of determination (Sq r linear) for the two correlations is shown.

between mtDNA and nDNA. The twenty-seven positions described by the authors as mtDNA mutations, however, are not reported as non-identical position between mtDNA and the nuclear regions that might potentially be co-amplified.

Finally, 123 positions described in 15 studies (Aikhionbare et al., 2007; Fliss et al., 2000; Gasparre et al., 2007; Habano et al., 1999; Hervouet and Godinot, 2006; Jin et al., 2007; Jones et al., 2001; Kassauei et al., 2006; Lievre et al., 2005; Mithani et al., 2007; Nagy et al., 2003; Parrella et al., 2001; Polyak et al., 1998; Tan et al., 2006; Tzen et al., 2007) are more likely to be authentic, since the primers used to amplify the regions that encompass such mutations did not suggest the possibility of nDNA co-amplification (positions signaled with ** in Table II of Online Supplementary Material I).

4. Discussion

In population and clinical genetics, as well as in the design of evolutionary, phylogenetic and epidemiological studies, it is important to

Table 2
Percentage identity for all chromosomes.

Chromosome		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
Mean		78.76	80.20	79.42	81.24	82.02	80.77	77.67	77.88	78.22	78.31	78.27	83.40	80.61	80.40	79.23	77.37	81.69	84.61	72.92	81.43	75.82	75.74	76.05	79.00
Median		75.73	78.93	79.80	80.43	80.73	81.74	75.83	76.32	77.30	76.98	75.76	88	80.72	79.13	77.65	76.72	80.41	83.96	69.65	77.88	74.05	72.16	73.83	74.57
Minimum		64.71	65.04	68.14	68.71	66.35	64.69	68.46	67.77	68.02	69.08	66.46	67.79	65.41	73.55	71.83	68.26	68.54	74.08	63.50	69.49	65.59	65.67	64.82	70.15
Maximum		100	98.48	97.44	96.77	97.67	93.33	97.09	98.81	90.20	92	98.61	98.88	99.22	93.07	95.35	85.45	97.06	95.57	86.59	100	98.75	100	96.30	100
Percentiles	25	71.43	74.58	73.16	76.19	76.47	77.78	72.22	72.37	73.97	74.14	74.44	77.58	72.41	74.75	75.15	73.53	77.06	75.89	67.43	72.40	69.09	69.32	69.78	71.38
	50	75.73	78.93	79.80	80.43	80.73	81.74	75.83	76.32	77.30	76.98	75.76	88	80.72	79.13	77.65	76.72	80.41	83.96	69.65	77.88	74.04	72.16	73.83	74.57
	75	84.03	85.47	83.56	84.86	88.36	85	82.02	82.35	80.71	82.08	81.67	88	88.59	86.02	83.37	81.73	83.42	94.08	79.04	94.10	79.71	80.31	80.95	82.35

Table 3
Frequency and identity of insertions in relation to the size of insertion (bp).

Size (bp)	Identity		Total	Percentage
	≥80%	<80%		
<500	276	245	521	69.00
[500–1500]	21	116	137	18.15
[1500–2500]	5	43	48	6.36
[2500–4000]	3	21	24	3.18
≥4000	4	21	25	3.31
Total	309	446	755	100

have a complete overview of the total number of NUMTs, their location and variation. In this study, a NUMT database obtained using the latest version of the Human Genome — GRCh37 draft is reported. This is, to our knowledge, the first database that is based on this latest version of the human genome. Moreover, it is also the first where the non-identical positions between mtDNA and nDNA are listed for each NUMT.

