Aproximació molecular a l'estudi dels primats: Evolució dels gens *RPS4Y* i aplicació d'SNPs en conservació

Memòria presentada per
Olga Andrés Viñas

Per optar al grau de
Doctora

Barcelona, 4 d'abril del 2007

OBJECTIUS

La situació actual de gran part de les espècies de primat és crítica, principalment com a conseqüència de l'activitat humana, ja que la destrucció de l'hàbitat i la persecució de què són víctimes estan conduint a una reducció dràstica de les poblacions, amb pèrdua de diversitat genètica i disminució de l'eficàcia biològica. La diversitat genètica és necessària perquè les poblacions puguin evolucionar i adaptar-se als possibles canvis ambientals que es van produint contínuament, de manera que el manteniment de la diversitat és el focus principal de la conservació biològica. Cada vegada és més evident, doncs, el paper fonamental que pot jugar la genètica en la conservació de les espècies. D'altra banda, des d'un punt de vista antropocèntric, l'estudi dels primats no humans és de gran transcendència en biomedicina i evolució, ja que conté els tàxons més propers a l'home, i la genòmica comparada pot revelar informació sobre els nostres orígens, sobre els gens que ens han donat les característiques que ens fan humans i sobre els gens alterats en les malalties genètiques.

Davant d'aquesta situació, el treball de tesi presentat aquí és una aproximació a l'estudi dels primats des d'un punt de vista molecular amb l'objectiu principal de contribuir a la conservació de les espècies. És per això que els estudis realitzats es basen, per una banda, en el desenvolupament i assaig d'eines de biologia molecular que poden ser fonamentals en genètica de la conservació i, per altra banda, en l'obtenció d'informació genètica bàsica per avançar en el coneixement evolutiu dels primats. Els objectius concrets d'aquesta tesi són:

- *Determinar l'aplicabilitat de la tècnica MDA sobre mostres no invasives de primat.*

  Amb aquest treball es pretén estudiar la possibilitat d'aplicar la tècnica d'amplificació total de genomes *Multiple Displacement Amplification* (MDA) sobre mostres de semen de macaco recol·lectades de manera no invasiva. Per determinar si els productes amplificats són aptes per dur a terme estudis genètics s'ha realitzat una comparació entre seqüències de la regió control del DNA mitocondrial obtingudes a partir de les mostres originals i a partir dels productes de la reacció d'MDA.

- *Contribuir amb el desenvolupament d'un* microarray *de miniseqüència a l'estudi i conservació de poblacions de ximpanzé.*

  Per desenvolupar el *microarray* en aquest estudi s'ha dut a terme el descobriment de nous SNPs en el cromosoma Y de l'espècie *Pan troglodytes* amb la tècnica de la *Denaturing high-performance liquid chromatography* (DHPLC) i, a més, s'ha realitzat una cerca d'SNPs en les seqüències de gens del DNA mitocondrial presents a les bases de dades. El *microarray* de miniseqüència ha de permetre l'estudi automatitzat de poblacions captives i salvatges de ximpanzé, de manera que s'ha testat els dos tipus de poblacions. A més, per contribuir a l'estudi de la variabilitat genètica de dues poblacions salvatges, s'ha seqüenciat la regió control

del DNA mitocondrial per comparar els resultats amb la informació obtinguda de l'aplicació del *microarray*.

- *Estudiar l'evolució dels gens* RPS4Y *dins la filogènia dels primats.*

A partir de l'amplificació i seqüenciació d'introns dels gens *RPS4Y*, aquest treball pretén aclarir la història evolutiva dels gens *RPS4Y* en primats, confirmar si la primera còpia del gen *RPS4Y* està present en tots els infraordres de primat i identificar quins infraordres posseeixen la segona còpia lligada a cromosoma Y, per tant, determinar el moment en què es va produir la duplicació que va generar el gen *RPS4Y2*. D'altra banda, les seqüències de cDNA d'espècies de primat presents a les bases de dades públiques han permès estudiar, amb mètodes de màxima versemblança, la possible acció de la selecció positiva en l'evolució dels gens *RPS4Y* i, en cas d'haver-hi selecció, caracteritzar els llinatges i les posicions afectades. Finalment, s'han inferit els mecanismes evolutius que possiblement han actuat després de la duplicació per mantenir les dues còpies del gen *RPS4* en el cromosoma Y.

RESULTATS

CAPÍTOL 1

Article

Títol: ***Sequence quality is maintained after multiple displacement amplification of non-invasively obtained macaque semen DNA***

Autors: O. Andrés, A.C. Rönn, A. Ferrando, M. Bosch, X. Domingo-Roura

Referència: *Biotechnology Journal*, 2006, 1: 466–469

Resum

Els protocols d'amplificació de genomes sencers estan revolucionant els camps de la biologia molecular i de la genètica de la conservació ja que aporten la possibilitat d'obtenir un nombre elevat de còpies d'un genoma complet a partir de quantitats mínimes de mostra. L'Amplificació per Desplaçament Múltiple (MDA) és una tècnica d'amplificació de genomes sencers basada en les propietats de la polimerasa de DNA phi29, que dóna una representació uniforme del genoma amb unes taxes d'error molt baixes. En el nostre estudi s'ha aplicat la tècnica de l'MDA en 28 mostres de DNA extretes de sang o de semen recollit de manera no invasiva i s'ha obtingut la seqüència de la regió control del DNA mitocondrial abans i després de l'MDA. S'ha pogut demostrar que el número de posicions no resoltes és comparable abans i després d'aplicar l'MDA, tot i que la longitud de les seqüències llegibles ha resultat ser més llarga en les mostres originals que en els productes d'MDA. Es pot concloure que la tècnica de l'MDA és útil per incrementar la quantitat de DNA de mostres de semen recol·lectades de manera no invasiva ja que proporciona productes adequats per realitzar estudis de seqüenciació de regions mitocondrials.

Aportació personal al treball

Tots els experiments de seqüenciació dels productes obtinguts d'MDA han estat realitzats per mi per tal de comparar els resultats amb seqüències de les mostres originals que s'havien obtingut anteriorment. Les anàlisis comparatives també han estat dutes a terme per mi. A més, he realitzat l'escriptura de l'article, amb la col·laboració de tots els autors.

DOI 10.1002/biot.200600011 Biotechnol. J. 2006, 1, 466–469

Short Communication

Sequence quality is maintained after multiple displacement amplification of non-invasively obtained macaque semen DNA

*Olga Andrés[1,2], Ann-Charlotte Rönn[3], Aïnhoa Ferrando[1,4], Montserrat Bosch[1,2] and Xavier Domingo-Roura[1]**

[1]Genètica de la Conservació, Institut de Recerca i Tecnologia Agroalimentàries, Cabrils, Spain
[2]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain
[3]Molecular Medicine, Department of Medical Sciences, Uppsala University, Academic Hospital, Uppsala, Sweden
[4]Departament de Biologia Cel·lular, de Fisiologia i d'Immunologia, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

Whole genome amplification protocols are revolutionizing the fields of molecular and conservation biology as they open the possibility of obtaining a large number of copies of a complete genome from minute amounts of sample. Multiple displacement amplification (MDA) is a whole genome amplification technique based on the properties of the phi29 DNA polymerase, which leads to a uniform representation of the genome with very low error rates. In this study we performed MDA on 28 macaque DNA samples extracted from blood or non-invasively collected semen from which we obtained mitochondrial control region sequences both before and after MDA. The length of the readable sequences was longer for the original samples than for the MDA products, but the number of unresolved positions was comparable both before and after MDA. We conclude that the MDA technique is useful for increasing the amount of DNA for sequencing mitochondrial regions in the case of non-invasively collected semen samples.

Received 26 January 2006
Revised 15 February 2006
Accepted 17 February 2006

Keywords: Japanese macaque · Mitochondrial DNA · Non-invasive samples · Semen · Whole genome amplification

Non-invasive genetic sampling is becoming a crucial issue (*e.g.*, [1]), especially when working on endangered, elusive or socially complex species. It has been possible to successfully obtain DNA from different non-invasive sources, such as hair [2], feces [3], urine [4], buccal cells [5] and semen [6]. However, the total amounts of DNA recovered this way are usually very small. Whole genome amplification (WGA) techniques, which allow the acquisition of large numbers of DNA copies covering the whole genome starting from minute amounts of original DNA, could contribute to overcome this limitation. Several WGA methods have been developed based on PCR. These include degenerate oligonucleotide-primed PCR [7] and primer extension pre-amplification [8]. However, these techniques tend to be characterized by an imbalanced amplification of microsatellite [9] and single nucleotide polymorphism alleles [10], in addition to providing partial coverage of the genome in the amplified products [11]. Moreover, the use of degenerate PCR primers in these methods can cause artificial sequence variation in the amplified products. More recently the multiple displacement amplification (MDA) procedure [11], a new method based on the activity of the bacteriophage phi29 DNA polymerase [12], has been described. The phi29 DNA polymerase is highly progressive and has DNA strand-displacing activity. This activity is used in MDA to extend denatured genomic DNA from annealed random hexamer primers, and to thereby form up to 100-kb DNA products. This process implies a very low error rate, on account of

Correspondence: Olga Andrés, Genètica de la Conservació, Institut de Recerca i Tecnologia Agroalimentàries, Crta. de Cabrils s/n, 08348 Cabrils, Spain
Fax: +34-937533954
E-mail: olga.andres@irta.es; olga.andres@gmail.com

Abbreviations: MDA, multiple displacement amplification; **WGA**, whole genome amplification

* Dr. Xavier Domingo-Roura passed away on November 17th 2005, after a brave fight against cancer.

the 3'→5' exonuclease activity of the polymerase [13]. Following this isothermal reaction, thousands or even millions of copies of the whole genome can be generated from small amounts of initial genomic DNA – a few nanograms of source DNA are sufficient [14] –, and the resulting genome coverage is greater than that obtained using PCR-based methods [11].

MDA seems to be a possible solution to overcome the difficulty of low DNA amounts obtained from non-invasive sampling. Nevertheless, before establishing MDA as the key, it is necessary to prove that this technique does not introduce errors or indeterminations in the sequences. Thus, the objective of our study was to test the feasibility of employing MDA to obtain a permanent source of DNA from non-invasive samples suitable for sequencing. We performed MDA on DNA from primate blood and non-invasively collected semen samples. We obtained mitochondrial DNA sequences from both original DNA and MDA products, and statistically evaluated whether they were different.

Twenty-eight DNA extracts of Japanese macaque (*Macaca fuscata*) were included in this study (Table 1). All samples were used in conformity with the current Spanish legislation and their trade was in accordance with the Convention for the International Trade of Endangered Species (CITES). Twelve samples were extracted from blood using standard phenol-chloroform protocols, whereas 16 samples were extracted from non-invasively collected semen as described in [15]. All DNA extracts were stored at 4°C for more than 10 years. MDA was performed on all samples using the GenomiPhi™ DNA Amplification kit (GE Healthcare, Uppsala, Sweden) following manufac-

Table 1. Samples, DNA sources and sequencing results

	Original sequence			WGA sequence			
Sample code	**Readable sequence length (bp)[a)]**	**Number of unresolved nucleotides**	**Unresolved nucleotides in common sequence**	**Readable sequence length (bp)[a)]**	**Number of unresolved nucleotides**	**Unresolved nucleotides in common sequence**	**DNA source**
Yaku2	137	10	10	194	9	0	Semen
Yaku3	213	12	7	194	11	11	Semen
Yaku14	214	9	0	140	5	5	Semen
Yaku27	214	13	11	206	6	6	Semen
Yaku63	214	13	8	199	12	12	Semen
Yaku73	214	8	3	160	9	9	Semen
Yaku72	214	10	10	207	8	8	Semen
Yaku26	214	7	1	142	9	9	Semen
Yaku30	214	6	3	199	16	16	Semen
Yaku38	167	4	4	140	14	14	Semen
Yaku40	214	15	15	214	16	16	Semen
Yaku41	214	16	16	214	17	17	Semen
Yaku43	214	15	6	178	6	6	Semen
Yaku50	215	8	8	214	10	10	Semen
Yaku61	214	16	16	214	11	11	Semen
Yaku64	214	18	18	214	16	16	Semen
Hak866	355	3	2	340	5	5	Blood
Hak867	355	7	6	340	0	0	Blood
Hak868	355	8	8	340	2	2	Blood
Hak870	355	7	7	353	2	2	Blood
Aw1149	352	8	5	240	7	7	Blood
Aw1151	355	7	7	340	4	4	Blood
Aw1153	355	4	2	315	7	7	Blood
Aw1156	355	9	4	223	3	3	Blood
Ko1586	355	8	4	340	3	3	Blood
Ko1631	350	17	16	352	4	4	Blood
Ko1649	355	1	0	349	3	3	Blood
Ko1683	355	8	6	280	2	2	Blood
Mean	269.71	9.54	7.25	244.32	7.75	7.43	
SD	76.39	4.43	5.14	73.85	4.93	5.13	

a) After performing forward and reverse sequencing.

turer's instructions. To test the reliability of MDA, original DNA templates and MDA products obtained from those templates were sequenced under the same conditions. We sequenced 390 bp of the mitochondrial DNA control region using human primers L15996 and H16401 [16] as in [17], but in an ABI Prism 3100 automated DNA sequencer (Applied Biosystems, Foster City, CA, USA). Sequences were found to be identical, excluding unresolved nucleotides from the analysis.

To compare results before and after MDA, we determined both the length of the readable sequence and the number of unresolved bases obtained in each case (Table 1). Analyses were based on 28 individuals (n=28) when all samples were considered, and on 16 individuals (n=16) when only DNA extracted from semen was tested.

A Kolmogorov-Smirnov (K-S) test was first performed to determine if the results followed a normal distribution. The K-S test did not detect any significant evidence to reject normality in the number of unresolved nucleotide positions, either before amplification (Z=0.827; p=0.501; n=28) or after MDA (Z=0.585; p=0.883; n=28). Therefore, a Levene test was performed to confirm variance homogeneity (F=0.109; p=0.743; n=28). As we found normality and variance homogeneity, we carried out the parametric paired-samples t test. The test was not significant (t=–0.163; p=0.872; n=28), indicating that there were no statistically significant differences between the number of unresolved nucleotides in the overlapping region of original and MDA sequences. This lack of significance was also maintained when only DNA extracted from semen was compared (t=–1.257; p= 0.228; n=16).

For readable sequence length, K-S test showed normal distribution for MDA data (Z=1.221; p=0.101; n=28) but not for original data (Z=1.770; p=0.004; n=28), so we carried out a non-parametric Wilcoxon's signed-ranks test. This test was significant (Z=–3.682; p< 0.001; n=28), indicating that it was possible to sequence more base pairs with the original samples than with the MDA products. The same trend was obtained when only semen extracts were compared, although in this case the test was significant at a 0.05 level of probability (Z=–2.276; p=0.023; n=16). All statistical tests were performed using SPSS version 9 software (SPSS Inc, Chicago, IL, USA).

In this study we demonstrated that it is possible to successfully apply MDA on old DNA from non-invasively collected macaque semen samples. MDA products and original DNA provided identical mitochondrial sequences of similar quality, although sequences from MDA products were shorter. Other sources of non-invasively obtained samples, such as feces, hair or material stored in museums under diverse conditions, should now be tested by MDA. Moreover, MDA might be a very useful tool to overcome the constraints of low copy number when obtaining nuclear sequences from non-invasive samples.

In summary, the present study highlights the utility of MDA technique and the possibility – and potential importance – of applying it on non-invasively collected samples. The MDA technique can contribute to the preservation of DNA collections and, in combination with non-invasively collected samples, can help to avoid disturbances when conducting studies on wild populations.

We thank A. C. Syvänen, J. Marmi, J. F. López-Giráldez, R. Lecis and B. Guallar for providing useful comments and references and F. Sanz for advice concerning statistical analyses. Financial support was provided by the European Commission under contract QLRI-CT-2002-01325 (INPRIMAT project). O. Andrés and A. Ferrando are supported by scholarships from the DURSI, Generalitat de Catalunya (Ref. 2003FI-00787 and 2002FI-00280, respectively). We also thank O. Takenaka and the Primate Research Institute, Kyoto University, Japan for supplying some of the samples.

References

[1] Taberlet, P., Griffin, S., Goossens, B., Questiau, S., Manceau, V. *et al.,* Reliable genotyping of samples with very low DNA quantities using PCR. *Nucleic Acids Res.* 1996, *24*, 3189–3194.

[2] Morin, P.A., Moore, J.J., Woodruff, D.S., Identification of chimpanzee subspecies with DNA from hair and allele-specific probes. *Proc. Biol. Sci.* 1992, *249*, 293–297.

[3] Hoss, M., Kohn, M., Paabo, S., Knauer, F., Schroder, W., Excrement analysis by PCR. *Nature* 1992, *359*, 199.

[4] Hayakawa, S., Takenaka, O., Urine as another potential source for template DNA in polymerase chain reaction. *Am. J. Primatol.* 1999, *48*, 299–304.

[5] Takasaki, H., Takenaka, O., Paternity testing in chimpanzees with DNA amplification from hairs and buccal cells in wadges: a preliminary note, in: Ehara, A., Kimura, T., Takenaka, O., Iwamoto, M. (Eds.), *Primatology Today*. Elsevier Science, Amsterdam 1991, pp. 613–616.

[6] Yoshimi, I., Takasaki, H., Long distance mobility of male Japanese macaques evidenced by mitochondrial DNA. *Primates* 2003, *44*, 71–74.

[7] Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjold, M., Ponder, B.A. *et al.,* Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* 1992, *13*, 718–725.

[8] Zhang, L., Cui, X., Schmitt, K., Hubert, R., Navidi, W. *et al.,* Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl. Acad. Sci. USA* 1992, *89*, 5847–5851.

[9] Cheung, V.G., Nelson, S.F., Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proc. Natl. Acad. Sci. USA* 1996, *93*, 14676–14679.

[10] Grant, S.F., Steinlicht, S., Nentwich, U., Kern, R., Burwinkel, B. *et al.*, SNP genotyping on a genome-wide amplified DOP-PCR template. *Nucleic Acids Res.* 2002, *30*, e125.

[11] Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., *et al.*, Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 5261–5266.

[12] Blanco, L., Bernad, A., Lazaro, J.M., Martin, G., Garmendia, C., Highly efficient DNA synthesis by the phage phi 29 DNA polymerase.

