IDENTIFICATION OF HUMAN SPECIFIC GENE DUPLICATIONS RELATIVE TO OTHER PRIMATES BY ARRAY CGH AND QUANTITATIVE PCR

Gemma Armengol[a,b]*, Sakari Knuutila[b], Juan-José Lozano[c], Irene Madrigal[d], María-Rosa Caballín[a].

[a] Biological Anthropology Unit, Department of Animal Biology, Plant Biology and Ecology, Faculty of Biosciences, Universitat Autònoma de Barcelona (UAB), Barcelona, Spain

[b] Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland

[c] CIBER de Enfermedades Hepáticas y Digestivas (CIBEREHD), Hospital Clínic, Barcelona, Spain.

[d] CIBER de Enfermedades Raras (CIBERER) and Biochemistry and Molecular Genetics Department, Hospital Clínic, Barcelona, Spain

*Corresponding author: Gemma Armengol, Biological Anthropology Unit, Department of Animal Biology, Plant Biology and Ecology, Facultat de Biociències, Universitat Autònoma de Barcelona (UAB), 08193-Bellaterra, Barcelona, Spain. E-mail: gemma.armengol@uab.cat. Tel: +34 93 5812049, Fax: +34 93 5811317.

ABSTRACT

In order to identify human lineage specific (HLS) copy number differences (CNDs) compared to other primates, we performed pair wise comparisons (human *vs.* chimpanzee, gorilla and orangutan) by using cDNA array comparative genomic hybridization (CGH). A set of 23 genes with HLS duplications were identified, as well as other lineage differences in gene copy number specific of chimpanzee, gorilla and orangutan. Each species has gained more copies of specific genes rather than losing gene copies. Eleven of the 23 genes have only been observed to have undergone HLS duplication in Fortna et al. (2004) and in the present study. Then, seven of these 11 genes were analyzed by quantitative PCR in chimpanzee, gorilla and orangutan, as well as in other six primate species (*Hylobates lar*, *Cercopithecus aethiops*, *Papio hamadryas*, *Macaca mulatta*, *Lagothrix lagothricha*, and *Saimiri sciureus*). Six genes confirmed array CGH data, and four of them appeared to have *bona fide* HLS duplications (*ABCB10*, *E2F6*, *CDH12*, and *TDG* genes). We propose that these gene duplications have a potential to contribute to specific human phenotypes.

Key words: human, primates, genetics, copy number differences, array CGH

1. INTRODUCTION

It has long been postulated that the differences between human and great ape genomes may be based on differences in gene sequences, gene expression and/or gene complement of a few coding genes, given the rather stable mammalian genome. Just a single Robertsonian fusion and a few chromosomal rearrangements differentiate the karyotypes of great apes [1]. Therefore, gene copy number differences (CNDs) are considered to be one of the significant sources of DNA variation between primates as the analysis of the human, chimpanzee and macaque genomes has revealed [2-5]. It is well known that the gene complement of an organism can be altered by sequence gain and loss events, resulting in phenotypic variation susceptible to selection pressures [6]. According to Kehrer-Sawatzki and Cooper [7], the myriad submicroscopic rearrangements in primate genomes, particularly those involving copy number variation, are unlikely to represent exclusively neutral changes and hence promise to facilitate the identification of genes that have been important for human evolution.

One of the best methods to detect DNA copy number changes is comparative genomic hybridization (CGH), which is able to compare directly the entire genomes of two different species [8]. The array CGH approach with human high-density arrays has increased the resolution of this technique and has allowed identifying sequence differences between humans and other great apes. Several studies have reported DNA copy number differences between human and nonhuman primates [5, 6, 9-14]. However, only two of them used a gene-based approach by employing cDNA array-based CGH [10, 13]. The use of cDNA clones has the advantage of permitting the analysis of individual genes over the use of BAC clones. As abovementioned, knowledge of changes in individual gene copy number can provide direct data regarding the primary molecular mechanisms underlying genome evolution. Fortna et al. [10] compared five hominoid species (human, bonobo, chimpanzee, gorilla, and orangutan) and identified 1,005 genes that gave genetic signatures unique to one or more of the hominoid lineages, including 134 and 6 human lineage-specific increases and decreases, respectively. Some of these genes detected to be amplified in human are thought to have possible neuronal functions. Recently, the same research group extended the study across 10 primate species, including human, and found 84 genes with human-specific copy number changes [13].

In an effort to understand the genetic history and evolution of our species, we wanted to investigate further the gene copy number gains and losses between humans and chimpanzee,

gorilla and orangutan in pair wise comparisons (human-chimpanzee, human-gorilla and human-orangutan) by using cDNA array-based CGH. A set of *bona fide* 23 genes were identified with human lineage-specific duplication, among other lineage differences in gene copy number specific of other hominoids. Seven of these genes were chosen to be validated by quantitative PCR (qPCR) because their duplication has never been studied in detail. Four of them were confirmed to be duplicated in humans *vs.* chimpanzee, gorilla, orangutan, and other six simians.