In this study, 755 nuclear insertions of mtDNA origin were reported. A comparison of the total number of NUMTs in previously published works clearly shows significant discrepancies. The number of reported NUMTs ranges from 190 (Lascaro et al., 2008) to 1105 (Tourmen et al., 2002), with reports of 206 (Richly and Leister, 2004), 211 (Ricchetti et al., 2004), 247 (Mishmar et al., 2004), 452 (Hazkani-Covo and Graur, 2007), 612 (Woischnik and Moraes, 2002) and 871 (Hazkani-Covo et al., 2010) insertions. The reasons for such discrepancies have been discussed by others previously (Lascaro et al., 2008). According to Lascaro et al. (2008), the incautious usage of bioinformatics methods and the application of methods to an as yet not correctly assembled nuclear genome, are the main factors underlying the disagreement observed between studies. In this sense, the Human mtDNA Reference Sequence (NC_012920) was compared against the human RefSeq Genomic database at NCBI, using the RemoteBlast package in Bioperl bundle (Stajich et al., 2002) with the parameters set to restrict the search to the human organism and the E-value of 10^{-3}. Moreover, regions with less than 20 bp were excluded and only hits in the most recent version of the Human genome draft — GRCh37 primary reference assembly were considered.

Another factor that can induce differences in the number of reported insertions is whether or not the post-insertional processes were evaluated in order to differentiate original insertions from those that result from processes that alter the initially inserted sequence. In this study, post-insertional processes were not considered, since we aimed to obtain an accurate database with all the nuclear insertions and their specific variation (relatively to human mtDNA).

In accordance to other authors [for a review see Hazkani-Covo et al. (2010)], the reported frequency of insertions, as well as their length and distribution along the genome is variable. Most insertions have less than 500 bp. This agrees with the most parsimonious explanation for the origin of NUMT which assumes that mtDNA would be used as a "patch" in the double-strand break repair from non-homologous end joining repair mechanism (Hazkani-Covo et al., 2010; Hazkani-Covo and Covo, 2008; Leister, 2005).

As previously mentioned, 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). In this context, large insertions with a high percentage of identity with mtDNA would be particularly problematic. In fact, one of these insertions, located on chromosome 1, was previously reported to be a source of several errors and misinterpretations (Yao et al., 2008). We characterized the 33 large insertions (with more than 500 bp) that present a percentage identity greater than 80%. Its size makes it difficult to perform primer designs to prevent co-amplification between nuclear insertions and mtDNA (Ramos et al., 2009; Ramos et al., 2011).

This study reports for the first time all identical and non-identical positions between the NUMTs and the GRCh37 version of the nuclear

Table 4
Paralogous sequences with more than 500 bp and that present a similarity equal or higher than 80%.

Chromosome	% Identity	nDNA length	mtDNA length	mtDNA region	nDNA region
1	98.53	5844	5845	3911–9755	564,461–570,304
1	84.03	517	521	987–1507	142,792,601–142,793,117
1	83.87	518	521	987–1507	143,243,224–143,243,741
1	83.78	515	519	987–1505	143,344,679–143,345,193
2	84.03	516	521	987–1507	95,564,754–95,565,269
2	83.95	728	729	6270–6998	50,815,826–50,816,553
2	81.78	1350	1349	4854–6202	156,119,971–156,121,320
3	95.39	1323	1323	1396–2718	96,336,032–96,337,354
3	83.31	623	623	7151–7773	120,440,870–120,441,492
3	80.38	1621	1622	1418–3039	40,293,638–40,295,258
5	95.05	2114	2121	577–2697	79,945,841–79,947,954
5	94.06	5219	5219	10,269–15,487	134,258,999–134,264,217
5	88.36	9108	9067	6117–15,183	99,381,642–99,390,749
5	87.24	3463	3463	12,662–16,124	93,903,161–93,906,623
5	84.54	962	961	577–1537	123,096,499–123,097,460
6	90.70	527	527	2408–2934	62,284,008–62,284,534
7	83.66	2514	2517	577–3093	142,373,012–142,375,525
7	81.27	562	561	2740–3300	141,504,769–141,505,330
9	83.85	2517	2517	577–3093	33,656,612–33,659,128
10	84.62	533	533	8281–8813	101,817,140–101,817,672
11	94.24	2394	2396	577–2972	10,529,434–10,531,827
11	81.68	560	568	596–1163	87,524,440–87,524,999
14	93.07	1021	1024	5583–6606	32,953,304–32,954,324
17	96.17	653	653	6818–7470	51,183,094–51,183,746
17	83.42	10,477	10,536	577–11,112	22,021,365–22,031,841
17	81.73	621	619	2656–3274	19,504,577–19,505,197
17	80.57	1514	1521	577–2097	19,501,874–19,503,387
17	80.41	2206	2205	14,365–16,569	22,018,521–22,020,726
17	80.24	575	582	2084–2665	19,503,699–19,504,273
21	83.78	515	519	987–1505	9,735,524–9,736,038
X	93.68	554	554	10,606–11,159	125,606,714–125,607,267
X	93.07	749	750	6553–7302	125,605,687–125,606,435
Y	83.08	517	521	987–1507	13,290,151–13,290,667