Symmetrical mode of DNA replication. *J. Biol. Chem.* 1989, *264*, 8935–8940.

[13] Esteban, J.A., Salas, M., Blanco, L., Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *J. Biol. Chem.* 1993, *268*, 2719–2726.

[14] Lasken, R.S., Egholm, M., Whole genome amplification: abundant supplies of DNA from precious samples or clinical specimens. *Trends Biotechnol.* 2003, *21*, 531–535.

[15] Domingo-Roura, X., Marmi, J., Andres, O., Yamagiwa, J., Terradas, J., Genotyping from semen of wild Japanese macaques (*Macaca fuscata*). *Am. J. Primatol.* 2004, *62*, 31–42.

[16] Vigilant, L., Pennington, R., Harpending, H., Kocher, T.D., Wilson, A.C., Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl. Acad. Sci. USA* 1989, *86*, 9350–9354.

[17] Marmi, J., Bertranpetit, J., Terradas, J., Takenaka, O., Domingo-Roura, X., Radiation and phylogeography in the Japanese macaque, *Macaca fuscata*. *Mol. Phylogenet. Evol.* 2004, *30*, 676–685.

CAPÍTOL 2

Article

Títol: ***A microarray system for Y-chromosomal and mitochondrial SNP analysis in chimpanzee populations***

Autors: O. Andrés*, A.C. Rönn*, M. Bonhomme, T. Kellermann, B. Crouau-Roy, G. Doxiadis, E. Vershoor, B. Goossens, X. Domingo-Roura, M. Bruford, M. Bosch, A.C. Syvänen, the INPRIMAT consortium

Referència: Enviat a *Molecular Ecology*

*Primera autoria compartida amb l'estudiant de doctorat Ann-Charlotte Rönn, de la Universitat d'Uppsala, a Suècia.

Resum

Les poblacions de ximpanzé estan disminuint a gran velocitat a conseqüència de l'activitat humana i actualment es troben en perill d'extinció. En el context dels programes de conservació, les dades genètiques poden aportar informació essencial, per exemple sobre la diversitat genètica i l'estructura de les poblacions amenaçades. Els polimorfismes d'una sola base (SNPs) són marcadors bial·lèlics, àmpliament utilitzats en estudis moleculars en humans, que poden ser implementats en sistemes de *microarray* molt eficients. Aquesta tecnologia ofereix la possibilitat de genotipar molts SNPs simultàniament també en altres organismes, però el seu ús no és comú en estudis poblacionals. En aquest treball es descriu la caracterització de nous SNPs en regions intròniques del cromosoma Y de ximpanzé i també s'identifiquen SNPs en gens mitocondrials amb l'objectiu de desenvolupar un sistema de *microarray* que permeti estudiar de manera simultània tant el llinatge patern com el matern. El sistema consta de 42 SNPs pel cromosoma Y i 45 pel genoma mitocondrial. S'ha demostrat l'aplicabilitat d'aquest *microarray* en una població captiva, ja que els genotips obtinguts reflecteixen correctament la seva àmplia genealogia. A més, també s'han analitzat dues poblacions salvatges i els resultats suggereixen que el *microarray* pot ser una eina molt útil al costat dels microsatèl·lits, ja que aporta informació complementària sobre l'estructura poblacional i l'ecologia. El genotipatge d'SNPs amb la tecnologia del *microarray* és, doncs, una estratègia molt prometedora i pot esdevenir una eina essencial en genètica de la conservació per ajudar en la gestió i estudi de poblacions

captives i salvatges. A més, *microarrays* que combinessin SNPs de diferents regions genòmiques podrien arribar a desplaçar el genotipatge amb microsatèl·lits en el futur.

Aportació personal al treball

La meva tasca en aquest treball ha estat la relacionada amb el descobriment de nous SNPs en el cromosoma Y de ximpanzé, des del disseny dels encebadors a la realització de les anàlisis de DHPLC (fetes a la Universitat d'Uppsala, sota la supervisió d'A.C. Syvänen) i de seqüenciació. He dut a terme també l'obtenció i anàlisi de seqüències de regions mitocondrials de les bases de dades per detectar SNPs mitocondrials. D'altra banda, he realitzat la seqüenciació de la regió control del genoma mitocondrial, tant en les mostres de ximpanzé utilitzades per al descobriment dels SNPs com en els individus de la població salvatge. A més, he participat en les anàlisis inicials de variabilitat i he contribuït a la redacció de l'article.

A microarray system for Y-chromosomal and mitochondrial SNP analysis in chimpanzee populations

Olga Andrés[1, 2, #], Ann-Charlotte Rönn[3, #], Maxime Bonhomme[4], Thomas Kellermann[2,5], Brigitte Crouau-Roy[4] , Gaby Doxiadis[6], Ernst J Verschoor[7], Benoît Goossens[8], Xavier Domingo-Roura[1*], Michael W Bruford[8], Montserrat Bosch[1,2], Ann-Christine Syvänen[3,§] and the INPRIMAT consortium

[1] *Genètica de la Conservació Animal, Institut de Recerca i Tecnologia Agroalimentàries, Carretera de Cabrils Km2, 08348 Cabrils, Spain*
[2]*Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Carrer Doctor Aiguader 80, 08003 Barcelona, Spain*
[3]*Molecular Medicine, Department of Medical Sciences, Uppsala University, University Hospital, Entr. 70, 3rd floor, Res. Dep. 2, 75185 Uppsala, Sweden*
[4]*Evolution et Diversité Biologique UMR 5174 CNRS, Université Paul Sabatier Toulouse, 118 route de Narbonne, Bat IV R3 (b2), 31062 Toulouse cedex 4, France*
[5] *Institut für Immungenetik, Charité-Universitätsmedizin Berlin, Campus Virchow-Klinikum, Humboldt-Universität zu Berlin, Spandauer Damm 130, 14050 Berlin, Germany*
[6]*Department of Comparative Genetics and Refinement, Biomedical Primate Research Centre, Lange Kleiweg 139, 2288 GJ Rijswijk, The Netherlands*
[7]*Department of Virology, Biomedical Primate Research Centre, Lange Kleiweg 139, 2288 GJ Rijswijk, The Netherlands*
[8]*Cardiff School of Biosciences, Cardiff University, PO Box 915 Cathays Park, Cardiff CF10 3TL, UK*

Abstract

Chimpanzee populations are diminishing as a consequence of human activities, and as a result this species is now endangered. In the context of conservation programs, genetic data can add vital information, for instance on the genetic diversity and structure of threatened populations. Single nucleotide polymorphisms (SNPs) are biallelic markers that are widely used in human molecular studies and can be implemented in efficient microarray systems. This technology offers the potential of robust, multiplexed SNP genotyping at low reagent cost in other organisms than humans, but it is not commonly used yet in wild population studies. Here we describe the characterization of new SNPs in Y-chromosomal intronic regions in chimpanzees and also identify SNPs from mitochondrial genes, with the aim of developing a microarray system that permits the simultaneous study of both paternal and maternal lineages. Our system consists of 42 SNPs for the Y-chromosome and 45 SNPs for the mitochondrial genome. We demonstrate the applicability of this microarray in a captive population where genotypes accurately reflected its large pedigree. Two wild-living populations were also analyzed and the results show that the microarray will be a useful tool alongside microsatellite markers, since it supplies complementary information about population structure and ecology. SNP genotyping using microarray technology, therefore, is a promising approach and may become an essential tool in conservation genetics to help in the management and study of captive and wild-living populations. Moreover, microarrays that combine SNPs from different genomic regions could replace microsatellite typing in the future.

# These two authors contributed equally to the study
+This paper is dedicated to the memory of Dr. Xavier Domingo-Roura who passed away in November 2005
§ Correspondence: Ann-Christine Syvänen. Fax: +46 18 553601. E-mail: Ann-Christine.Syvanen@medsci.uu.se

Keywords: microarray, minisequencing, chimpanzee, sex linked SNPs, population genetics, conservation genetics.

Introduction

Chimpanzee (*Pan troglodytes*) is a species that, as many other primates, requires ever-increasing conservation measures, since wild-living populations are diminishing as a consequence of human activities (Walsh *et al.* 2003). Logging and fragmentation are altering and reducing chimpanzees' habitat; bushmeat trade and trade of infants to be used as pets lead to chimpanzee poaching; and human diseases contaminate chimpanzee populations. Conservation programs are therefore needed to protect this species and other great apes, as recently recognized by UNEP, through its Great Ape Survival Project (GRASP: http://www.unep.org/grasp/). Genetic analyses in the chimpanzee are an invaluable support in conservation programs as they allow detection of variation of wild or captive populations and identification of individuals by genetic tests, which is basic, for example, in breeding management, translocation programs, or subspecies recognition (Goossens *et al.* 2002; Goossens *et al.* 2005). In addition, the chimpanzee is the closest living relative to the human species, and therefore of special interests for understanding human evolution, biology and genetics: as a result its preliminary genome was recently published (Mikkelsen *et al.* 2005).

Single Nucleotide Polymorphisms (SNPs) are ideal as genetic markers owing to their abundance and random genomic distribution, low mutation rate and the possibility of incorporating them into robust genotyping assays with high multiplexing levels (for a review, see Syvänen 2001). Consequently, SNPs have now become the most widely used markers in human disease genetic and population genetic studies (see e.g. Nielsen 2004; Suh & Vijg 2005). SNPs have the potential to become the markers of choice also for studies of evolution, population ecology, and conservation of wildlife species (for a review, see Morin *et al.* 2004), and indeed in the recent past, more SNP-based population studies in a variety of non-model organisms are appearing in the literature (e.g. Gilchrist *et al.* 2006; Smith *et al.* 2005).

However, so far genetic studies of the chimpanzee have been based on sequence analysis of the hypervariable region of mitochondrial DNA (mtDNA) (e. g. Deinard & Kidd, 1999; Yu *et al.* 2003), and on genotyping nuclear short tandem repeat markers amplified using primers from human sequences (e. g. Goossens *et al.* 2002; Gusmao *et al.* 2002). Only a few studies of limited scale have applied genotyping of Y-chromosomal and autosomal SNPs in chimpanzees to date (Smith *et al.* 2004; Stone *et al.* 2002). One reason for this was probably that the scarce availability of genomic sequence information from the chimpanzee until recently hampered the development of assays for identification and genotyping of large numbers of SNPs (Aitken *et al.* 2004; Hellborg & Ellegren 2003).

In this study we established a multiplexed genotyping system for 45 SNPs in the coding region of the chimpanzee mtDNA and for 42 SNPs in the male-specific region of the Y-chromosome (MSY). As neither mtDNA nor MSY recombine and both are inherited uniparentally our genotyping system allows separate determination of both the maternal and paternal lineages in chimpanzee populations based on mitochondrial and Y-chromosomal haplotypes. The genotyping system is based on fluorescent minisequencing primer extension using "tagged" primers and fluorescent nucleotides followed by capture of the reaction products on microarrays carrying complementary tag-sequences (Lindroos *et al.* 2002; Lovmar *et al.* 2003). Here we demonstrate the feasibility of the system by genotyping Y-chromosomal and mitochondrial SNPs in a large pedigree from a captive chimpanzee population and from individuals from two wild-living chimpanzee populations from Sub-Saharan Africa.

Materials and methods

Samples and DNA extraction

Tissue samples from 61 captive, unrelated male chimpanzees were obtained from ten zoos or research institutions (Supplementary Table 1) for the discovery of Y-chromosomal SNPs. DNA from these individuals was extracted from 100 µl of blood taken during routine veterinary analysis, from 25 mg of tissue from carcasses, or from ten rooted hairs, using the DNeasy Tissue Kit (Qiagen, Hilden, Germany). In some cases purified DNA was acquired from a primate centre. To avoid exhaustion of the DNA, the samples were subjected to whole genome amplification using the GenomiPhi DNA Amplification Kit (GE Healthcare, Uppsala, Sweden) as previously described (Rönn *et al.* 2006). In order to test the applicability of the microarray system, DNA samples from peripheral blood lymphocytes or immortalized B cells from members of a

chimpanzee pedigree with two sires, four dams and 12 offspring were provided by the Biomedical Primate Research Centre (Rijswijk, the Netherlands). Another set of 59 chimpanzee samples from wild-living populations were also analyzed. These chimpanzees were part of a release project of illegally poached animals, seized by the authorities of the Republic of Congo, originating from the regions of Cabinda (Angola), Kouilou and L´Ekoumou and Niari as described in detail by Goossens *et al.* (2002). From these samples, DNA was extracted from plucked hairs by buffer extraction as described by Vigilant (1999), or by the Chelex-100 extraction method (Walsh *et al.* 1991). In addition, DNA from one female chimpanzee and one human male were used as control samples.

Identification of Y-chromosomal SNPs

Fifteen primer pairs for amplification of introns of the jumonji, AT rich interactive domain 1D (JARID1D) gene and 22 primer pairs for amplification of introns of the Y-linked protein kinase (PRKY) gene were designed based on chimpanzee and human sequences for analysis by denaturing high performance liquid chromatography (DHPLC). The primer sequences and PCR fragment sizes are given in Supplementary Table 2. The fragments amplified by these primers cover a total of 7947 non-overlapping intronic sequences in the JARID1D and PRKY genes. To ensure specific Y-chromosomal amplification, the primers were positioned so that the two last nucleotides of the 3'-end of at least one primer of each pair differed from the paralogous X-chromosomal sequence. Two µl of a 1:50 dilution of the whole genome amplified material from 61 captive chimpanzees, with an approximate concentration of 25-50 ng/µl, were used for PCR amplification. The PCR reaction mixtures contained 0.17 µM primers, 0.32 mM dNTPs, 1.5 mM $MgCl_2$, and 0.85 U of Taq DNA polymerase (Ecogene) in 15µl of buffer. A touchdown PCR program described by Hellborg & Ellegren (2003) with a 0.5 ºC incremental decrease in elongation temperature from 65ºC to 55ºC during 20 cycles, followed by 20 cycles at 55ºC, was used for all primer pairs except one pair (SMCY14), for which the initial elongation temperature was 68ºC and the final temperature after 20 cycles was 58ºC.

The optimal temperature for DHPLC was selected for each PCR fragment using the DHPLC Melt Program from Stanford Genome Technology Center (http://insertion.stanford.edu/melt.html) (Oefner & Underhill 1998), and was verified experimentally. For each fragment, PCR products from 3 to 6 samples were pooled for analysis by DHPLC. An equal amount of PCR product was included in the pools according to visual determination of the DNA concentration of the fragments after separation by agarose gel electrophoresis. DHPLC was run for the fragment pools at the optimal temperature (Supplementary Table 2) on a Wave DNA Fragment Analysis 3500 System (Transgenomic, Omaha, NE, USA). SNPs were identified in a fragment when the elution profile from a pooled sample showed two or more peaks. Fragments containing putative SNPs were then amplified by PCR from 5 ng of the corresponding original genomic DNA sample at the conditions given above, and the PCR products were purified using a GFX PCR DNA and Gel Band Purification Kit (Amersham Biosciences UK Limited, Buckinghamshire, UK). Both DNA strands of each fragment were sequenced in a final volume of 10 µl with 10 ng of the purified DNA, 1.6 µM of forward or reverse primer and 2 µl of the reagent mixture from the ABI Prism™ Big Dye Terminator Cycle Sequencing 3.1 kit, followed by electrophoresis on an ABI 3100 automated DNA sequencer (Applied Biosystems, Foster City, CA, USA).

Identification of mitochondrial SNPs

A GenBank search of chimpanzee mtDNA sequences was performed to allow identification of SNPs in the 12S rRNA (MT-RNR1), 16S rRNA (MT-RNR2), NADH dehydrogenase 2 (MT-ND2), cytochrome c oxidase I (MT-CO1), cytochrome c oxidase II (MT-CO2), NADH dehydrogenase 5 (MT-ND5) and cytochrome b (MT-CYB) genes. The sequences were analysed using BioEdit version 6.0.7 (Hall 1999), and multiple alignments were performed using Clustal W (Thompson *et al.* 1994).

Table 1 Identified Y-chromosomal SNPs and haplotypes

PRKY															
	SNP[1]														
	alleles	G/-	C/T	C/A	T/C	A/-	G/A	G/A	A/-	A/G	C/G	T/C	C/T	G/A	C/T
	MAF[2]	0.082	0.082	0.033	0.049	0.082	0.016	0.033	0.393	0.033	0.115	0.016	0.230	0.033	0.082
Haplot.	N[3]	PY03	PY04	PY05	PY06.1	PY06.2	PY10	PY11-1	PY11-2	PY14	PY18	PY19.1	PY19.2	PY19.3	PY19.4
HP01	35	G	C	C	T	A	G	G	A	A	C	T	C	G	C
HP02	5	-	T	C	T	A	G	G	-	A	C	T	C	G	T
HP03	5	G	C	C	T	A	G	G	-	A	G	T	T	G	C
HP04	2	G	C	C	T	A	G	G	-	A	G	T	T	G	C
HP05	2	G	C	C	T	A	G	G	A	A	C	T	T	G	C
HP06	2	G	C	C	C	-	G	G	-	A	C	T	T	G	C
HP07	2	G	C	C	T	-	G	G	-	A	C	T	C	A	C
HP08	2	G	C	C	T	A	G	G	-	G	C	T	C	G	C
HP09	2	G	C	C	T	A	G	G	-	A	C	T	C	G	C
HP10	1	G	C	A	T	A	G	A	-	A	C	C	T	G	C
HP11	1	G	C	C	C	-	G	G	-	A	C	T	C	G	C
HP12	1	G	C	A	T	A	G	A	-	A	C	T	T	G	C
HP13	1	G	C	C	T	A	A	G	-	A	C	T	T	G	C

JARID1D										
	SNP									
	allele	T/G	T/G	C/T	A/G	A/G	C/T	G/A	C/T	T/C
	MAF[2]	0.082	0.033	0.082	0.164	0.082	0.082	0.066	0.098	0.164
Haplot.	N[3]	JY01	JY02	JY03	SMCY8.1	SMCY8.2	JY05.1	JY05.2	JY05.3	JY08
HJ01	36	T	T	C	A	A	C	G	C	T
HJ02	5	T	T	C	G	A	C	G	C	C
HJ03	5	T	T	T	A	A	T	G	C	T
HJ04	5	G	T	C	G	A	C	G	C	C
HJ05	5	T	T	C	A	G	C	G	T	T
HJ06	2	T	T	C	A	A	C	A	C	T
HJ07	2	T	G	C	A	A	C	A	C	T
HJ08	1	T	T	C	A	A	C	G	T	T

Genotyping Y-chromosomal and mitochondrial SNPs

Primers for multiplexed PCR were designed using the Autoprimer software (Beckman Coulter Inc., Fullerton, CA, USA) based on the chimpanzee sequence assembly available at Ensembl in March 2006 (release PanTro 1.0; http://www.ensembl.org). To ensure specificity for the Y-chromosome, the sequences of the Y-chromosomal fragments were aligned to their X-chromosomal paralogs. A 25-plex amplification reaction for the fragments containing 45 mitochondrial SNPs and a separate 17-plex amplification reaction for the fragments containing 42 Y-chromosomal SNPs were designed. Primers for single base extension ("minisequencing") were designed using the NetPrimer software (http://www.premierbiosoft.com/netprimer/) based on the chimpanzee sequence assembly. Minisequencing primers were designed to anneal immediately adjacent to the nucleotide position to be detected. The 5'-end of each minisequencing primer contained a unique 20 bp tag-sequence from the Affymetrix GeneChip® Tag collection (Affymetrix, Santa Clara, CA, USA) (Fan *et al.* 2000). Oligonucleotides complementary to the tag sequences were immobilized covalently on CodeLink™ Activated Slides (GE Healthcare, Uppsala, Sweden) in an "array-of-arrays" conformation (Pastinen *et al.* 2000) as described in Lovmar *et al.* (2003). This conformation allows detection of all the 87 SNPs originally included in the study in 80 samples simultaneously on each slide.