2. RESULTS AND DISCUSSION

2.1 Detection of lineage-specific changes

In this study, over 16,000 human genes across four hominoid species (human, common chimpanzee, gorilla, and orangutan) were compared using array CGH, leading to the identification of various CNDs (array CGH plots are shown in Supplementary Figure 1). We selected as positive for CNDs those sites in which both replicas were above or under the $\log_2$ ratios ± 2 Standard Deviations of the thresholds of middle 50% quantile of data (±0.38, ±0.46, and ±0.56 in chimpanzee, gorilla and orangutan, respectively). Sixty two genes or transcribed sequences showed array CGH signatures unique to humans (Supplementary Table 1). Then, our data were combined with Fortna et al. [10] results, given the high similarity between the arrays. With such approach we were able to narrow down the *bona fide* human lineage specific (HLS) CNDs. The fact of analyzing together our results with those of a previous report and the use of very conservative criteria provide more confidence with regard to predictions of true LS changes, even though they are likely to be underestimates of the actual list of genes with CNDs between hominoids. For this second analysis, only sites displaying CNDs in our and Fortna et al. [10] raw data were scored as positive and 23 HLS CNDs were identified. Interestingly, all of them corresponded exclusively to gains in the human genome (Table 1). In addition, genes with CNDs specific to one or more non-human great ape lineages *vs.* human genome were also detected (Figure 1). Chimpanzee presented 22 increases and 4 decreases, gorilla 28 gene copy number gains and 4 losses, and orangutan 27 gains and 23 losses. Common chimpanzee LS (CLS), gorilla LS (GLS) and orangutan LS (OLS) CNDs are shown in Supplementary Tables 2, 3, and 4, respectively. A total of 156 copy number changes were detected at 150 loci. Some genes are duplicated in one species and lost in another. In these cases, the differences were considered to be independent events.

Although the arrays were spotted with human genomic sequences and hybridization becomes less reliable when more distant species are evaluated, the species compared in the array CGH study have a high degree of sequence conservation. Moreover, cDNA microarrays should be less sensitive than other types of arrays to small sequence differences across species, and therefore, in principle, should be more suitable for comparative studies [15]. Finally, the use of nonfixed cutoffs but quantile-based cutoffs offers very conservative thresholds and adapted to sequence divergence, allowing the detection of *bona fide* CND even though with the risk of losing other small differences. Gilad et al. [16] reported the approach of conservative thresholds to be the only context in which cross-species comparisons can be reliable. Moreover, six out of the 23 HLS CNDs identified in the present study were confirmed by qPCR, and 12 had been previously reported as duplicated in humans *vs.* other great apes (see below).

As expected, the estimated time of divergence of each species from the common ancestral lineage (about 6 million years for chimpanzee, 8.2 million years for gorilla and 14 million years for orangutan) had a very good correlation with the number of CNDs observed. Human and chimpanzee had the fewest number of CNDs (23 in human and 26 in chimpanzee); gorilla showed 32 CNDs; and orangutan had 50. This gives about 4 CNDs per million years of age (average of the four species is 3.85 and Standard Deviation is 0.3). These findings agree with previous array CGH studies [6, 10], where the quantity of variant sites detected in each great ape species was in proportion to the estimated divergence time of each species.

Moreover, 7 gene copy number decreases and 1 increase were detected both in gorilla and orangutan *vs.* human genome. The most parsimonious explanation is that these CNDs would correspond to 7 increases and 1 decrease probably originated in the common ancestor of humans and chimpanzees, and therefore would identify changes specific for *Homini* (HominiLS), which are shown in Supplementary Table 5. This agrees with the 2.2 million years of separation of *Gorillini* from *Hominae* (human, chimpanzee and gorilla ancestor). However, following the same explanation, the 27 gains and 23 losses detected in orangutan could be also 27 losses and/or 23 gains in *Hominae*, in this case with the same theoretical probability. Thus, it can not be completely discarded that some of the OLS CNDs are alterations originated during the evolution from *Hominidae* (human, chimpanzee, gorilla and orangutan ancestor) to *Hominae* (HominaeLS). If this is the case, then the number of CNDs in

orangutan genome would be lower than what one may expect by its estimated time of divergence from the common ancestral lineage. Goidst et al. [11] observed something similar and hypothesized that the orangutan genome might have a more conserved nature at submicroscopic level, as has been observed at karyotypic level.