genome and the mtDNA Reference Sequence. This information can be used as a tool in the authentication of mtDNA mutations reported in different fields. Notwithstanding, it is worth mentioning that, in this compilation, the variation of NUMTs was not considered since no studies of NUMT sequence variation had yet been performed to date. Moreover, the mtDNA variation was only taken into account for human mtDNA, although the original insertion could in fact represent mtDNA sequence from very different species.

The described NUMT database was used as a tool to assess the impact of NUMT contamination on cancer studies. Several authors detected innumerable deficiencies in the medical literature related to the analysis and interpretation of mtDNA data [Salas et al. (2005), Yao et al. (2004) among others]. Concerning cancer studies, Salas et al. (2005) showed the consequences of poor experimental design, both in the misinterpretation of the role of mtDNA in the complex tumoral process, as well as in the comparison of results. Also, Yao et al. (2004) showed that researchers may still ignore the possibility of a NUMT contribution when a seemingly novel mtDNA sequence is encountered. Using the NUMT database, we evaluated 256 mtDNA positions that were previously reported to be mutated in cancer samples and that present zero hits in the updated mtDNA phylogeny (Van Oven and Kayser, 2009). From those positions, 220 (identified in 21 studies) could represent changes in one or more nuclear insertions of mtDNA since they are non-identical between mtDNA and nDNA. Thus, we raise the hypothesis that these changes could represent nuclear variations instead of mtDNA variation. To test this hypothesis, we further validated *in silico* the primer pairs used in several original studies if they were accessible in the publication or if they were provided by the authors after request. After primer *in silico* validation, from a total of 16 works for which primers were available, only six studies (Habano et al., 1999; Jones et al., 2001; Kassauei et al., 2006; Nagy et al., 2003; Parrella et al., 2001; Tzen et al., 2007) did not evidence the possibility of nDNA co-amplification. In the remaining studies (Table III of Online Supplementary Material I and references therein), some of the positions reported are located in regions that could be the result of co-amplification. However, with the exception of position 11,203 reported by Allegra et al. (2006), these positions are not reported as non-identical between the potentially co-amplified mtDNA and nDNA regions. Even so, we must take into account the fact that the variability of NUMTs is unknown, and that the non-identical position database was generated using the mtDNA and nDNA Reference Sequences. Thus, the differences between mtDNA and nDNA reported in our database are surely underestimated and the above mentioned positions could be non-authentic mtDNA mutations. On the other hand, it is possible that these changes actually represent authentic mtDNA mutations; however, this hypothesis seems extremely difficult to test.

Using cancer studies as an example application, we aimed to highlight the applicability of the described NUMT database as a new tool to validate *in silico* mtDNA mutations described in different contexts; it is clear that there is a need to follow standardized protocols that avoid co-amplification between mitochondrial and nuclear paralogous sequences. Moreover, and due to the amount of information provided for each nuclear insertion, this database could be applied as a tool to design evolutionary, phylogenetic and epidemiological studies.

To our knowledge, this is the first work to report the non-identical positions between the mtDNA and nDNA. These positions are the basis for the application to cancer studies reported in this work, and knowledge of their distribution is useful for the authentication of mtDNA mutations reported in different kinds of study, especially those occurring in heteroplasmy.

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mito.2011.08.009.

Acknowledgments

This work was supported by MEC (project: CGL2009-08205) and by the Generalitat de Catalunya (Ref. 2009 SGR 566).

We are in debt with Dr. Czarnecka, Dr Liu and Dr. Maitra, the corresponding authors of cancer studies for supplying the information necessary for the analysis.