Oligonucleotides were synthesized by Integrated DNA Technologies (Coralville, IA, USA). The sequences of the PCR and minisequencing primers are provided in Supplementary Table 3.

The multiplex PCRs for Y-chromosomal and mitochondrial SNPs were performed in 15 µl reaction mixtures containing 2 - 4 µl of DNA extract, 50 nM primers, 100 µM dNTPs, 1.5 U of

Smart Taq Thermostable DNA Polymerase (Naxo, Tartu, Estonia), 10 mM Tris-HCl, 50 mM KCl, 0.8% Nonidet 40, 5 mM $MgCl_2$, and 7.5 µg bovine serum albumin. Individual PCRs were performed with all primer pairs for the Y-chromosomal fragments using female chimpanzee DNA to verify absence of X-chromosomal amplification. To remove remaining dNTPs and primers, 7 µl aliquots of the PCR-products were treated with 0.1 U/µl shrimp alkaline phosphatase (USB Corporation, Cleveland, OH, USA) and 0.5 U/µl exonuclease I (Fermentas International, Burlington, Canada) in 10.5 µl of 6.7 mM $MgCl_2$, 50 mM Tris–HCl pH 9.5 for 60 min at 37°C with subsequent deactivation of the enzymes at 95°C for 15 min. Cyclic minisequencing reactions were performed with 0.1 µM ddATP-Texas Red, ddCTP-Tamra and ddGTP-R110, 0.15 µM of ddUTP-Cy5 (Perkin-Elmer Life Sciences, Boston, MA, USA), 10 nM of each minisequencing primer, and 0.067 U/µl KlenThermase DNA Polymerase (GeneCraft GmbH, Lüdinghausen, Germany) in 15 µl of buffer containing 0.02% Triton-X100. The cyclic extension reaction was performed in a Thermal Cycler PTC-225 (MJ Research, Watertown, MA, USA) with initial denaturation at 96°C for 3 min followed by 55 cycles at 95°C and 55°C for 30 s each. Separate minisequencing reactions were performed using the multiplex PCR product from the Y-chromosome and the mtDNA as template, and the reaction products from the same DNA sample were pooled prior to hybridisation to the tag sequences on the microarray. The slides were scanned using the ScanArray® Express instrument (Perkin-Elmer Life Science), the fluorescence signal intensities were determined using the QuantArray® analysis 3.1 software (Perkin-Elmer Life Sciences), and the genotypes were called using scatter plots of the fluorescent signals. SNPs that gave successful genotype calls in >90% of the samples were accepted for further data analysis.

Sequencing of the mitochondrial D-loop

In samples from the wild-living chimpanzee population, 498 bp of the mitochondrial D-loop control region were also sequenced to analyse polymorphisms using primers described by Lacoste *et al.* (2001). The PCR reaction mixtures were the same as those for the amplification of Y-chromosomal fragments, and the PCR conditions consisted of an initial cycle at 94°C for 5 min followed by 30 cycles of 30 s each at 94°C, 55°C and 72°C, and a final extension at 72°C for 5 min.

Statistical analysis

The D-loop sequences were aligned using Clustal W (Thompson *et al.* 1994). For SNP and sequence data, number of haplotypes, gene diversity (i.e. heterozygosity $h = 1-\Sigma p_i^2$, Nei, 1987) and nucleotide diversity π (average number of nucleotide differences per site between any two DNA sequences chosen randomly from the sample population) were calculated using arlequin ver. 2.000 (Schneider *et al.* 2000). Population structure was estimated by analysis of molecular variance (AMOVA) and by pairwise F_{ST} (Reynolds *et al.* 1983; Slatkin 1995), using arlequin. The significance of the F_{ST} was tested using permutations. A mismatch analysis was conducted on the whole chimpanzee sample to test for the sudden expansion model (Harpending 1994; Rogers 1995). The validity of the estimated parameters (θ_0 and θ_1, the scaled mutation parameter before and after the expansion, and the expansion parameter τ) for the sudden expansion demographic model was tested using a bootstrap approach on the SSD (sum of squared deviations) between simulated and observed data, and by calculating the raggedness index. Tajima's *D* (Tajima 1989) and Fu's F_S (Fu 1997) were computed to test for neutrality, though they provide complementary tests for demographic inferences when applied to neutral markers. Significance was determined using permutation tests implemented in arlequin.

Results

Y-chromosomal and mitochondrial SNPs

To establish a genotyping system for Y-chromosomal SNPs in the chimpanzee, we first used DHPLC-analysis followed by sequencing to screen for SNPs in introns of the PRKY and JARID1D genes in the unrelated chimpanzees. This analysis identified 23 previously unknown SNPs –including three single nucleotide insertion/deletion polymorphisms–, 14 of which were located in the PKRY gene and 9 were in the JARID1D gene. The SNPs are distributed over 2 introns of *PRKY* and 7 introns in *JARID1D*. Their exact genomic positions are defined by the minisequencing primers provided in Supplementary Table 3. Table 1 shows the SNPs,

their allele frequencies and haplotype designations. Thirteen and eight different haplotypes were defined for the PRKY and JARID1D genes, respectively. In both genes, the SNPs defined one major haplotype, with a frequency of 0.57 in *PRKY*, and a frequency of 0.59 in *JARID1D*. Four of the haplotypes in the PRKY gene and one of the haplotypes in the JARID1D gene were found in a single individual only (f = 0.016). In addition to these SNPs, we included 19 Y-chromosmal SNPs previously described by Stone *et al.* (2002) in the Y-chromosomal SNP panel. The mitochondrial SNPs to be included in the panel for genotyping, identified by comparisons of published sequences, produced 45 SNPs with allele frequencies > 0.07 that were flanked by conserved sequences to allow primer design (Supplementary Table 4).

Genotyping

Multiplex PCRs and tag-array minisequencing assays were designed for the initially selected SNP panel that comprised 42 Y-chromosomal and 45 mitochondrial SNPs. The assays were successful for 9 Y-chromosomal SNPs in both DNA polarities and for 26 SNPs in one DNA polarity (see Supplementary Table 3), so that a total of 35 SNPs could be typed. The main reason for assay failures was background noise preventing SNP genotype assignment, and these assays were discarded. Successful genotyping assays were established for 37 out of 45 of the mitochondrial SNPs. Figure 1 shows four scatter plots of the fluorescence signals measured from the microarrays used for calling genotypes. As can be seen, the clustered signals are clearly distinguished from background in negative water control reactions both for the Y-chromosomal and mitochondrial SNPs.

The performance of the genotyping system was validated by genotyping the 35 functional Y-chromosomal SNPs and 37 mitochondrial SNPs in four chimpanzee individuals out of the 61 individuals originally used to identify the Y-chromosomal SNPs. All SNPs were typed successfully and all genotypes were correctly assigned.

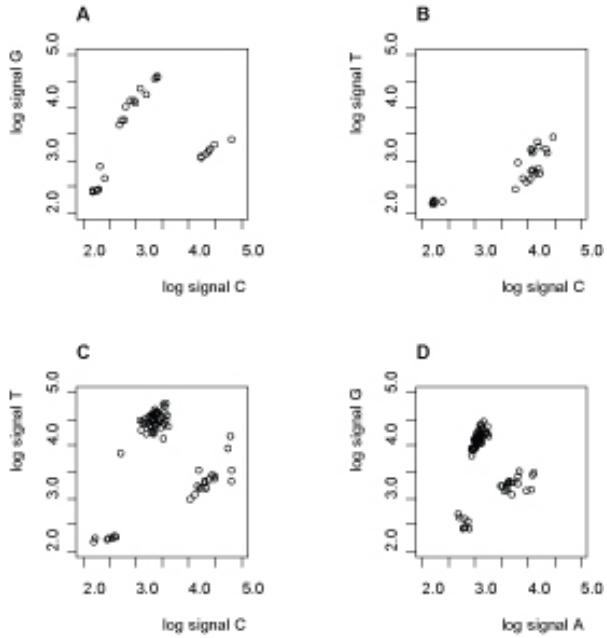

Fig 1. Examples of scatter plots of fluorescence signals from tag-array minisequencing used for genotyping Y-chromosomal (A, B) and mitochondrial (C, D) SNPs. A) SNP sY19 (defined in Stone et al. (2002) in 19 male samples, average signal/noise ratio (s/n) 150 for the C-allele and 51 for the G-allele; B) SNP PY19.3, that is monomorphic in 19 samples with s/n 82 for the G-allele, detected in reverse polarity as the C-signal. C) SNP 956 in 69 samples, s/n 104 for the C-allele and 184 for the T-allele. D) SNP 7188 in 69 samples, s/n 10 for the A-allele and 41 for the G- allele.

Applicability of the microarray on a captive population

To test a possible application of the microarray in the management of captive chimpanzee populations, a large chimpanzee pedigree from a primate research center was genotyped (Table 2). All males in the pedigree had the same Y-chromosomal haplotype formed by the major alleles of 33 of the SNPs and the minor alleles of PY11-2 (Table 1) and sY65 (defined by Stone et al. 2002). The mitochondrial haplotypes followed correct maternal inheritance in the pedigree (Figure 2).

Table 2 Mitochondrial haplotypes observed in the chimpanzee pedigree

		SNP position	721	737	833	956	975	1117	1419	1650	1663	3927	4147	4193	4301	4342	4681	4858	4892	5395	5539
Haplotype	N*	SNP alleles	AG	GA	GA	CT	TC	CT	CT	AG	CT	TC	CT	GA	TC	AG	TC	CT	GA	AG	TC
MT-H1	8		A	A	G	C	T	C	C	A	C	T	C	G	T	A	T	C	G	A	T
MT-H2	7		A	G	G	C	T	C	C	G	C	C	C	G	T	A	T	C	G	A	T
MT-H3	2		A	G	G	C	T	C	C	A	C	T	C	G	T	A	T	C	G	A	T
MT-H4	1		A	A	A	C	T	C	C	A	C	T	C	A	T	A	T	C	G	A	T

		SNP position	5614	5767	5848	6016	7188	11952	12135	12255	14173	14269	14392	14467	14515	14551	14728	14857	15088	15163
Haplotype	N*	SNP alleles	C/T	C/T	G/A	A/G	A/G	G/A	C/T	G/A	A/G	AC	T/C	C/T	C/T	G/A	T/C	C/T	T/C	T/C
MT-H1	8		T	T	G	A	A	G	C	G	A	A	T	C	T	G	T	C	T	T
MT-H2	7		C	C	G	A	A	A	T	A	G	A	T	C	T	G	C	C	T	C
MT-H3	2		C	C	G	A	A	A	T	A	G	A	T	C	T	G	C	C	T	C
MT-H4	1		T	T	G	A	A	G	C	G	A	A	T	C	T	G	T	C	T	C

*Number of samples

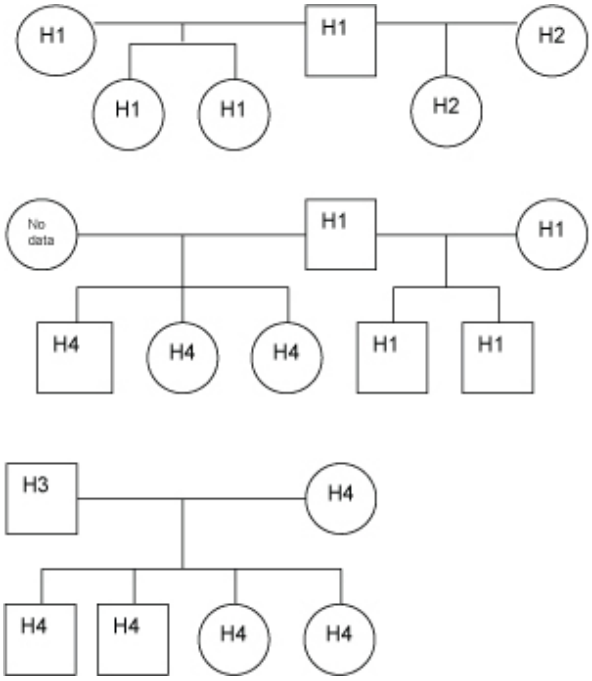

Fig. 2 Chimpanzee pedigree with mitochondrial haplotypes. The haplotypes are shown in Table 2.

Applicability of the microarray on wild-living populations

The 72 SNPs implemented on the microarray were genotyped in 59 DNA samples from wild-living chimpanzee populations from the Counkouati Reserve in Congo (Goossens *et al.* 2002) to demonstrate the value of its use in wild population studies and management. Taking into account the success rate requirement of genotype calling, all 37 mitochondrial SNPs and 28 of the Y-chromosomal SNPs fulfilled this requirement in 51 samples; eight samples were discarded due to poor genotyping success. The genotyping accuracy for the accepted genotypes was > 99.9% based on repeated genotyping of all samples and all SNPs in independent experiments.

Further statistical analysis was carried out for the three different marker types – D-loop sequences, mtDNA SNPs and Y-chromosomal SNPs – to study the wild-living chimpanzee population. Genetic diversity and haplotypic divergence (h and π) was higher for maternal lineages with the exception of the haplotypic divergence of mtDNA SNPs (Table 3), indicating either a higher N_e for female or variance in reproductive success in males. Tajima's D and Fu's F_S values from neutrality tests were not statistically different from 0, indicating neutrality or population stability in light of demographic history. The exception, however, was a slight signal of expansion with mtDNA D-loop data when using Fu's test, which is sensitive to population expansion (Table 3).

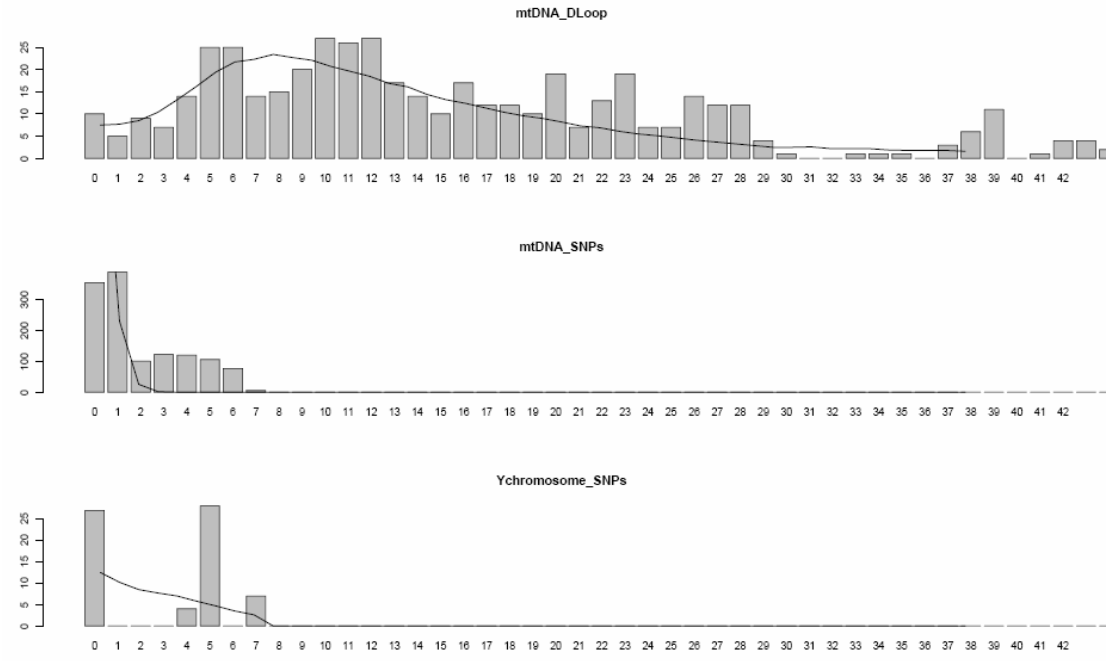

Fig 3 Mismatch distributions in 51 *Pan troglodytes* samples, of which 12 were male, for A) the mitochondrial D-loop, B) mitochondrial SNPs and C) Y-chromosomal SNPs.

Genetic structure was estimated between chimpanzees from Kouilou, including the Conkouati Reserve, and the adjacent regions of L'Ekoumou, Niari and Cabinda (see Goossens *et al.* 2002). For all three markers, F_{ST} values were not significant and nearly all of the genetic variation was found within populations, indicating very limited genetic structure (Table 3). Since little genetic differentiation was found for Y-chromosomal and mtDNA markers, one could infer that the chimpanzee sample belongs to a population characterized by high levels of gene-flow. However, the Y-chromosomal sample size is small –only 12 males were analysed (7 from Kouilou, 1 from Niari, 1 from DRC, and 3 from unknown origin)–, and sampling bias may therefore obscure the true demographic structure for male chromosomes in these populations.

The observed mismatch distributions are shown together with their expected shape under the sudden expansion model (Figure 3). For the mtDNA D-loop, the p-values of SSDs and raggedness index were not significant, and the mismatch distribution was unimodal, indicating an expansion of the female population, and then a high effective population size. For the Y-chromsome SNP dataset, the bimodal mismatch distribution and the significant p-values of the SSD and raggedness index seemed to reflect a more complex demographic history. However, for the mtDNA SNP dataset, the results were ambiguous: test of the SSD did not suggest an expansion while the raggedness index (not different from 0) did not suggest a stationary population (Supplementary Table 5). The number and polymorphism of the mitochondrial-SNPs tested here may not be high enough to allow inference of a precise population history, unlike the highly variable mtDNA D-loop.

Table 3 Genetic diversity, neutrality tests and analysis of molecular variance (AMOVA) in the wild-living chimpanzee populations

Markers	N_{ind}	N_{hap}	N_{sites}	h (s.d.)	π (s.d)	D (p-value)	F_S (p-value)	F_{ST} (p-value)	Percent Variation	
									Among popul.	Within popul.
Y- SNPs	12	4	8	0.68 (0.10)	3.11 (1.73)	0.69 (0.24)	2.08 (0.13)	-0.002 (p=0.56)	-0.15	100.15
mtDNA SNPs	51	9	9	0.75 (0.04)	1.93 (1.12)	-0.09 (0.53)	-1.26 (0.30)	-0.02 (p=0.73)	-2.39	102.39
D-loop sequence	31	25	82	0.98 (0.02)	15.49 (7.11)	-0.90 (0.18)	-6.13 (0.03)	0.01 (p=0.25)	1.14	98.86

N_{ind} = number of individuals; N_{hap} = number of haplotypes; N_{sites} = number of polymorphic sites
h = gene diversity; s.d. = standard deviation;π = mean number of pairwise nucleotide differences; D = Tajima's D; F_S= Fu's F_S.
Population structure: comparison between Kouilou(+Conkouati Reserve) and the other locations (L'Ekoumou, Niari, Cabinda, DRC, and unknown origin).

Discussion

Genetic markers for population studies should ideally allow data acquisition of many loci scored in large population samples, and genotyping of these markers should be technically easy to perform with high reproducibility. Moreover, the possibility of applying these markers to non-invasively collected samples, which typically contain low amounts of degraded DNA, is crucial for wild population studies. SNP genotyping using tag-array minisequencing meets these demands, as SNPs are easily scored with high accuracy and allow the use of very short DNA fragments.

The minisequencing microarray system developed in this study included SNPs from Y-chromosomal intronic regions and mitochondrial genes, and thus allowed genotyping of both paternal and maternal lineages simultaneously in the same experiments. This is especially useful when analyzing species in the wild that have different male and female behavior patterns, since they may present sex differences in genetic structure (e.g. Hammond *et al.* 2006). This information cannot be directly obtained simultaneous by other methods, such as genotyping microsatellites or sequencing nuclear genes. Moreover, as the SNPs in this study were from haploid genomic regions, the microarray did not suffer from the technical difficulties with scoring microsatellites, such as null alleles and allelic dropout, which may require many DNA-exhausting repetitive experiments for accurate results,

especially when working with non-invasively collected samples.

The microarray developed in this study performed accurately in the analysis of a chimpanzee captive population in that it was possible to obtain the real pedigree of the population. This demonstrates the potential of this microarray system in captive population management since it allows pedigree reconstruction and paternity testing. Moreover, this tool could also give support to other captive areas, such as the design of breeding plans and the study of translocation programs, for which the analysis of genetic variability is crucial.

Chimpanzees are primates with a classical multimale/multifemale social system; males usually remain in the community of origin, while females disperse once they reach adulthood (Morin *et al.* 1994). These sexual differences in the migration pattern should be reflected in a differentiation of the genetic structure between sexes in wild populations. We applied the microarray and carried out sequencing of the mitochondrial D-loop to study individuals from a wild chimpanzee population. Results of the population structure and demographic history analyses were similar to the results obtained by Goossens *et al.* (2002) in the same samples using microsatellite markers. They found F_{ST} values of 0.004 (p=0.027, a significant if very small value), and inferred a high migration rate in these populations. Both mitochondrial DNA D-loop sequences and mitochondrial gene SNPs corroborated these observations since lower differentiation for mitochondrial (female) markers than for microsatellites or Y-chromosomal DNA was detected, and females are known to be the migrating sex. However, we found a surprising lack of genetic structure for lineages belonging to males, which are philopatric in chimpanzees (Morin *et al.* 1994), although this is possibly due to sampling bias. The high genetic diversity and population expansion pattern for mtDNA data compared to a small genetic diversity for Y-chromosome data may be due to a higher female than male effective population size, as found in Stone *et al.* (2002), perhaps due to female-biased dispersal or, less likely, variance in reproductive success in males.

The microarray system was accurate for the captive population study. For the wild population study, however, the system presented some limitations. On the one hand, the sampling may not have allowed a thorough detection of male geographic subdivision using male specific genetic data due to the low number of male samples; this is a constraint that may be present when studying wild-living populations and cannot always be overcome. On the other hand, the number of SNPs implemented in the microarray might have been too low. Depending on allele frequency, 30-50 independently inherited autosomal SNPs should be sufficient for human individual identification (Chakraborty *et al.* 1999). However, we must take into account that the SNPs analyzed here are linked, and do not segregate independently, since mitochondrial DNA and Y-chromosome do not recombine, and thus a greater number of markers would be needed. In order to improve the applicability of our microarray system on wild-living population studies, therefore, the system could be ameliorated by adding more SNPs. Moreover, the lower success rate for the Y-chromosomal SNPs than the mitochondrial SNPs in hair samples is likely to be due to degradation of the DNA in the non-invasively collected samples. This limitation could perhaps be solved by re-designing some of the longer PCR fragments (>200 bp) so that degraded DNA could be genotyped for all the SNPs.

The results of the wild population study show that this microarray may be a helpful complementary tool to microsatellites, as it provides complementary information about population structure and ecology. With the addition of autosomal SNPs to the SNPs from mitochondrial genes and Y-chromosome introns, an enhanced microarray assay would permit the analyses of a population genetic variation in a wider range of DNA regions from the same experiments. The microarray technology, therefore, could replace microsatellite typing in the future and SNPs may become a marker of choice due to their many advantages.

Acknowledgements

We thank Raul Figueroa for technical assistance. This study was funded by the European Commission (INPRIMAT Consortium, contract QLRI-CT-2002-01325), the Knut and Alice Wallenberg Foundation (to A-C.S.), The Swedish Research Council for Science and Technology (to A-C.S), the DURSI of the Generalitat de Catalunya (to O.A. with ref. 2002FI-00280), the German Academic Exchange Service (T.K.). Six

chimpanzee DNA samples were made available by Dr P.A. Morin and Dr. S. Pääbo of the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, from cells obtained from *Pan t. troglodytes* held at the Centre International de Recherches Médicales de Franceville, Gabon and donated by Dr E. Jean Wickings, Unité de Génétique des Ecosystèmes tropicaux, CIRMF. We also thank H. Zischler, M. Rocchi, Worclaw Zoo, for providing chimpanzee samples, and the NGO HELP Congo, for providing access to the Congolese chimpanzee samples.

References

Aitken N, Smith S, Schwarz C, Morin PA (2004) Single nucleotide polymorphism (SNP) discovery in mammals: a targeted-gene approach. *Molecular Ecology,* **13**, 1423-1431.

Chakraborty R, Stivers DN, Su B, Zhong Y, Budowle B (1999) The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis,* **20**, 1682-1696.

Deinard A, Kidd K (1999) Evolution of a HOXB6 intergenic region within the great apes and humans. *Journal of Human Evolution* **36**, 687-703.

Fan JB, Chen X, Halushka MK*, et al.* (2000) Parallel genotyping of human SNPs using generic high-density oligonucleotide tag arrays. *Genome Research,* **10**, 853-860.

Fischer A, Pollack J, Thalmann O, Nickel B, Paabo S (2006) Demographic history and genetic differentiation in apes. *Current Biology,* **16**, 1133-1138.

Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics,* **147**, 915-925.

Gilchrist EJ, Haughn GW, Ying CC*, et al.* (2006) Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of Populus trichocarpa. *Molecular Ecology,* **15**, 1367-1378.

Goossens B, Funk SM, Vidal C*, et al.* (2002) Measuring genetic diversity in translocation programmes: principles and application to a chimpanzee release project. *Animal Conservation,* **5**, 225-236.

Goossens B, Setchell JM, Tchidongo E*, et al.* (2005) Survival, interactions with conspecifics and reproduction in 37 chimpanzees released into the wild. *Biological Conservation,* **123**, 461-475.

Gusmao L, Gonzalez-Neira A, Alves C*, et al.* (2002) Genetic diversity of Y-specific STRs in chimpanzees (*Pan troglodytes*). *American Journal of Primatology,* **57**, 21-29.

Hall TA (1999) BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/97/NT. *Nucleic Acids Symposium Series,* **41**, 95-98.

Hammond RL, Handley LJ, Winney BJ, Bruford MW, Perrin N (2006) Genetic evidence for female-biased dispersal and gene flow in a polygynous primate. *Proceedings of the Royal Society B: Biological Sciences,* **273**, 479-484.

Harpending HC (1994) Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. *Human Biology,* **66**, 591-600.

Hellborg L, Ellegren H (2003) Y chromosome conserved anchored tagged sequences (YCATS) for the analysis of mammalian male-specific DNA. *Molecular Ecology,* **12**, 283-291.

Hughes JF, Skaletsky H, Pyntikova T*, et al.* (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature,* **437**, 100-103.

Kuroki Y, Toyoda A, Noguchi H*, et al.* (2006) Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nature Genetics,* **38**, 158-167.

Lacoste V, Mauclere P, Dubreuil G*, et al.* (2001) A novel gamma 2-herpesvirus of the Rhadinovirus 2 lineage in chimpanzees. *Genome Research,* **11**, 1511-1519.

Lindroos K, Sigurdsson S, Johansson K, Ronnblom L, Syvanen AC (2002) Multiplex SNP genotyping in pooled DNA samples by a four-colour microarray system. *Nucleic Acids Research,* **30**, e70.

Lovmar L, Fredriksson M, Liljedahl U, Sigurdsson S, Syvanen AC (2003) Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. *Nucleic Acids Research,* **31**, e129.

Mikkelsen TS, Hillier LW, Eichler EE, the Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature,* **437**, 69-87.

Morin PA, Luikart G, Wayne RK, the SNP workshop group (2004) SNPs in ecology,

evolution and conservation *Trends in Ecology and Evolution,* **19**, 208-216.
Morin PA, Moore JJ, Chakraborty R*, et al.* (1994) Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science,* **265**, 1193-1201.
Nei M (1987) *Molecular Evolutionary Genetics* Columbia University Press, New York.
Nielsen R (2004) Population genetic analysis of ascertained SNP data. *Human Genomics,* **1**, 218-224.
Oefner PJ, Underhill PA (1998) DNA mutation detection using denaturing high-performance liquid chromatography (DHPLC). In: *Current Protocols in Human Genetics*, Supplement 19, 17.10.11-17.10.12. Wiley & Sons, New York.
Pastinen T, Raitio M, Lindroos K*, et al.* (2000) A system for specific, high-throughput genotyping by allele-specific primer extension on microarrays. *Genome Research,* **10**, 1031-1042.
Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. *Genetics,* **105**, 767-779.
Rogers AR (1995) Genetic evidence for a Pleistocene population explosion. *Evolution,* **49**, 608-615.
Ronn A-C, Andrés O, Bruford MW*, et al.* (2006) Multiple displacement amplification for generating an unlimited source of DNA for genotyping in nonhuman primate species *International Journal of Primatology,* **27**, 1145-1169.
Schneider S, Roessli D, Excoffier L (2000) ARLEQUIN ver. 2.000: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva, Switzerland.
Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics,* **139**, 457-462.
Smith CT, Elfstrom CM, Seeb LW, Seeb JE (2005) Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. *Molecular Ecology,* **14**, 4193-4203.
Smith S, Aitken N, Schwarz C, Morin PA (2004) Characterization of 15 single nucleotide polymorphism markers for chimpanzee (Pan troglodytes). *Molecular Ecology Notes,* **4**, 348-351.
Stone AC, Griffiths RC, Zegura SL, Hammer MF (2002) High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proceedings of the National Academy of Sciences of the United States of America,* **99**, 43-48.
Suh Y, Vijg J (2005) SNP discovery in associating genetic variation with human disease phenotypes. *Mutation Research,* **573**, 41-53.
Syvanen AC (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics,* **2**, 930-942.
Syvanen AC (2005) Toward genome-wide SNP genotyping. *Nature Genetics,* **37 Suppl**, S5-10.
Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* **123**, 585-595.
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 Research,* **22**, 4673-4680.
Walsh PD, Abernethy KA, Bermejo M*, et al.* (2003) Catastrophic ape decline in western equatorial Africa. *Nature,* **422**, 611-614.
Walsh PS, Metzger DA, Higuchi R (1991) Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques,* **10**, 506-513.
Vigilant L (1999) An evaluation of techniques for the extraction and amplification of DNA from naturally shed hairs. *Biological Chemistry,* **380**, 1329-1331.
Yu N, Jensen-Seaman MI, Chemnick L*, et al.* (2003) Low nucleotide diversity in chimpanzees and bonobos. *Genetics*,**164**, 1511-1518.

Supplementary Table 1 *Pan troglodytes* samples used for detecting Y-chromosomal SNPs.

Sample provider	No of samples	DNA source
Yerkes Primate Center, USA	14	Blood
Biomedical Primate Research Centre, the Netherlands	13	PBMC*
Max Planck Institute, Germany	8	B-cell line
Zoo, Germany	6	Liver and muscle tissue**
Zoo Madrid, Spain	5	Hair
Worclaw Zoo, Poland	4	Hair
Biomedical Primate Research Centre, the Netherlands	2	Serum
Biomedical Primate Research Centre, the Netherlands	2	Serum
London Zoo,	2	Blood
Parc de les Aus, Spain	2	Hair
Stanford University, USA	2	Cell line
Budapest Zoo, Hungary	1	Cell line
Universitá degli Studi di Bari, Italy	1	Cell line
Universitá degli Studi di Bari, Italy	1	Cell line***
Universitat Autònoma de Barcelona, Spain	1	Brain

*PBMC, Peripheral blood mononuclear cells

**Liver and muscle tissue were obtained during diagnostic necropsies

***Female sample

Supplementary Table 2 PCR primers, fragment size and optimal running temperature for denaturing high performance liquid chromatoghaphy

Fragment	Forward primer 5' – 3'	Reverse primer 5' – 3'	Fragment size (bp)	DHPLC T (ºC)
SMCY8*	TTGGATCTCAATCTAGCAGT	CGCTCAGCCTCCGTGACG	384	58
SMCY14*	TGCTGGGGCTGTCTGCA	CTTCTCCTTCTGTTCCCCT	267	60
JY01	TGCGAAAATAAGGCCCATAG	GGAATGCGGCGATCTGATAC	240	59
JY02	TTCTGACCCGAACAAGCTG	CAAAGTAAGAAAACACCAGTCCCA	262	56
JY03	AGCTTGTGTGTTAACCCCTG	ACAGCCCAGGCTTATTTCAG	223	56
JY04	CGTAGACCATTTTGGTTTGCT	GCCATTATCTCTTCCCACTCC	271	57
JY05	CATCAGATTGGGGATGTCAAG	AGACAGAGATGGTATGGGAACA	191	57
JY06	AACAGGTTCTTTCCTGGAAGTG	GAAATCAACCAACCAACCAATC	178	57
JY07	TAAGACACTGGCTCTTGGGAAT	GGCCCTCTGTTTACCTAGAATG	180	56
JY08	AAAGTGGGTTTGGAAAAAGAAA	TCCACCTGTTCCAGGACATC	247	55
JY09	CCTAGTCCCTTTGCTCCATTC	GGAGCCATGTTCTCAGGACT	233	58
JY10	GCTAGAAGCTGTGTATGAGTTAGGG	TTGGACTTAAAGGTGGCTTGAC	200	55
JY11	GCTGGTCAGTGGTCAGGTG	TTGCTGTTGTTTTGAGATGGA	229	60
JY12	GCAGACATCAGAATTTCCACAA	ATGCAGCAATGACAAAACCA	232	55
JY13	TGAAACTTGCAGGACAGGAA	AAACCCCCATGACACAAATG	248	55
PY01	CGAGAGGTGGCGGAGGATAC	GTGCGCTCCAGTTCTCTGAG	298	66
PY02	CCAGCGGCAGGGTCCTT	TGCCCATTCGAGGTTAAGTT	245	60
PY03	CCGTTTCCTTTGTCGTCTGT	GACAAACAGGAGTTAAGAGGGTT	310	60
PY04	TCAAGAAAATAAATGGGAGAGC	GTCCCACTCTGTCCTTCCAA	298	61
PY05	CCACGAGGAAAGCCTGAGT	GAGCAAGACAGTAAGGGAACAGAC	204	59
PY06	TGGAAATACAACTCGACAATGG	CCCCTGTCACACGAGAGCT	192	59 & 64
PY07	CGGCAGAGGTGAGGAGC	GGTCCTGGTCTCCTGTTGC	279	60
PY08	CCCAAGGGTTTTAGAGTCTTCA	CTGCATACAGATCCCCTCCA	293	59
PY09	CCCTCATTCCTAATTCTCTCAG	TTTCAGTAGAGATGGGGTTTTG	259	59
PY10	ACGTGGTGGCTCATGCTG	TGCAGTGCAATGCTGTGATT	249	62
PY11	CTTTGGATCAGAACAGCCATATG	GCCAACATGCCCAACTACCTAT	297	55
PY12	CAATCTTCCCACTTAAAATGACTG	AAGCTCACGACACCATACCC	228	56 & 61
PY13	GCCACTGCTTGTTTCTCATG	GTACGGGTCCAGGAACTTTG	214	57
PY14	TGCATTACCCAGGCTGATCT	CCCAGGCAATATGTTTACCCA	217	59
PY15	GATTCTCAGCTTTTAAAGTCATGC	GGCACTCTTTTGATTCACCA	294	56
PY16	GGTGCCAGGATCATAAATAAAGTT	ATTGGGGCAGAAGTGAGCTA	298	58
PY17	CCAAACACCAAAAGAAAAACAA	GGCAGGCAGTGATGTGTAAA	192	56
PY18	AGGGGCAGCATGTCAAAATA	CCCATCTGATTGGAAATTTAACC	264	58
PY19	CTGGTTAAATTTCCAATCAGATG	AACAATCCAATACCCATGACTG	211	60
PY20	CAACCCAGCCTTGCATAAACT	TTCTGTCAACCACGAGCAGT	230	58
PY21	TTCCACCTTTGGCTTTTGTT	GAATTCTATCAAATCCGGGAATAC	211	54
PY22	TCATGGACGTTTTGTAGGCTG	GCTTCCACGACTGGGTCTG	271	58

*PCR primer sequences are from Hellborg *et al*. (2003)

Supplementary Table 3 PCR and minisequencing primers for the Y-chromosomal SNPs.