References

Aikhionbare, F.O., Mehrabi, S., Kumaresan, K., Zavareh, M., Olatinwo, M., Odunsi, K., Partridge, E., 2007. Mitochondrial DNA sequence variants in epithelial ovarian tumor subtypes and stages. J. Carcinog. 6, 1.

Allegra, E., Garozzo, A., Lombardo, N., De Clemente, M., Carey, T.E., 2006. Mutations and polymorphisms in mitochondrial DNA in head and neck cancer cell lines. Acta Otorhinolaryngol. Ital. 26, 185–190.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Bensasson, D., Zhang, D., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. 16, 314–321.

Blanchard, J.L., Lynch, M., 2000. Organellar genes: why do they end up in the nucleus? Trends Genet. 16, 315–320.

Caro, P., Gomez, J., Arduini, A., Gonzalez-Sanchez, M., Gonzalez-Garcia, M., Borras, C., Vina, J., Puertas, M.J., Sastre, J., Barja, G., 2010. Mitochondrial DNA sequences are present inside nuclear DNA in rat tissues and increase with age. Mitochondrion 10, 479–486.

Fliss, M.S., Usadel, H., Caballero, O.L., Wu, L., Buta, M.R., Eleff, S.M., Jen, J., Sidransky, D., 2000. Facile detection of mitochondrial DNA mutations in tumors and bodily fluids. Science 287, 2017–2019.

Gasparre, G., Porcelli, A.M., Bonora, E., Pennisi, L.F., Toller, M., Iommarini, L., Ghelli, A., Moretti, M., Betts, C.M., Martinelli, G.N., Ceroni, A.R., Curcio, F., Carelli, V., Rugolo, M., Tallini, G., Romeo, G., 2007. Disruptive mitochondrial DNA mutations in complex I subunits are markers of oncocytic phenotype in thyroid tumors. Proc. Natl. Acad. Sci. U.S.A. 104, 9001–9006.

Goios, A., Prieto, L., Amorim, A., Pereira, L., 2008. Specificity of mtDNA-directed PCR-influence of NUclear MTDNA insertion (NUMT) contamination in routine samples and techniques. Int. J. Legal Med. 122, 341–345.

Goldin, E., Stahl, S., Cooney, A.M., Kaneski, C.R., Gupta, S., Brady, R.O., Ellis, J.R., Schiffmann, R., 2004. Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV. Hum. Mutat. 24, 460–465.

Habano, W., Sugai, T., Yoshida, T., Nakamura, S., 1999. Mitochondrial gene mutation, but not large-scale deletion, is a feature of colorectal carcinomas with mitochondrial microsatellite instability. Int. J. Cancer 83, 625–629.

Hazkani-Covo, E., 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. Mol. Biol. Evol. 26, 2175–2179.

Hazkani-Covo, E., Covo, S., 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet. 4, e1000237.

Hazkani-Covo, E., Graur, D., 2007. A comparative analysis of numt evolution in human and chimpanzee. Mol. Biol. Evol. 24, 13–18.

Hazkani-Covo, E., Sorek, R., Graur, D., 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J. Mol. Evol. 56, 169–174.

Hazkani-Covo, E., Zeller, R.M., Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. 6, e1000834.

Hervouet, E., Godinot, C., 2006. Mitochondrial disorders in renal tumors. Mitochondrion 6, 105–117.

Jensen-Seaman, M.I., Wildschutte, J.H., Soto-Calderón, I.D., Anthony, N.M., 2009. A comparative approach shows differences in patterns of numt insertion during hominoid evolution. J. Mol. Evol. 68.

Jin, X., Zhang, J., Gao, Y., Ding, K., Wang, N., Zhou, D., Jen, J., Cheng, S., 2007. Relationship between mitochondrial DNA mutations and clinical characteristics in human lung cancer. Mitochondrion 7, 347–353.

Jones, J.B., Song, J.J., Hempen, P.M., Parmigiani, G., Hruban, R.H., Kern, S.E., 2001. Detection of mitochondrial DNA mutations in pancreatic cancer offers a "mass"-ive advantage over detection of nuclear DNA mutations. Cancer Res. 61, 1299–1304.