Fragment	Forward PRC primer 5' -3'	Reverse PCR primer 5' - 3'	SNP	Forward strand minisequencing primer	Reverse strand minisequencing primer
PY03	TTTCGAGTTCTCAGTGCGTT	TTTAAATCCAGCGCCCAC	PY03	ACGCGCTGGTTTATGGGG	
PY04	TTGGGTGTGTTGGTGGGG	TAAGGTGTGACCTTAAGCCCT	PY04	AAACCCTCTTAACTCCTGTTTGTC	
PY05	CTCCCTCCTCCTATCTCATTCT	AAAGCAGAGAGAAGGCAGG	PY05		AGAAATGCAGATACACGAACAGG
PY06	TTGATGAGGTCCAGCCGT	TAAGCTGACTTGATGTCAAACTG	PY06.1	GGGCACCCACATGGGTCC	
			PY06.2	GAGCAGGATGCGGGCAGG	
PY11-1	ATTTAATAGTTTTAGGAACTTTGGATCA	TTTTCCTTTTTTAAAGCACAATTG	PY11-1	AATTTCTTGTTTATATAACCAAGCC	
PY11-2	TAGTCCCAGGTATTTGGGAAG	TAACTAAGAGCACAGGTATGCG	PY11-2	CGACCCTCATCTCTTAAAAAAAAA	
PY14	TGTCCCTGGTTGTCTGTGA	TTTACCCAGTCAAATGTTAAAAGC	PY14	GCTAGAATTTTGATTCTCAGCTTTT	TGCAGGAACCAGCATGACTT
PY18	GGTAAGTACTGTAAAGGTTTCTAACTTTTC	CTAAGGCTGGGATTTTGTAACC	PY18	GATGTTTATATTCCCCACCCTACT	
PY19	TAATTCAGAAAGCACGGCA	AACAATCCAATACCCATGACTG	PY19.1	GAGATTCACAATCTAGATGGTGGAG	
			PY19.2		AGCCAGGTGATGTCCCAGG
			PY19.3**		GGTGTTAGCATCTAGGCTCTCC
			PY19.3**		CAGTGTTAGCATCTAGGCTCTCC
			PY19.4		AACAATCCAATACCCATGACTGC
JY01	TTTGAAACAAAATTTGGGAATG	CAGCTTGTTCGGGTCAGA	JY01	TTTACACTCATGACCAGAAACCCT	AAAGTTAAATGAGAAATTTGGTGCT
JY02	TGTAGGGACGAAAATATAAGATTCA	AAGCTTATATAACTTGCTTCTGTCATAG	JY02	TATAGTTTAGGCATACACAAGCTTT	CTATTTTTTTGCGGTTAAATGTAAT
JY03	TTAGATGTATGTAGCAAATATAGCAAA	ATAGCAACCCAAAATGGTCTAC	JY03	GGAGTGATGTGGCATCCTTTG	
JY05	TAAGAGGGGCCTGGAAAG	AGTTTTTAATTATTTCAGTATGGAAGGA	JY05.1	CTGGAAAGGTGGAATTTTTGTTAA	
			JY05.2**	TTAACAGCAAATACTCAGGAAGTTT	GGTTCTCAAGTTTTCCTGACATGT
			JY05.2**	TTAATAGCAAATACTCAGGAAGTTT	
			JY05.3	AGTGGCGGAGATATGTGTTCC	CATGCTGTAAAGACAGAGATGGTAT
JY08	TTACTAGATGGATAAGATTGTATTAAAGTGG	TATTGCTTAACACTTATTCTCAACCA	JY08		ACAAAGGCCCTCTGTTTACCTAG
SMCY8	ATTTAGTCACATATATTACATTACAGAGACAC	TAAGGACAGCTGTAGAGAAATTGG	SMCY8.2	TTTGAAATGTAAAGACAGAGCCTGT	AATTGGAAGAAAAAATATAAACACA
sY19*	TCACACATTTTTCAGGAGGC	ATAAATAGGCAAAATGCTACCAGT	sY19.1		TGGACTGGGACCTGCCCTA
			sY19.2	AGGGAGGAGGTTTGACGGG	TTGAAGGTACAACAGGAGAAGAAGT

			sY19.3		GGAAGAGCCACCAGGTAAATAA
sY65*	ATTATTTCTCTTTCTCTTTCCATG	AGTCTAAGAAACAGAGTGAGATCCC	sY65		AGATCCCCATCTCTAACAAACAAA
sY67*	AAGTAGCTGGGATTACAGGCA	ACATGATCCCTGTCCTCAGA	sY67	ATGCCTGGCTATTATTTCATAGC	GATGCTAGGGATGAACTCAGAAAG
sY84*	TCATTCCCCCAGTGCCCATA	TACTACCTGGAGGCTTCATCAG	sY84.1	CCCTTAGTCTGCTGTAGGCATG	
			sY84.2		CCACTGAATCTCCAGCCCAT
sY85*	AGGTGCTTTTTAGAGGCTGA	GCATCTGTATTAACAACTCTGGG	sY85		CTGGGGACACAGTAATCTATTTTAC
sY123*	CACCAACATAGGGCATAATTT	AAAGTTCCCAATAATTCATGCTAA	sY123.1	CTCCTGGGAGATGAAGGTTTT	CAGCTTTTAATAGAACATATGATGA
			sY123.2		GAAAGGACTAAGCACAAGAGAAAG
SMCYintron2*	AATATTTGTGTTAGGTTTTGATTAAGCT	TTCTTTCATGTTACATTCAATATCTC	SMCYi2.1	GGTGAGCTACTATACAAGAACGATT	
			SMCYi2.2	TACTATACAAGAACGATTGGACACA	
			SMCYi2.3		TAACATGCTTCATGCCATTTGA
			SMCYi2.4		CATTCAATATCTCATGCAATAATTC

*SNPs from Stone *et al.* (2002).

**Degenerate primers for the same SNP, differences in sequence are underlined.

Supplementary Table 4 SNPs in mitochondrial DNA identified by sequence comparisons.

Gene[1]	SNP position	SNP alleles [2]	MAF[3]	N[4]	Gene[1]	SNP position	SNP alleles [2]	MAF[4]	N[3]
MT-RNR1	721	A/G	0.50	8	MT-CO1	5395	A/G	0.11	28
MT-RNR1	737	G/A	0.38	8	MT-CO1	5539	T/C	0.13	32
MT-RNR1	833	G/A	0.25	8	MT-CO1	5614	C/T	0.44	32
MT-RNR1	956	C/T	0.13	31	MT-CO1	5767	C/T	0.44	32
MT-RNR1	975	T/C	0.13	31	MT-CO1	5848	G/A	0.13	32
MT-RNR2	1117	C/T	0.13	31	MT-CO1	6016	A/G	0.17	12
MT-RNR2	1419	C/T	0.13	31	MT-CO2	7188	A/G	0.40	5
MT-RNR2	1650	A/G	0.36	31	MT-ND5	11952	G/A	0.29	7
MT-RNR2	1663	C/T	0.13	31	MT-ND5	12135	C/T	0.29	7
MT-RNR2	2131	A/G	0.13	31	MT-ND5	12255	G/A	0.43	7
MT-ND2	3893	T/C	0.12	25	MT-CYB	14173	A/G	0.33	18
MT-ND2	3927	T/C	0.08	26	MT-CYB	14269	AC	0.21	19
MT-ND2	4057	C/T	0.46	26	MT-CYB	14299	T/C	0.40	20
MT-ND2	4147	C/T	0.19	26	MT-CYB	14392	T/C	0.16	43
MT-ND2	4193	G/A	0.23	26	MT-CYB	14467	C/T	0.33	51
MT-ND2	4301	T/C	0.19	26	MT-CYB	14515	C/T	0.33	51
MT-ND2	4342	A/G	0.19	26	MT-CYB	14551	G/A	0.08	49
MT-ND2	4619	T/C	0.19	26	MT-CYB	14728	T/C	0.42	19
MT-ND2	4681	T/C	0.19	26	MT-CYB	14857	C/T	0.39	13
MT-ND2	4721	T/C	0.46	26	MT-CYB	15004	T/C	0.23	13
MT-ND2	4858	C/T	0.19	26	MT-CYB	15088	T/C	0.46	13
MT-ND2	4892	G/A	0.12	26	MT-CYB	15115	C/T	0.46	13
					MT-CYB	15163	T/C	0.33	12

[1]Gene name according to the HUGO nomenclature committee

[2]Major allele/minor allele

[3]Major allele frequency

[4]Number of sequences

Supplementary Table 5 Parameters for the sudden expansion model, estimated from Y-chromosomal and mtDNA data, and test of good raggedness index in the wild-living chimpanzee populations

Marker	N_{ind}	$\tau = 2\mu t$ (99% CI)	$\theta_0 = 2\mu N_0$ (99% CI)	$\theta_1 = 2\mu N_1$ (99% CI)	SSD (p-value)	raggedness index (p-value)
Y-chromosomal SNPs	12	6.46 (0.0-95.5)	0.002 (0.0-7.8)	4.39 (0.5-10^5)	0.22 (p=0.045)	0.51 (p=0.02)
mtDNA SNPs	51	0.20 (0.0-0.9)	0.00 (0.0-0.4)	10^5 (16.6-10^5)	0.34 (p=0.001)	0.06 (p=1)
D-loop sequence	31	5.96 (0.5-42.5)	12.75 (0.0–40.7)	65.07 (18.0–10^5)	0.004 (p=0.85)	0.006 (p=0.79)

N_{ind} = number of individuals; τ = expansion parameter; μ = mutation rate per locus per generation; θ_0 and θ_1 = the scaled mutation parameter before and after the expansion, respectively; N_0 and N_1 = effective population size before and after the expansion, respectively; SSD = sum of squared deviations.

CAPÍTOL 3

Article

Títol: ***Molecular evolution of primate* RPS4Y *gene family***

Autors: O. Andrés, T. Kellermann, F. López-Giráldez, J. Rozas, X. Domingo-Roura, M. Bosch

Referència: En revisió a *Molecular Biology and Evolution*

Resum

El gen *RPS4* codifica per la proteïna ribosomal S4, que és una proteïna molt conservada present en tots els regnes. En mamífers, aquesta proteïna es troba codificada per un gen lligat al cromosoma X, el gen *RPS4X*. En primats, a més, el gen *RPS4* presenta una altra còpia funcional, lligada a cromosoma Y (*RPS4Y*). Recentment s'ha descobert que el gen *RPS4Y* es troba duplicat al cromosoma Y dels humans i aquesta tercera còpia (*RPS4Y2*) és també funcional, però d'expressió restringida a testicle i pròstata. L'objectiu d'aquest estudi és descriure la història evolutiva de la família *RPS4Y* en primats, que encara no està ben estudiada, especialment la del gen *RPS4Y2*. Les anàlisis de la seqüència nucleotídica de regions intròniques i de cDNAs d'aquests gens en espècies que cobreixen els quatre infraordres de primats han demostrat que l'esdeveniment de duplicació que va originar la segona còpia del cromosoma Y va tenir lloc després de la divergència de les mones del Nou Món, uns 35 milions d'anys enrere. Les anàlisis de màxima versemblança de les substitucions sinònimes i no sinònimes revelen que la selecció positiva ha actuat sobre el gen *RPS4Y2* en el llinatge humà. Aquesta és la primera vegada que s'identifica de manera inambigua l'acció de la selecció positiva en una proteïna ribosomal. Les substitucions aminoacídiques que estan sota selecció positiva afecten els tres dominis de la proteïna, però el canvi en el domini KOW és especialment interessant perquè afecta l'única posició invariable d'aquest domini, de manera que pot ser que comporti conseqüències dràstiques per a la funció. Pel que fa als mecanismes evolutius de retenció de les diferents còpies funcionals, sembla que el gen *RPS4Y1* s'ha preservat per compensar la dosi proteica de RPS4 en els ribosomes dels primats mascles, mentre que la supervivència del gen *RPS4Y2* respon a un procés d'especialització funcional o de neofuncionalització, possiblement lligada a funcions relacionades amb espermatogènesi.

Aportació personal al treball

La major part del treball de laboratori ha estat dut a terme per mi, amb l'ajut inicial de l'estudiant Thomas Kellermann al laboratori de Biologia Evolutiva de la Facultat de Ciències Experimentals i de la Salut de la Universitat Pompeu Fabra. Les anàlisis estadístiques també han estat realitzades per mi, amb el suport del Dr. Rozas. A més, he dut a terme la redacció de l'article, amb la col·laboració de tots els autors.

Molecular evolution of primate *RPS4Y* gene family

Olga Andrés[1,2], Thomas Kellermann[2,3], Francesc López-Giráldez[1], Julio Rozas[4], Xavier Domingo-Roura[1,#], Montserrat Bosch[1,2].

[1]Genètica de la Conservació Animal, Institut de Recerca i Tecnologia Agroalimentàries, Crta. de Cabrils km2, 08348 Cabrils, Spain
[2]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain
[3] Institut für Immungenetik, Charité-Universitätsmedizin Berlin, Campus Virchow-Klinikum, Humboldt-Universität zu Berlin, Spandauer Damm 130, 14050 Berlin, Germany
[4]Departament de Genètica, Universitat de Barcelona, Av. Diagonal 645, 08028 Barcelona, Spain

Genètica de la Conservació Animal, Crta. de Cabrils km 2, 08348 Cabrils, Spain
e-mail: montse.bosch@irta.es

This paper is dedicated to the memory of Dr. Xavier Domingo-Roura.

The RPS4 gene codifies for ribosomal protein S4, a very well-conserved protein present in all kingdoms. In primates, RPS4 is codified by two functional genes located on both sex chromosomes: the *RPS4X* and *RPS4Y* genes. In humans, *RPS4Y* is duplicated and the Y chromosome therefore carries a third functional paralog: *RPS4Y2*. Y-linked copies are the longest ribosomal protein genes of the human genome, and *RPS4Y2* presents a testis-specific expression pattern. In the present study, we shed new light on the evolutionary history of the *RPS4Y* gene family, especially on that of *RPS4Y2*. DNA sequence analysis of the intronic and cDNA regions of these genes from species covering the four primate infraorders showed that the duplication event leading to the second Y-linked copy occurred after the divergence of New World monkeys, about 35 million years ago. Maximum likelihood analyses of the synonymous and non-synonymous substitutions revealed that positive selection was acting on *RPS4Y2* gene in the human lineage, which represents the first evidence of positive selection on a ribosomal protein gene. Putative positive amino acid replacements affected the three domains of the protein: one of these replacements, the one located in the KOW protein domain, affected the only invariable position of this motif, and might thus have a dramatic effect on the protein function.

Key words: RPS4Y, primate, gene duplication, positive selection, Y chromosome

Introduction

RPS4 genes encode for the ribosomal protein small subunit 4 (29kD; 263 amino acids), a protein involved in mRNA binding and located at the 40S/60S subunit interface of the small ribosomal subunit (Nygard and Nika 1982). The RPS4 protein is well-conserved in prokaryotes and eukaryotes, which suggests strong functional constraints on structural evolution (Bergen et al. 1998).

RPS4 is found on autosomes in all vertebrates except mammals, which all have an X-linked copy (*RPS4X*). Fisher et al. (1990) found a Y-linked copy (*RPS4Y*) in humans, located in position p11.31, and Watanabe et al. (1993) showed that this gene is functional and functionally interchangeable with *RPS4X*. Despite the lower expression level of *RPS4Y*, both copies appeared to be necessary for correct development (Zinn et al. 1994; Lambertsson 1998). Omoe and Endo (1996) postulated that *RPS4Y* was primate specific and Bergen et al. (1998) found an increased substitution rate in great ape *RPS4Y* than in the X-linked copies, showing fewer functional constraints on the Y genes. Thus, RPS4X and RPS4Y proteins are both found in primate male ribosomes while primate female *RPS4X* genes escape inactivation. However, Jegalian and Page (1998) found a Y-linked copy in a non-primate species, *Monodelphis domestica*, which led the authors to conclude that the duplication event occurred before mammalian radiation. Moreover, another Y-linked copy has recently been discovered on the human Y chromosome and has been named *RPS4Y2* (Skaletsky et al. 2003) in order to distinguish it from

the first copy, which is now called *RPS4Y1*. *RPS4Y2* is located in position q11.223, a region associated with infertility (AZFb). While *RPS4Y1* is ubiquitously expressed, *RPS4Y2* shows a testis-specific expression pattern (Skaletsky et al. 2003; Rozen, personal communication). The existence of two paralogous copies is a unique feature of human RPS4 compared to other ribosomal proteins (Fisher et al. 1990), and the presence of three copies is even more surprising. This feature is also present in *Pan troglodytes*, based on Ensembl information (Hubbard et al. 2007).

Gene duplication is a major force for the rise of new gene functions in evolution. The average rate of gene duplication is 0.01 per gene per million years (Lynch and Conery 2003). The classical model of Ohno (1970) proposes that after gene duplication both copies can be preserved if one of the copies maintains the original function while the other acquires new functionalities after experiencing beneficial amino acid substitutions; this process is called neofunctionalization. This model also states that the most common fate of duplicate pairs is that one of the copies becomes a pseudogene by the fixation of deleterious mutations, and will finally be lost in the genome. Many other models have been proposed to explain how duplicates are retained in the genomes (Clark 1994; Force et al. 1999). In the classical subfunctionalization model, the two genes become specialized in different tissues or at different developmental stages due to selection on beneficial mutations (Ferris and Whitt 1979). Subfunctionalization can also lead to functional specialization, based on the fact that the ancestral gene had several functions and therefore the duplicates tend to become specialized for certain of these functions. Force et al. (1999) proposed a different subfunctionalization model, the duplication-degeneration-complementation (DDC) model, which explains the conservation of duplicated genes by fixation of complementary loss-of-function mutations in independent subfunctions, such that both copies are required in order to preserve the original functions (See Prince and Pickett 2002 for a review).

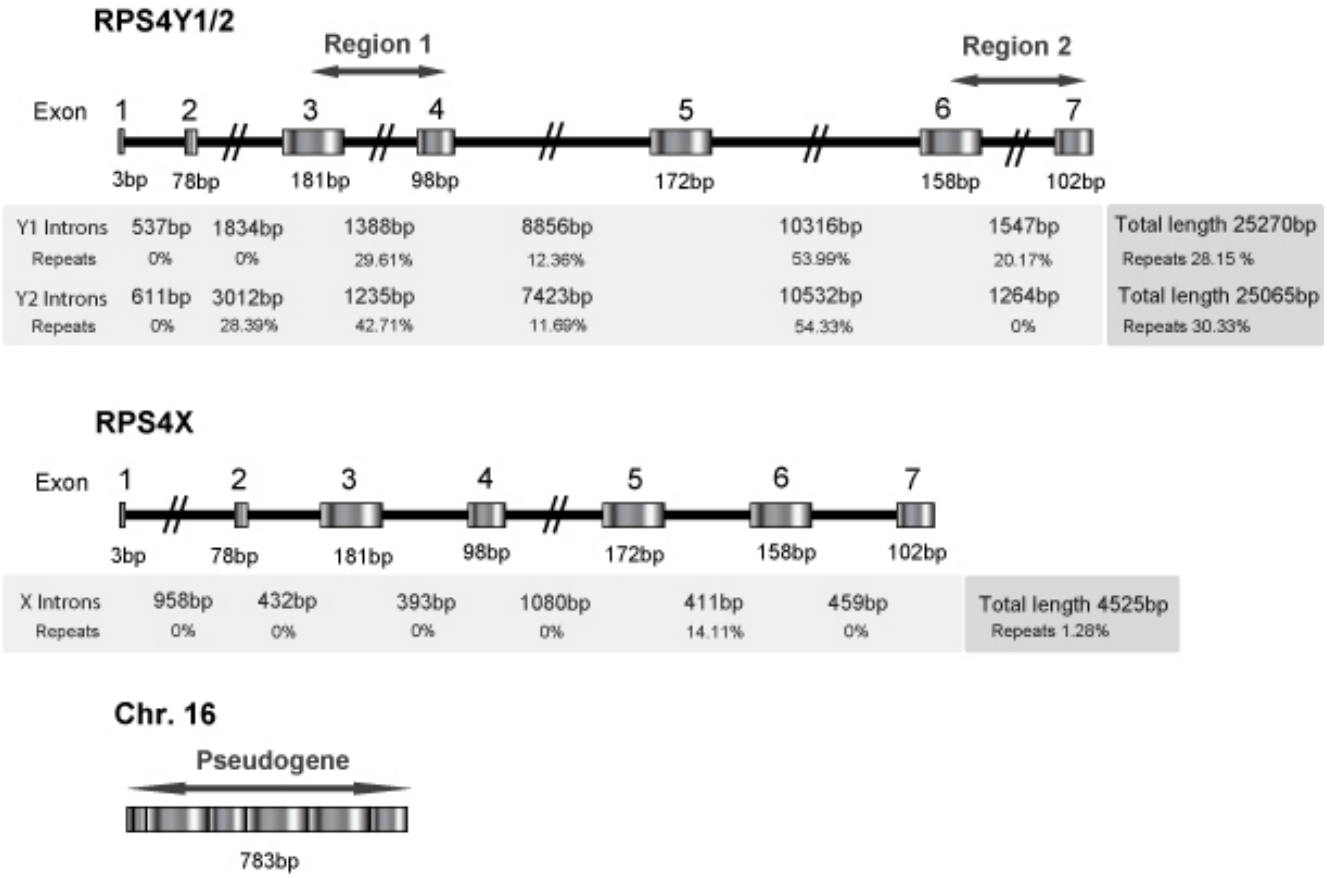

FIG. 1. Genomic structure of human *RPS4* genes. Exonic and intronic lengths and repeat content are shown. Amplified loci on the Y chromosome (regions 1 and 2) and on the pseudogene are shown with arrows.

The DDC model is a selectively neutral process that does not involve positive selection. It has, however, been found that positive selection plays an important role in duplicated gene retention in mammalian genomes (Shiu et al. 2006) and is active in neofunctionalization and some subfunctionalization events. It is necessary to distinguish between the positive selection and both the random fixation of neutral mutations and the relaxation of selective constraints. The detection of positively selected positions among interspecific sequences has usually been achieved by comparing the number of non-synonymous and synonymous substitutions per site; this approach can, however, generate false positives. Even so, there are persuasive examples of positive selection detected by this method (e.g. Yang and Bielawski 2000; Liberles et al. 2001; Liberles and Wayne 2002), particularly, after a gene duplication event. For example, Zhang, Rosenberg and Nei (1998) showed that adaptive selection acted on the divergence of the duplicated eosinophil cationic protein (*ECP*) and eosinophil-derived neurotoxin (*EDN*) genes in primates; Zhang, Zhang and Rosenberg (2002) also showed that the pancreatic ribonuclease gene 1B, which derives from the duplication of pancreatic ribonuclease gene 1, evolved under positive selection in *Pygathrix nemaeus*. Here we describe the evolutionary history of the *RPS4Y* gene family in primates. The study was conducted by analyzing DNA sequences from different species covering the four primate infraorders. Our aim was to elucidate the evolutionary mechanisms operating in the retention of these duplicated genes and the possible role of positive selection in this evolution. We also estimated the age of the main duplication events. Finally, we have discussed the functional implications of RPS4Y2 protein evolution.

Materials and Methods

Samples

Samples of *Homo sapiens* (Hsp), great apes –*Pan troglodytes* (Ptr), *Gorilla gorilla* (Ggo), and *Pongo pygmaeus* (Ppy) –, Old World monkeys (OWM) –*Macaca fuscata* (Mfu), and *Mandrillus sphinx* (Msp) –, New World monkeys (NWM) –*Saimiri sciureus* (Sbo), *Callithrix jacchus* (Cja), and *Callicebus moloch* (Cmo) –, and strepsirrhines –*Eulemur fulvus* (Efu), and *Eulemur macaco* (Ema)– were provided from the INPRIMAT sample collection. For tissue and blood samples DNA was extracted using the Qiagen tissue kit (Qiagen, Valencia, CA, USA) following the manufacturer's instructions. Initial amounts were 25mg for muscle tissues and 100µl for blood samples. DNA from cell lines was also provided by INPRIMAT DNA collection (supplementary table 1).

Amplification and sequencing

We designed male-specific primers on exonic sequences of *RPS4Y1* and *RPS4Y2* to amplify intron 3 and intron 6 (fig. 1) in different primate species. We also designed another pair of primers to amplify a complete mRNA *RPS4Y* pseudogene (fig. 1). The names and sequences of the oligonucleotides are shown in supplementary table 2. We have described the PCR conditions, fragments resulting from the use of different primer combinations, and species specificity in supplementary table 2.

PCR products were purified using the GFX PCR DNA and Gel Band Purification Kit (Amersham Biosciences UK Limited, Buckinghamshire, UK). Both strands were sequenced from the purified products using forward and reverse PCR primers (sequencing conditions are described in supplementary table 2).

Sequence analysis

Genomic sequences and information concerning human and chimpanzee *RPS4* genes were taken from Ensembl (Hubbard et al. 2007). We used the RepeatMasker v3.1.6. program (Smit, Hubley and Green, unpublished data; www.repeatmasker.org) (Smit *et al.*) to detect interspersed repeats in Ensembl genomic sequences.

We handled DNA sequences from this study and protein and cDNA sequences obtained from GenBank (see supplementary tables 3 and 4 for accession numbers) with BioEdit v6.0.7 (Hall 1999). Multiple alignments were obtained by Clustal W (Thompson et al. 1994) or DiAlign2 (Morgenstern 1999) and subsequently manually edited to minimize the number of gaps.

Once aligned, we applied GBlocks (Castresana 2000) to eliminate poorly aligned positions and divergent regions of our intronic sequences in order to obtain reliable blocks that were suitable for phylogenetic analysis. We used DnaSP v4.0 (Rozas et al. 2003) to estimate nucleotide diversity and descriptive statistics to examine the different sequence sets.

For each alignment, we selected the nucleotide substitution model that best fitted the data among 56 different evolutionary models based on the Akaike Information Criteria approach using Modeltest 3.6 (Posada and Crandall 1998). We constructed a

phylogenetic tree (fig. 2) based on the neighbor-joining (NJ) method (Saitou and Nei 1987), using PAUP*v4.0b10 (Swofford 2002). Confidence in the resulting relationships was assessed using 10,000 bootstrap replicates (Felsenstein 1985). We also performed relative rate tests on the trees by applying the RRTree program (Robinson-Rechavi and Huchon 2000), which compares substitution rates between lineages of DNA sequences, relative to a particular outgroup. TreeView (Page 1996) was used to visualize trees.

Analysis of the impact of positive or negative selection on DNA coding region was conducted using Phylogenetic Analysis by Maximum Likelihood (PAML) v3.12 (Yang 1997) and DnaSP (Rozas et al. 2003) software. We applied the different codon substitution models implemented in codeml (Branch Models, Site Models and Branch-Site Models).

We estimated the time of the duplication event (Td) from the mean number of synonymous substitutions per site ($\overline{K}_s$) among all paralogous combinations. For each paralogous copy, the synonymous substitution rate (r) was estimated for all possible pairs of species, as $r = K_s/2*Ts$, where K_s is the number of synonymous substitutions per site and Ts is the divergence time for each pair of species. For Ts, we took the minimum and maximum values from Goodman et al. (1998). Average rates ($\check{r}$) for both the minimum and maximum values were obtained from the slope of a regression analysis. We then applied the equation $Td = \overline{K}_s/2*\check{r}$ to find the estimated duplication time range.

Results

Genomic sequences from *RPS4X, RPS4Y1* and *RPS4Y2* genes were obtained from Ensembl. Surprisingly, human and chimpanzee *RPS4* genes on the Y chromosome were approximately 5 times longer than *RPS4X* (fig. 1). Despite these differences, all three RPS4 proteins had the same number of exons and these were of exactly the same length (263 amino acids); the described protein domains –S4, ribosomal-S4e and KOW– were also conserved. Variations among *RPS4* copies were explained by different repeat content in intronic regions. When Repeatmasker was applied, *RPS4X* evidenced hardly any interspersed repeats (< 1.3%), while *RPS4Y1* and *RPS4Y2* showed about 30% repeat content (fig. 1). Besides *RPS4* genes, a *RPS4Y* retrotransposed pseudogene, containing the 7 exons disrupted by stop codons, was also detected on both human chromosome 16 and its chimpanzee homologue chromosome (fig. 1).

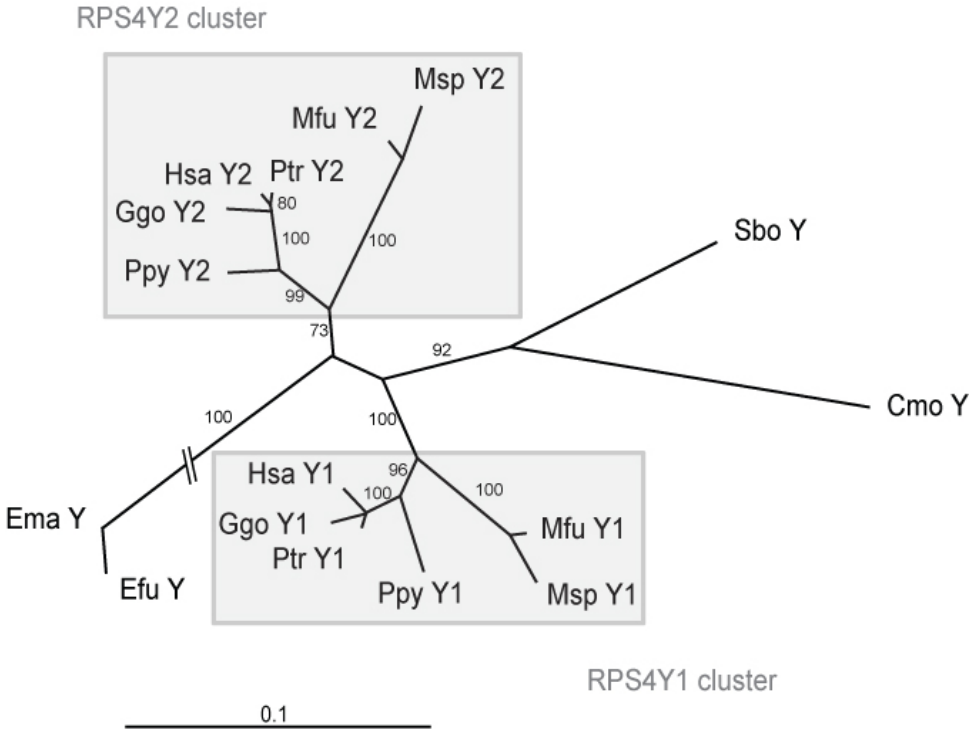

FIG. 2. Neighbor-joining tree with model TVM+G of concatenated region1 and region2 for *RPS4Y1* and *RPS4Y2* built with PAUP after GBlocks region selection (787 bp). Cja was excluded when performing the analyses since its sequence was too short. Boostrap values (10,000 replicates) are shown on each branch.

We were able to amplify the *RPS4Y1* and *RPS4Y2* intronic regions 1 and 2 in all great apes (Ptr, Ggo, Ppy) and OWM (Mfu, Msp). New World monkeys (Sbo, Cja, Cmo) and strepsirrhines (Efu, Ema) produced sequences from only a single copy. *RPS4Y* duplication should therefore have been generated before the divergence of OWM; this pattern was confirmed by FISH analyses (data not shown). The phylogenetic tree from concatenates of intronic regions 1 and 2 showed two different well-defined clusters, one for each duplicated gene copy, suggesting the independent evolution of *RPS4Y1* and *RPS4Y2* after the duplication event; the *RPS4Y* copy of NWM fell in the *RPS4Y1* cluster, while the strepsirrhine *RPS4Y* copy stayed out of the clusters (fig. 2). When we analyzed region 1 and region 2 independently, we obtained similar results. We estimated that the duplication event probably took place about 35 million years ago, which fitted in with the time between NWM and OWM divergence, estimated at between 25 and 40 million years ago (Goodman et al. 1998).

Pseudogene sequences were identified in great apes, OWM and NWM, but not in strepsirrhines. These findings suggested that the *RPS4Y* pseudogene was generated after the divergence of strepsirrhines but before the divergence of NWM. Since NWM presented these sequences, it can be inferred that the pseudogene derived from the ancestral *RPS4Y* copy (fig. 3).

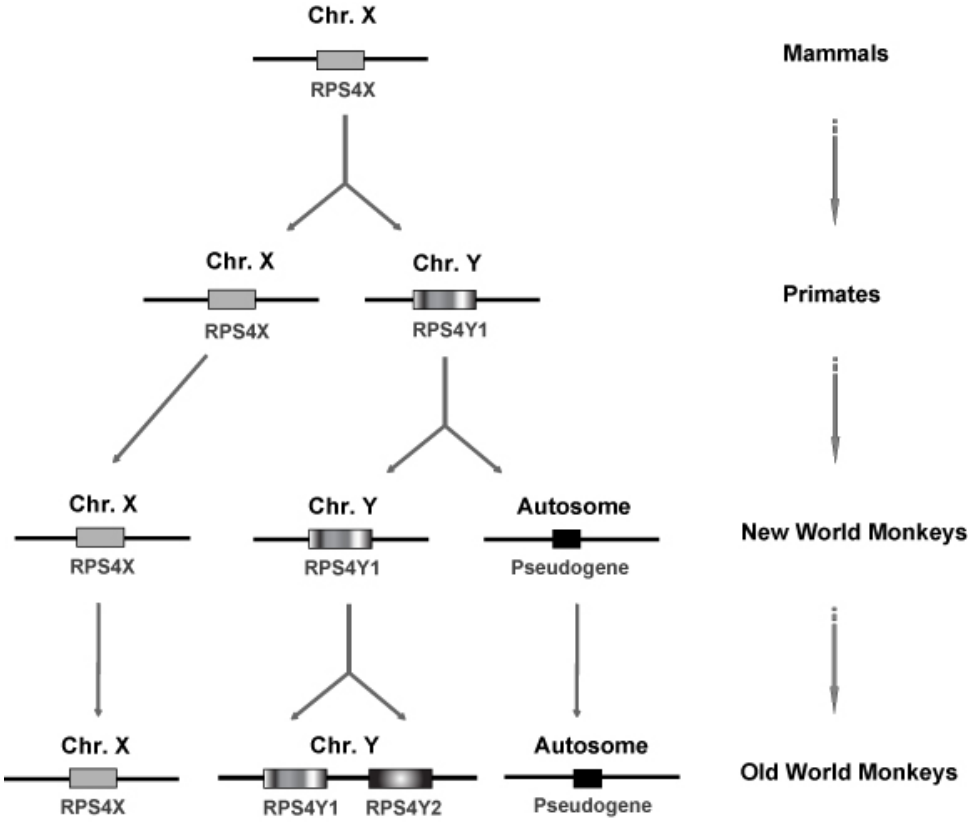

FIG. 3. Scheme of mammalian *RPS4* evolutionary history.

RPS4Y1, *RPS4Y2* and *RPS4X* GenBank cDNA sequences from great apes and OWM produced a phylogenetic tree with 3 distinct clusters (fig. 4). When we included pseudogene sequences from this study in the analysis, a phylogenetic tree with 4 well-defined clusters was generated; *RPS4X* sequences constituted a distantly separated cluster while pseudogene sequences were included in *RPS4Y1* group (data not shown).