Kassauei, K., Habbe, N., Mullendore, M.E., Karikari, C.A., Maitra, A., Feldmann, G., 2006. Mitochondrial DNA mutations in pancreatic cancer. Int. J. Gastrointest. Cancer 37, 57–64.

Kleine, T., Maier, U.G., Leister, D., 2009. DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. Annu. Rev. Plant Biol. 60, 115–138.

Lascaro, D., Castellana, S., Gasparre, G., Romeo, G., Saccone, C., Attimonelli, M., 2008. The RHNumtS compilation: features and bioinformatics approaches to locate and quantify Human NumtS. BMC Genomics 9, 267.

Leister, D., 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet. 21, 655–663.

Lievre, A., Chapusot, C., Bouvier, A.M., Zinzindohoue, F., Piard, F., Roignot, P., Arnould, L., Beaune, P., Faivre, J., Laurent-Puig, P., 2005. Clinical value of mitochondrial mutations in colorectal cancer. J. Clin. Oncol. 23, 3517–3525.

Liu, V.W., Shi, H.H., Cheung, A.N., Chiu, P.M., Leung, T.W., Nagley, P., Wong, L.C., Ngan, H.Y., 2001. High incidence of somatic mitochondrial DNA mutations in human ovarian carcinomas. Cancer Res. 61, 5998–6001.

Lorenc, A., Bryk, J., Golik, P., Kupryjanczyk, J., Ostrowski, J., Pronicki, M., Semczuk, A., Szolkowska, M., Bartnik, E., 2003. Homoplasmic MELAS A3243G mtDNA mutation in a colon cancer sample. Mitochondrion 3, 119–124.

Mishmar, D., Ruiz-Pesini, E., Brandon, M., Wallace, D.C., 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. Hum. Mutat. 23, 125–133.

Mithani, S.K., Taube, J.M., Zhou, S., Smith, I.M., Koch, W.M., Westra, W.H., Califano, J.A., 2007. Mitochondrial mutations are a late event in the progression of head and neck squamous cell cancer. Clin. Cancer Res. 13, 4331–4335.

Nagy, A., Wilhelm, M., Kovacs, G., 2003. Mutations of mtDNA in renal cell tumours arising in end-stage renal disease. J. Pathol. 199, 237–242.

Parfait, B., Rustin, P., Munnich, A., Rotig, A., 1998. Co-amplification of nuclear pseudogenes and assessment of heteroplasmy of mitochondrial DNA mutations. Biochem. Biophys. Res. Commun. 247, 57–59.

Parr, R.L., Dakubo, G.D., Crandall, K.A., Maki, J., Reguly, B., Aguirre, A., Wittock, R., Robinson, K., Alexander, J.S., Birch-Machin, M.A., Abdel-Malak, M., Froberg, M.K., Diamandis, E.P., Thayer, R.E., 2006a. Somatic mitochondrial DNA mutations in prostate cancer and normal appearing adjacent glands in comparison to age-matched prostate samples without malignant histology. J. Mol. Diagn. 8, 312–319.

Parr, R.L., Maki, J., Reguly, B., Dakubo, G.D., Aguirre, A., Wittock, R., Robinson, K., Jakupciak, J.P., Thayer, R.E., 2006b. The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. BMC Genomics 7, 185.

Parrella, P., Xiao, Y., Fliss, M., Sanchez-Cespedes, M., Mazzarelli, P., Rinaldi, M., Nicol, T., Gabrielson, E., Cuomo, C., Cohen, D., Pandit, S., Spencer, M., Rabitti, C., Fazio, V.M., Sidransky, D., 2001. Detection of mitochondrial DNA mutations in primary breast cancer and fine-needle aspirates. Cancer Res. 61, 7623–7626.

Petros, J.A., Baumann, A.K., Ruiz-Pesini, E., Amin, M.B., Sun, C.Q., Hall, J., Lim, S., Issa, M. M., Flanders, W.D., Hosseini, S.H., Marshall, F.F., Wallace, D.C., 2005. mtDNA mutations increase tumorigenicity in prostate cancer. Proc. Natl. Acad. Sci. U.S.A. 102, 719–724.