We conducted a number of relative rate tests to seek possible deviations from the molecular clock expectations. For cDNA sequences, the results suggested that, using *RPS4X* sequences as the outgroup, only combinations involving pseudogene sequences produced statistically significant results (p-value < 0.04). With respect to intronic concatenated sequences, macaque was the only lineage to generate differences when comparing *RPS4Y2* sequences and using NWM sequence as the outgroup (p-value < 0.04).

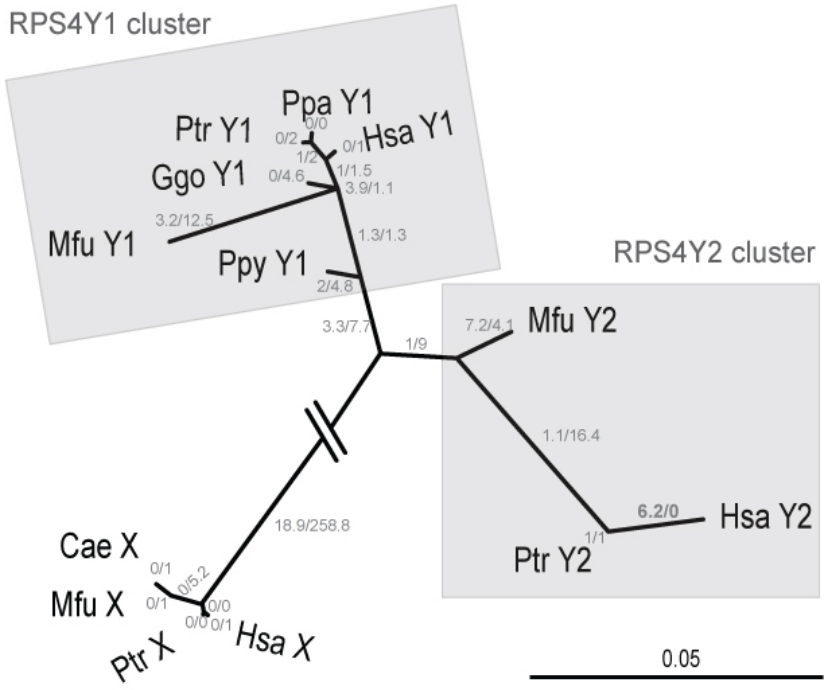

FIG. 4. Neighbor-joining tree with model TrNef+G of *RPS4X*, *RPS4Y1*, *RPS4Y2* cDNAs built with PAUP. Numbers of non-synonymous and synonymous substitutions are shown on each branch. The sequences analyzed were 789 bp long.

To determine the effect of natural selection on coding gene regions, we applied a number of maximum likelihood codon models implemented in the PAML software package (Yang 1997).

Branch models: We applied the one ratio (M0) model –which assumes the same ω ratio for the complete tree– and the free-ratio (FR) model –which allows for different ω ratios across tree-branches. We compared the two models using likelihood ratio test (LRT) to determine whether our data was compatible with homogeneous selective pressure across the branches. Whenever *RPS4Y2* sequences were considered, FR fitted the data significantly better than M0. Yet for *RPS4Y1* and *RPS4X*, the M0 model could not be rejected (table 1). When human sequences were not considered in the analyses, M0 fitted the data better than FR, even when *RPS4Y2* sequences were included (table 1), showing that the human *RPS4Y2* lineage had evolved under a different ω ratio.

Site models: Site-specific likelihood models (Nielsen and Yang 1998; Yang et al. 2000) make it possible to detect variable selective pressures across sites. Three nested models were performed: M1 (neutral) and M2 (selection); M0 (one ratio) and M3 (discrete); and M7 (beta) and M8 (beta&ω). None of the LRT tests was significant; we therefore found no evidence for sites evolving under positive selection across the different species.

Branch-site models: If positive selection only affects certain sites of specific lineages, previous models would probably not have been able to detect them. We therefore applied the more realistic branch-site models A and B (Yang and Nielsen 2002), which allow ω ratio to vary among sites and lineages. Test 1 for Model A compared this model with the site-specific model M1 (neutral); test 2 compared model A against null model A. Model B was compared with the site-specific model M3 (discrete, $K = 2$). We chose human and macaque *RPS4Y2* as possible foreground branches based on the exclusive amino acid substitutions of each protein observed in the protein alignment. We found that models A (test 1) and B only fitted the data significantly better than simpler models when human *RPS4Y2* lineage was used as the foreground branch (table 2 for model A); this indicates that positive selection acted on some sites of the human *RPS4Y2* gene. Test 2 confirmed this result when *RPS4Y1* and *RPS4Y2* sequences were considered.

When only *RPS4Y2* sequences were considered, Bayes Empirical Bayes (BEB) analysis in model A identified the amino acids located at positions 68, 70, 87, 108 and 185 as positively selected sites. When *RPS4Y1* sequences were also considered, only amino acids at positions 70 and 185 were chosen as positively selected. Finally, when all the sequences were considered (*RPS4Y1*, *RPS4Y2* and *RPS4X*) the same 5 positions plus amino acid 104 were identified as putative targets of positive selection (table 2). Positions 70 and 185 were identified in all sequence sets and had the highest posterior probabilities; moreover, these two positions were corroborated by conservative test 2. Amino acid replacements in the human *RPS4Y2* lineage involved all 3 RPS4 protein domains (S4, Ribosomal-S4e and KOW) (table 3; supplementary fig. 1). All changes in the S4 domain (amino acids 68, 70, 87) were conservative while non-synonymous amino acid substitutions in the Ribosomal-S4E (104, 108) and KOW (185) domains were radical.

Table 1
Parameter estimates for the one ratio and free-ratio branch models.

Data Set	Model	*f*	l	Estimated parameters	κ
RPS4Y1	One ratio (M0)	11	-1275.00	$\omega = 0.161$	3.40
	Free-ratio (FR)	19	-1269.87	Ggo $\omega = 1.27$	3.39
RPS4Y2	One ratio (M0)	5	-1239.64	$\omega = 0.278$	3.78
	Free-ratio (FR)*	7	-1233.58	Hsa[a]	3.77
RPS4X	One ratio (M0)	7	-1100.83	$\omega = 0.0001$	99.00[b]
	Free-ratio (FR)	11	-1100.83	Ptr $\omega = 15.81$	99.00 [b]
RPS4Y1/Y2	One ratio (M0)	17	-1571.22	$\omega = 0.192$	4.55
	Free-ratio (FR)*	31	-1555.16	Ptr Y1 $\omega = 1.71$; Hsa Y2[a]; Ggo/Mfu Y1 $\omega = 1.41$	4.55
RPS4Y1/Y2/X	One ratio (M0)	25	-1994.26	$\omega = 0.069$	2.95
	Free-ratio (FR)*	47	-1961.77	Hsa Y2[a]; Ggo/Mfu $\omega = 1.30$	3.29
Y1Y2X No Hsa *Y2*	One ratio (M0)	23	-1937.8	$\omega = 0.058$	2.95
	Free-ratio (FR)	43	-1917.62	Ggo/Mfu $\omega = 1.30$	3.24
Y1Y2 No Hsa *Y2*	One ratio (M0)	15	-1518.97	$\omega = 0.1520$	4.55
	Free-ratio (FR)	27	-1510.59	Ppa $\omega = 1.25$; Ggo/Mfu Y1 $\omega = 1.46$	4.56
Y1Y2X Hsa	One ratio (M0)	5	-1630.42	$\omega = 0.051$	2.85
	Free-ratio (FR)*	7	-1620.58		3.14

f for the number of free parameters; l for likelihood values; and κ for transitions/transversions
[a] ω cannot be estimated since there are no synonymous substitutions.
[b] κ cannot be estimated since there are no transversions.

Table 2
Parameter estimates for branch-site model A from maximum likelihood analyses.

Data Set	Foreground branch	*f*	l	Estimated parameters	κ	Positive selection (BEB)
RPS4Y1	Hsa *Y1*	14	-1275.00	p_0=1 p_1=0 (p_2+p_3=0), ω_2=1	3.40	
RPS4Y2	Hsa *Y2*	8	-1233.73	p_0=0 p_1=0 (p_2+p_3=1), ω_2=139.62	3.78	68H*, 70L*, 87I*, 108C*, 185A*
RPS4X	Hsa *X*	10	-1100.83	p_0=1 p_1=0 (p_2+p_3=0), ω_2=1	99.00 [b]	
RPS4 Y1/Y2	Hsa *Y2*	20	-1562.15	p_0=0 p_1=0 (p_2+p_3=1), ω_2 [a]	4.61	70I*, 185G*
	*Y2*cluster	20	-1570.33	p_0=0 p_1=1 (p_2+p_3=1), ω_2=139.62	4.59	
RPS4 Y1/Y2/X	Hsa *Y2*	28	-1971.13	p_0=0 p_1=0 (p_2+p_3=1), ω_2 [a]	3.21	68R*, 70I*, 87M*, 104D*, 108R*, 185G*
	*Y2*cluster	28	-1986.19	p_0=0.942 p_1=0.058 (p_2+p_3=0), ω_2=1	3.12	
Y1Y2X No Hsa *Y2*	*Y2*cluster	26	-1925.91	p_0=0.939 p_1=0.061 (p_2+p_3=0), ω_2=1	3.16	
Y1Y2 No Hsa *Y2*	*Y2*cluster	18	-1516.95	p_0=0.901 p_1=0.099 (p_2+p_3=0), ω_2=1	4.62	
Y1Y2X Hsa	Hsa *Y2*	8	-1619.84	p_0=0.647 p_1=0.021 (p_2+p_3=0.332), ω_2=1	3.03	68R, 70I, 87M, 104D, 108R, 180L, 185G, 205F, 222L (P > 0.71)

f for the number of free parameters; l for likelihood values; and κ for transitions/transversions.
* Posterior probability >0.95.
[a] ω cannot be estimated since there are no synonymous substitutions.
[b] κ cannot be estimated since there are no transversions.

We also checked for possible deviations in the correlation between synonymous and non-synonymous variation in the human *RPS4* gene family genealogy, which would have been expected according to the neutral theory. We analyzed the number of synonymous and non-synonymous substitutions for each branch (from the PAML Free Ratio model, using *RPS4X* as the ancestral sequence) (fig. 4). When we compared *RPS4Y1* and *RPS4Y2* substitution rates we did not find any significant differences (Fisher exact test, p = 0.54), while statistically significant differences were detected when comparing *RPS4X* with *RPS4Y1* and with *RPS4Y2* (p < 0.0005). The two Y-linked *RPS4* copies were therefore evolving at different rates to the X copy.

Table 4 presents nucleotide divergences between human and chimpanzee *RPS4Y1*, *RPS4Y2* and *RPS4X* cDNA, intronic sequences and pseudogene. It should be highlighted that the K_a/ K_s ratio of *RPS4X* was zero due to the absence of non-synonymous changes; this was a result of the strong functional constraints in this gene. As far as *RPS4Y1* was concerned, the increase in both the K_a/K_s ratio (ω = 0.062) and K_s (0.0270) with respect to *RPS4X* pointed to a relaxation in purifying selection. This relaxation could also be observed in the intronic

sequences, as the intronic divergence estimate (K_i) was higher for *RPS4Y1* (and similar to *RPS4Y2* and pseudogene, with K_i values of 0.0142, 0.0152 and 0.0181, respectively) than for *RPS4X* (K_i=0.0076). Finally, and as expected, *RPS4Y2* exhibited a K_a/K_s ratio that was larger than 1 (ω = 2.9477) due to the elevated K_a value (0.0129) since the K_s value (K_s=0.0044) was similar to that of *RPS4X* (K_s=0.0038). These results could not be explained by a relaxation in purifying selection (i.e. by a reduction in functional constraints) and pointed towards the action of positive selection in human *RPS4Y2* reinforcing the former PAML analyses.

Table 3
Positive positions selected by Bayes Empirical Bayes analysis in model A of the PAML.

			Change type				
BEB	Ancestral amino acid	Hsa RPS4Y2	Charge	Polarity	Polarity & Volume	Data Sets	Domain affected
68*	R	H	Cons.	Cons.	Cons.	a, c	S4
70*	I	L	Cons.	Cons.	Cons.	a, b, c	S4
87*	M	I	Cons.	Cons.	Cons.	a, c	S4
104*	D	N	Rad.	Cons.	Cons.	c	Ribosomal-S4E
108*	R	C	Rad.	Cons.	Rad.	a, c	Ribosomal-S4E
185*	G	A	Cons.	Rad.	Cons.	a, b, c	KOW

Data Sets: a) for RPS4Y2; b) for RPS4Y1 and RPS4Y2; and c) for RPS4Y1, RPS4Y2 and RPS4X.
Cons. for conservative change; *Rad*. for radical change.
* Posterior probability >0.95.

Discussion

First efforts to describe mammalian *RPS4* phylogeny suggested that *RPS4* moved to the X chromosome before mammalian radiation while *RPS4* Y-linked copy was primate specific (Bergen et al. 1998). However, the discovery of a *Rps4* Y-linked copy in the non-primate species *M. domestica* (Jegalian and Page 1998) and the location of human *RPS4Y* in an X-degenerate block (Skaletsky et al. 2003) suggest that *RPS4X* and *RPS4Y* were present in the ancestral mammalian sex chromosomes but were lost in the Y-chromosome of most lineages during the mammalian evolution (Graves 2006). In this study we have demonstrated that all primate infraorders maintained the *RPS4Y* gene and that the second Y-linked copy originated from the duplication of the *RPS4Y* gene after the divergence of NWM but before the radiation of OWM. These results shed new light on the evolution of mammalian *RPS4* gene family.

The *RPS4* gene family shows unusual characteristics for a ribosomal protein with structural functions. *RPS4Y* genes size is extraordinarily extended for a ribosomal gene (25 kb); Yoshihama et al. (2002) showed that *RPS4Y* is, in fact, the largest ribosomal protein gene in the human genome. This increase in size was caused by the insertion of repeat elements in the intronic sequences of the genes; this might be explained by their location, as the Y chromosome has an extremely high repeat content (Skaletsky et al. 2003). However, other X/Y gene duplication pairs – such as *SMCX/SMCY*, *PRKX/PRKY*, *AMELX/AMELY*, *UTX/UTY* (Hubbard et al. 2007) – do not present longer introns in the Y copy. Moreover, long introns cause a reduction in the level of expression since transcription cost increases. Purifying selection therefore also serves to maintain intron length (Castillo-Davis et al. 2002; Parsch 2003). A possible reduction in expression levels might explain why only 15% of the RPS4 proteins found in male ribosomes are from the *RPS4Y1* gene (Zinn et al. 1994).

Mammalian ribosomal proteins may present many pseudogenes but are typically encoded by a single functional gene (Zinn et al. 1994). However, RPS4 in primates is encoded by two different genes and even by three in some lineages. Davis and Petrov (2004) stated that conserved proteins (with low ω values) leave more functional duplicates in eukaryotic genomes than those that are less well-conserved. Since both *RPS4X* and *RPS4Y1* genes have an extremely well-conserved evolutionary pattern, them having left duplicates would seem consistent with this hypothesis. However, it should be noted that all ribosomal protein genes are very well-conserved and no others exhibit functional duplicates.

Table 4
Nucleotide divergence of Hsa and Ptr sequences estimated by pairwise comparisons using DnaSP and PAML software.

	DnaSP			PAML		
Data Set	K_a	K_s (K_i)	K_a/K_s	K_a	K_s	K_a/K_s
cDNA *Y1*	0.0017	0.0270	0.0619	0.0018	0.0224	0.0815
cDNA *Y2*	0.0118	0.0053	2.2147	0.0129	0.0044	2.9477
cDNA *X*	0	0.0053	0	0	0.0038	0
Introns *Y1*	-	0.0142	-			-
Introns *Y2*	-	0.0152	-			-
Introns *X*	-	0.0076	-			-
Pseudogene	-	0.0181	-			-

The preservation of the first *RPS4Y* copy in primates could be explained as a mechanism for compensating gene dosage. Meyuhas, Avni and Shama (1996) showed that the expression of ribosomal protein genes must be regulated in a coordinated way in order to ensure the equimolar assembly of the elements of the ribosomal complex. Since genes on the X chromosome are inactivated to overcome sex differences, ribosomal proteins on the X chromosome need a mechanism to achieve equimolarity. In non-primate mammals, the active *RPS4X* in females and the sole *RPS4X* in males therefore need to be more fully expressed than autosomal genes. The existence of a functional Y-linked copy in primates has led *RPS4X* to escape inactivation, as *RPS4Y* entails gene dosage compensation. In fact, in humans, the other three ribosomal X chromosome protein genes (*RPL10*, *RPL36A*, and *RPL39*) achieve equimolarity by using functional processed copies (*RPL10L*, *RPL36AL*, and *RPL39L*) elsewhere in the genome (Uechi et al. 2002).

We have shown that *RPS4Y2* emerged in the primate phylogeny between the divergence of NWM and OWM. However, nothing is known about *RPS4Y2* essentiality, or about its functionality, expression patterns in non-human primates, or about the mechanisms associated with its survival. We can undoubtedly discard a pseudogenization process since the gene has remained in the genome for approximately 35 million years, while for duplicated genes that finally disappear, half-lives tend to range from 1 to 17 million years (Lynch and Conery 2003). The testis and prostate-specific expression found in human *RPS4Y2* points to a subfunctionalization event. This hypothesis is supported by the fact that the human *RPS4Y2* promoter presents the oligopyrimidine tract as being disrupted by a mutation (data not shown). This tract is the only feature present in all ubiquitously expressed human ribosomal proteins (Yoshihama et al. 2002) and its disruption would account for the specificity of *RPS4Y2* expression in humans. Interestingly, the oligopyrimidine tract of the promoter in chimpanzee *RPS4Y2* has remained untouched, which would suggest ubiquitous expression in this species. Studies of *RPS4Y2* expression patterns in all primate lineages would elucidate whether this expression is testis-specific as in humans or ubiquitous as suggested by our observations relating to the chimpanzee *RPS4Y2* promoter.

The detection of positive selection and the relaxation of purifying selection suggest that *RPS4Y2* copy has either undergone a neofunctionalization process or been subject to a functional specialization, at least in the human lineage. The action of positive selection on ribosomal protein genes is rare. Only one study (Arbiza, Dopazo and Dopazo 2006) suggested that three ribosomal protein genes (one in the human genome and two in the chimpanzee genome) might be positively selected, but signs of positive selection were weak and it was not possible to distinguish between positive selection and a relaxation of selective constraints. We found that model A (test 1) pointed to six positively-selected positions, while test 2, which is more powerful, but very conservative, confirmed positive selection in only two of the amino acid positions. Since current statistical methods are very conservative at the

moment of detecting weak positive selection, some of the other positions identified by test 1 may also have been affected by positive selection. Moreover, there is other evidence that points to the involvement of positive selection in the *RPS4Y2* human lineage. First, branch models, which are usually very conservative, detected a different ω ratio in the human *RPS4Y2* branch. Second, nucleotide divergence information, and particularly the acceleration of non-synonymous substitutions in *RPS4Y2*, also pointed in this direction. Finally, it has been shown that positive selection is overrepresented in genes, like *RPS4Y2*, that are expressed in testis (Nielsen et al. 2005).

We detected six possible positions that were possibly affected by positive selection in the human RPS4Y2. It is not clear which specific activity of RPS4Y2 could be affected since all of the protein domains appear to be equally involved: there were three amino acids in the S4 domain, two in the ribosomal_S4E domain, and one in the KOW domain; all changes in the S4 domain were conservative, while changes in the ribosomal_S4E and KOW domains were radical. Moreover, the amino acid affected by positive selection in the KOW domain of RPS4Y2 (in position 185) is the only residue conserved in the KOW motif – a glycine in position 11 – (Kyrpides, Woese and Ouzounis 1996). In human RPS4Y2, this invariable glycine residue has been replaced by an arginine, and so the function of the domain may be dramatically affected. In order to elucidate the putative effects of the amino acid changes on the protein function, its interactions within the ribosomal complex, and the binding to RNA, it is necessary to carry out further biochemical studies, such as obtaining the three-dimensional structure.

In addition to their role in the protein synthesis, ribosomal proteins can also perform other functions (Wool 1996). Fisher et al. (1990) suggested that haploinsufficiency in RPS4 could contribute to Turner syndrome. This, in turn, led Wool (1996) to postulate that RPS4 could be involved in the regulation of development. Mutations in human *RPS4Y2* gene may therefore have improved an extra-ribosomal function that was already present in the gene. On the other hand, *RPS4Y2* location in the azoospermia region AZFb and its testis-specific expression pattern suggest a possible connection between RPS4Y2 and fertility. This feature suggests that RPS4Y2 may have acquired a new spermatogenesis-related function in human male lineage; this would be consistent with the observed excess of sperm-specific genes affected by positive selection (Nielsen et al. 2005). Complementation analyses of a rodent *Rps4* knockout mutant with human *RPS4Y2* gene would help to elucidate if this gene still conserves its original function or whether it has acquired a functional specialization related to an extra-ribosomal function or even a new function.

In conclusion, using comparative sequence analyses, we were able to establish the genealogy of *RPS4Y* genes in primate phylogeny; we corroborated the preservation of the first *RPS4Y* gene in all primate infraorders and were able to date the origin of *RPS4Y2* as occurring between the divergence of NWM and OWM. We detected that the human *RPS4Y2* gene probably evolved under positive selection and we have described the possible biological significance of this evolutionary force.

Supplementary Material

Supplementary tables 1 to 4 and figure 1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgements

We would like to thank M. Rocchi for his help with the FISH analyses. Financial support was provided by the European Commission under contract QLRI-CT-2002-01325 (INPRIMAT project). O. Andrés was supported by scholarships from the DURSI, Generalitat de Catalunya (Ref. 2003FI-00787) and T. Kellermann was supported by a scholarship from the German Academic Exchange Service (DAAD). We would also like to thank the INPRIMAT Consortium (www.inprimat.org) for supplying the samples.

References

Arbiza L, Dopazo J, Dopazo H. 2006. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput. Biol. 2:e38.

Bergen AW, Pratt M, Mehlman PT, Goldman D. 1998. Evolution of *RPS4Y*. Mol Biol Evol. 15:1412–1419.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nature. 31:415–418.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

Clark AG. 1994. Invasion and maintenance of a gene duplication. Proc Natl Acad Sci USA. 91:2950–2954.

Davis JC, Petrov DA. 2004. Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. 2:E55.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 39:783–791.

Ferris SD, Whitt GS. 1979. Evolution of the differential regulation of duplicate genes after polyploidization. J Mol Evol. 12:267–317.

Fisher EM, Beer–Romero P, Brown LG, Ridley A, McNeil JA, Lawrence JB, Willard HF, Bieber FR, Page DC. 1990. Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. Cell. 63:1205–1218.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531–1545.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP. 1998. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol. 9:585–598.

Graves JAM. 2006. Sex chromosome specialization and degeneration in mammals. Cell. 124:901–914.

Hall TA. 1999. BioEdit: a user–friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp. Ser. 41:95–98.

Hubbard TJP, Aken BL, Beal K, et al. (58 co–authors). 2007. Ensembl 2007. Nucleic Acids Res. 35(Database issue):D610–7.

Jegalian K, Page DC. 1998. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. Nature. 394:776–780.

Kyrpides NC, Woese CR, Ouzounis CA. 1996. KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. Trends Biochem Sci. 21: 425–426.

Lambertsson A. 1998. The *Minute* genes in *Drosophila* and their molecular functions. Adv Genet. 38: 69–134.

Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA. 2001. The adaptive evolution database (TAED). Genome Biol. 2:Research0028.1–Research0028.6.

Liberles DA, Wayne ML. 2002. Tracking adaptive evolutionary events in genomic sequences. Genome Biol. 3:Reviews1018.1–1018.4.

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. J. Struct. Funct. Genomics. 3:35–44.

Meyuhas O, Avni D, Shama S. 1996. Translational control of ribosomal protein mRNAs in eukaryotes. In: Hershey JWB, Mathews MB, Sonenberg N, editors. Translational Control. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press. p. 363–388.

Morgenstern B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics. 15:211–218.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV–1 envelope gene. Genetics. 148:929–936.

Nielsen R, Bustamante C, Clark AG, et al. (13 co–authors). 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Nygard, Nika. 1982. Identification by RNA-protein cross-linking of ribosomal proteins located at the interface between the small and the large subunits of mammalian ribosomes. EMBO J. 1:357–362.

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer.

Omoe K, Endo A. 1996. Relationship between the monosomy X phenotype and Y linked ribosomal protein S4 (*Rps4*) in several species of mammals: a molecular evolutionary analysis of *Rps4* homologs. Genomics. 31:44–50.

Page RDM. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput Appl Biosci. 12:357–358.

Parsch J. 2003. Selective constraints on intron evolution in Drosophila. Genetics. 165:1843–1851.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Prince VE, Pickett FB. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat Rev Genet. 3:827–37.

Robinson–Rechavi M, Huchon D. 2000. RRTree: Relative–rate tests between groups of sequences on a phylogenetic tree. Bioinformatics. 16:296–297.

Rozas J, Sánchez-Delbarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by

the coalescent and other methods. Bioinformatics 19:2496–2497.

Saitou N, Nei M. 1987. The neighbor–joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Shiu SH, Byrnes JK, Pan R, Zhang P, Li WH. 2006. Role of positive selection in the retention of duplicate genes in mammalian genomes. Proc Natl Acad Sci USA. 103:2232–2236.

Skaletsky H, Kuroda–Kawaguchi T, Minx PJ, et al. (40 co–authors). 2003. The male–specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825–837.

Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4.0b10. Sunderland (MA): Sinauer Associates.

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. 22:4673–4680.

Uechi T, Maeda N, Tanaka T, Kenmochi N. 2002. Functional second genes generated by retrotransposition of the X-linked ribosomal protein genes. Nucleic Acids Res. 30:5369–5375

Watanabe M, Zinn AR, Page DC, Nishimoto T. 1993. Functional equivalence of human X and Y–encoded isoforms of ribosomal protein S4 consistent with a role in Turner syndrome. Nat Genet. 4:268–271.

Wool IG. 1996. Extraribosomal functions of ribosomal proteins. TIBS 21:164–165.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 155:431–449.

Yang Z, Nielsen R. 2002. Codon–substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.

Yoshihama M, Uechi T, Asakawa S, et al. (11 co–authors). 2002. The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. Genome Res. 12:379-390

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA. 95:3708–3713.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf–eating monkey. Nat Genet. 30:411–415.

Zinn AR, Alagappan RK, Brown LG, Wool I, Page DC. 1994. Structure and function of ribosomal protein S4 genes on the human and mouse sex chromosomes. Mol Cell Biol. 14:2485–2492.

Supplementary Material

Supplementary tables

Supplementary table 1: Sample information: INPRIMAT code, species name, sex and DNA source are given.

Code	INPRIMAT code	Species	Sex	DNA source
Hsa_M		*Homo sapiens*	Male	Blood
Ptr_M	PTR00321IN	*Pan troglodytes* (Ptr)	Male	Cell line
Ptr_F		*Pan troglodytes* (Ptr)	Female	Cell line
Ggo_M	GGO00605IN	*Gorilla gorilla* (Ggo)	Male	Muscle
Ggo_F		*Gorilla gorilla* (Ggo)	Female	Cell line
Ppy_M		*Pongo pygmaeus* (Ppy)	Male	Muscle
Ppy_F	PPY00329IN	*Pongo pygmaeus* (Ppy)	Female	Cell line
Mfu_M		*Macaca fuscata* (Mfu)	Male	Blood
Mfu_F		*Macaca fuscata* (Mfu)	Female	Blood
Msp_M		*Mandrillus sphinx* (Msp)	Male	Blood
Msp_F		*Mandrillus sphinx* (Msp)	Female	Blood
Sbo_M	SBO00731IN	*Saimiri boliviensis* (Sbo)	Male	Blood
Sbo_F	SBO00688IN	*Saimiri boliviensis* (Sbo)	Female	Blood
Cja_M	CJA00376IN	*Callithrix jacchus* (Cja)	Male	Cell line
Cmo_M		*Callicebus moloch* (Cmo)	Male	Cell line
Efu_M	EFU00661IN	*Eulemur fulvus* (Efu)	Male	Muscle
Ema_M		*Eulemur macaco* (Ema)	Male	Cell line

Supplementary table 2: Resulting fragments of *RPS4Y1* and *RPS4Y2* amplification in different primate species using different primer combinations. E.g. *Reg1C1a* = *Reg1* for region 1 (intron3), *C1* for copy 1, *a* for the first amplified fragment. Nomenclature is analogue for region 2 (intron6). *CY* means there is no Y-linked copy specificity. Forward and reverse primer sequences are shown. Nomenclature: e.g. C1E3F1 = C1 (specific for *RPS4Y* copy 1), E3 (located in exon3), F (forward), and 1 (first primer designed in this location). CY refers to a primer that amplifies both Y-linked copies. mRNAYF and mRNAYR were used to amplify the pseudogene.
$MgCl_2$ [mM] column shows the concentrations used in the experiments and T_{Exp} is the experimental annealing temperature. (TD) means that the PCR program included touchdown cycles. Three-letter code indicates the species amplified with each primer pair.

Fragment	F Primer	Forward primer sequence 5' to 3'	R Primer	Reverse pimer sequence 5' to 3'	MgCl2 [mM]	T Exp (ºC)	Species amplified
Reg1C1a	C1E3F1	CTCAGGAATAGACTCAAGTATGCGT	C1E4R1	TCATAGACCAGGCGGAAATGT	1,5	60 (TD)	Ptr/Ppy/Sbo/Efu
Reg1C1b	CYE3F1	TCAGGAATAGACTCAAGTATGCGT	C1E4R2	CTTCCACTGTGATGCGGTGA	2	59	Mfu/Msp
Reg1C2a	C2E3F1	CCTCAGGAATAGACTCAAGTATGCA	C2E4R1	GCGGAAATGCTCACCTGTT	1,5	60 (TD)	Ptr/Ggo/Ppy
Reg1C2b	CYE3F1	TCAGGAATAGACTCAAGTATGCGT	C2E4R2	TCTTCCGCTGTGATACGATGA	2	59	Mfu/Msp
Reg1CY	CYE3F1	TCAGGAATAGACTCAAGTATGCGT	CYE4R1	CAAAACGGCCCTTGGTGTC	2	59	Ggo(c1) /Cja/Cmo/Ema
Reg2C1a	C1E6F1	GGGAAAGACATCCTGGTTCTTTT	C1E7R1	AGCCACTGCTCTGTTTGGTG	1,5	60 (TD)	Ptr/Ppy/Mfu
Reg2C1b	C1E6F2	TGTATGGTGATTGGTGGAGCC	C1E7R2	TTAGCCACTGCTCTGTTTGGTG	2	62	Ggo/Mfu
Reg2C2a	C2E6F1	GGAAAGACATCCTGGTTCTTGC	C2E7R1	GCCACTGCTCTGTTTGGCA	1,5	60 (TD)	Ptr/Ppy/Mfu
Reg2C2b	C2E6F2	TGGAGCTAACCTCGGTCGTG	C2E7R2	TCTCTCTTCAGCAATAGTAAGTCGG	2	60,5	Ggo/Ppy
Reg2CY	CYE6F1	CTTTTGATGTGGTGCATGTGAAG	CYE7R1	GGCAGGGAAATCCAAGGTTTA	1,5	60 (TD)	Sbo/Cja/Efu
mRNAY	mRNAYF	GGGCCCTAAGAAGCACTTG	mRNAYR	TTAGCCACTGCTGTTTGGTG	1,5	57	Hsa/Ptr/Ggo/Ppy/Mfu/Sbo/Cja

We followed standard PCR procedures in a final volume of 25µl containing a mixture of 0.17 uM of primers, 0.32 mM dNTPs, 2 mM $MgCl_2$, 0.034 U/µl Taq (Ecogene, Barcelona, Spain) and 0.6 ng/µl of DNA. PCR conditions included an initial cycle at 94°C for 5 minutes (min), 30 cycles divided into three steps of 45 seconds (sec) at 94°C, 45 sec at 57-60.5°C and 1 min at 72°C, and a final extension at 72°C for 5 min.

We used a touchdown PCR approach when normal conditions resulted in an unspecific smear or when the product of interest could only be very weakly amplified. Using this approach, DNA was amplified following the same conditions as for the standard PCR protocol, but with 1.5mM of $MgCl_2$. Amplification conditions consisted of an initial cycle at 95°C for 5 min followed by 20 cycles divided into three steps of 30 sec at 95°C, 30 sec from 70°C to 60ºC –the primer-specific annealing temperature was reduced by 0.5°C in each cycle– and 30 sec at 72°C; and 20 more cycles divided into three steps of 30 sec at 95°C, 30 sec at 60°C and 30 sec at 72°C to increase product yield, followed by a final extension at 72°C for 5 min.

Standard sequencing reactions contained 10 ng of purified DNA, 1.6 ng of primer and 2 µl of ABI Prism™ Big Dye Terminator Cycle Sequencing 3.1 kit in a final volume of 10 µl. Cycling conditions included an initial cycle at 94°C for 4 min, 25 cycles of 10 sec at 96°C, 5 sec at 50°C and 4 min at 60°C. Sequencing was performed on an ABI 3100 automated DNA sequencer (Applied Biosystems, Foster City, CA, USA).

Supplementary table 3: Accession numbers of intronic and pseudogenic sequences generated in this study. Code: Y1 for *RPS4Y1*, Y2 for *RPS4Y2*, Y for single *RPS4Y* gene, and ψ for pseudogene.

Name	Species	Intron3 accession number	Intron6 accession number	Pseudogene accession number
Ptr Y1	*Pan troglodytes*	EF408708	EF408722	-
Ggo Y1	*Gorilla gorilla*	EF408709	EF408723	-
Ppy Y1	*Pongo pygmaeus*	EF408710	EF408724	-
Mfu Y1	*Macaca fuscata*	EF408711	EF408725	-
Msp Y1	*Mandrillus sphinx*	EF408712	-	-
Ggo Y2	*Gorilla gorilla*	EF408713	EF408726	-
Ppy Y2	*Pongo pygmaeus*	EF408714	EF408727	-
Mfu Y2	*Macaca fuscata*	EF408715	EF408728	-
Msp Y1	*Mandrillus sphinx*	EF408716	-	-
Sbo Y	*Saimiri boliviensis*	EF408717	EF408729	-
Cja Y	*Callithrix jacchus*	EF408718	EF408730	-
Cmo Y	*Callicebus moloch*	EF408719	-	-
Efu Y	*Eulemur fulvus*	EF408720	EF408731	-
Ema Y	*Eulemur macaco*	EF408721	-	-
Ptr ψ	*Pan troglodytes*	-	-	EF408702
Ggo ψ	*Gorilla gorilla*	-	-	EF408703
Ppy ψ	*Pongo pygmaeus*	-	-	EF408704
Mfu ψ	*Macaca fuscata*	-	-	EF408705
Sbo ψ	*Saimiri boliviensis*	-	-	EF408706
Cja ψ	*Callithrix jacchus*	-	-	EF408707

Supplementary table 4: Accession numbers of nucleotide and protein sequences used from GenBank for cDNA and protein analyses. Y1 is used for *RPS4Y1*, Y2 is used for *RPS4Y2*, and X is used for *RPS4X* genes.

Name	Species	cDNA Accession Number	Protein Accession Number
Hsa Y1	*Homo sapiens*	NM_001008.3	NP_000999.1
Ptr Y1	*Pan troglodytes*	AY633110.1	Q861U9
Ppa Y1	*Pan paniscus*	AH012490.1	Q861V0
Ggo Y1	*Gorilla gorilla*	AH012492.1	Q861U8
Ppy Y1	*Pongo pygmaeus*	AH012493.1	Q861U7
Mfu Y1	*Macaca fuscata*	D50105.1	BAA21076.1
Hsa Y2	*Homo sapiens*	NM_001039567.2	Q8TD47
Ptr Y2	*Pan troglodytes*	AY633111.1	Q6GVM7
Mfu Y2	*Macaca fuscata*	AB024286.1	BAA87933.1
Hsa X	*Homo sapiens*	NM_001007.3	AAH71662.1
Ptr X	*Pan troglodytes*	XM521131.1	-
Mfu X	*Macaca fuscata*	AB024285.1	Q76MY1
Cae X	*Chlorocebus aethiops*	AB015610.1	Q76N24

Supplementary figure

Supplementary fig. 1: Primate RPS4 protein alignment. Four different Pfam domains are marked in different colors. Nucleotide positions under positive selection in the human RPS4Y2 lineage are highlighted in red squares.