Polyak, K., Li, Y., Zhu, H., Lengauer, C., Willson, J.K., Markowitz, S.D., Trush, M.A., Kinzler, K.W., Vogelstein, B., 1998. Somatic mutations of the mitochondrial genome in human colorectal tumours. Nat. Genet. 20, 291–293.

Ramos, A., Santos, C., Alvarez, L., Nogués, R., Aluja, M.P., 2009. Human mitochondrial DNA complete amplification and sequencing: a new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification. Electrophoresis 30, 1–7.

Ramos, A., Santos, C., Barbena, E., Mateiu, L., Alvarez, L., Nogues, R., Aluja, M.P., 2011. Validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification: a revision based on the New Human Genome Reference Sequence (GRCh37). Electrophoresis 32, 782–783.

Ricchetti, M., Tekaia, F., Dujon, B., 2004. Continued colonization of the human genome by mitochondrial DNA. PLoS Biol. 2, E273.

Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. Mol. Biol. Evol. 21, 1081–1084.

Salas, A., Carracedo, A., Macaulay, V., Richards, M., Bandelt, H.J., 2005. A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. Biochem. Biophys. Res. Commun. 335, 891–899.

Santos, C., Martínez, M., Lima, M., Hao, Y.-J., Simoes, N., Montiel, R., 2008. Mitochondrial DNA mutations in cancer: a review. Curr. Top. Med. Chem. 8, 1351–1366.

Stajich, J., Block, D., Boulez, K., Brenner, S., Chervitz, S., Dagdigian, C., Fuellen, G., Gilbert, J., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C., Osborne, B., Pocock, M., Schattner, P., Senger, M., Stein, L., Stupka, E., Wilkinson, M., Birney, E., 2002. The BioPerl toolkit: Perl modules for the life sciences. Genome Res. 12, 1611–1618.

Tan, D.J., Chang, J., Liu, L.L., Bai, R.K., Wang, Y.F., Yeh, K.T., Wong, L.J., 2006. Significance of somatic mutations and content alteration of mitochondrial DNA in esophageal cancer. BMC Cancer 6, 93.

Thalmann, O., Serre, D., Hofreiter, M., Lukas, D., Eriksson, J., Vigilant, L., 2005. Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. Mol. Ecol. 14, 179–188.

Tourmen, Y., Baris, O., Dessen, P., Jacques, C., Malthiery, Y., Reynier, P., 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics 80, 71–77.

Turner, C., Killoran, C., Thomas, N.S.T., Rosenberg, M., Chuzhanova, N.A., Johnston, J., Kemel, Y., Cooper, D.N., Biesecker, L.G., 2003. Human genetic disease caused by *de novo* mitochondrial-nuclear DNA transfer. Hum. Genet. 112, 303–309 (112, 303–309).

Tzen, C.Y., Mau, B.L., Wu, T.Y., 2007. ND4 mutation in transitional cell carcinoma: does mitochondrial mutation occur before tumorigenesis? Mitochondrion 7, 273–278.

Van Oven, M., Kayser, M., 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 30, E386–E394.

Woischnik, M., Moraes, C.T., 2002. Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Res. 12, 885–893.

Yao, Y.G., Bravi, C.M., Bandelt, H.J., 2004. A call for mtDNA data quality control in forensic science. Forensic Sci. Int. 141, 1–6.

Yao, Y.G., Kong, Q.P., Salas, A., Bandelt, H.J., 2008. Pseudomitochondrial genome haunts disease studies. J. Med. Genet. 45, 769–772.

Zhou, S., Kachhap, S., Sun, W., Wu, G., Chuang, A., Poeta, L., Grumbine, L., Mithani, S.K., Chatterjee, A., Koch, W., Westra, W.H., Maitra, A., Glazer, C., Carducci, M., Sidransky, D., McFate, T., Verma, A., Califano, J.A., 2007. Frequency and phenotypic implications of mitochondrial DNA mutations in human squamous cell cancers of the head and neck. Proc. Natl. Acad. Sci. U.S.A. 104, 7540–7545.