All lineages showed more gene copy number gains than losses. Almost 76% of all CNDs corresponded to increases in copy number. If we exclude the OLS CNDs from this analysis (for the above-mentioned reasons), then the percentage increases to 86%, and if then we consider just the species-specific CNDs, then the percentage goes up to 90%. It seems that some of the phenotypic differences among apes might be due to an increase of gene copy number in such a way that each species has gained more copies of specific genes compared to the ancestral common lineage rather than losing gene copies. This agrees with previous results from Fortna et al. [10] and Goidts et al. [11]. However, it is worth noting that the design of these array CGH experiments impairs the detection of complete gene losses that may have occurred during human speciation (only genes from the human genome were arrayed).

Other studies have performed genome-wide surveys of interhominoid copy number variation using either computational analyses [9], BAC-based array CGH [6, 11, 12] or oligonucleotide array CGH [14] with different coverage of the genome. Comparisons with our work are difficult because all these studies have analyzed the whole genome, contrary to the present study and Fortna et al. [10] and Dumas et al. [13] studies, which have focused only on the coding portion of the genome. Unfortunately, Dumas et al. [13] data could not be combined with ours because there was a high disparity in the arrayed sequences. Apart from array CGH reports, other studies have described few genes with duplications in the human lineage [17-21]. From all the genes with HLS CNDs identified in all these reports [6, 9, 11-14, 17-21] (using array CGH or other methods), 12 agreed with our results, taking into account that only the genes present in our and in Fortna et al. [10] arrays were considered for the comparison (see Table 1). Apart from this, there were three genes in which discrepancies were observed due to the experimental procedure, such as *SPANX* [17] and *MGC8902* [13, 21], which were flagged as outliers by our microarray analysis software, and *HECT domain and RLD2* [12], with different results in the two replicas. Finally, two genes, the *ANAPC1* [11] and *KIAA0514* [14] did not show CNDs in the present study. Interestingly, the gene coding for amylase was HLS according to Wilson et al. [12] after studying the human genome relative to chimpanzee

and gorilla, but we observed that orangutans have the same number of copies as humans, and therefore we would not consider it as HLS.

2.2. Gene copy number HLS variations

Special focus was paid to the 23 genes with HLS CNDs obtained after combining our data with Fortna et al. [10] results. Their characteristics are shown in Table 1. After a detailed analysis of function and tissue expression of these genes, we observed that 16 of them have neuronal function and/or are expressed in neurons or in tissues related to human-specific characteristics (parathyroid and cochlea), although in some genes these functions/tissue expressions are not exclusive. Moreover, 11 of these 23 genes (*ABCB10*, *E2F6, CDH12, PMP2, TDG, USP10, PAIP1*, and four *Homo sapiens* transcribed sequences) have only been observed to have undergone HLS duplication in Fortna et al. [10] and in the present study.
A detailed analysis of these 11 new HLS CNDs was performed, especially regarding gene functions. Five of these genes are involved in neuronal function and/or expression. First, E2F6 has a crucial role in the cell cycle. It appears to regulate a subset of E2F-dependent genes, whose products are required for entry into the cell cycle, but not for normal cell cycle progression. Interestingly, this gene is expressed in developing brain [22]. *CDH12* encodes a cadherin expressed specifically in the brain (its alias are brain cadherin and neuronal cadherin). These cadherins provide a mechanism for selective neuronal guidance and recognition. Another gene with neuronal function is *PMP2*, coding for one of the major proteins of peripheral myelin. Myelin coated axons make possible fast flowing information, which is required in large brains, such as those of humans. Moreover, TDG is an enzyme that repairs DNA mismatches caused by the spontaneous deamination. These genetic mutations are the only types that can occur naturally in non-dividing cells such as neurons and it has been observed that TDG is responsible for DNA repair in the adult rodent brain [23]. Finally, USP10 is a protease that is involved in synaptic growth. The loci corresponding to USP10 is located at SLI1 (Specific language impairment) region [24].

Moreover, two of the transcribed sequences with unknown function (AI088089 and H15704) are expressed in tissues related to the human-specific phenotype. AI088089 cDNA is highly expressed in parathyroid gland. The purpose of parathyroid glands is to regulate the calcium level within a very narrow range so that the neurons and muscular systems can function properly. Moreover, H15704 cDNA, which is similar to keratinocyte growth factor gene

(*FGF7*), is mostly expressed in cochlea. Interestingly, genes involved in the development of hearing have been reported to undergo positive selection in human evolution [25]. These authors speculated that fine-tuning of the human cochlea may have been required for understanding spoken language. The other transcribed sequences with unknown function were H40480 and H57306, which are located in the SMA (spinal muscular atrophy) region of chromosome 5 (see below).

The other two genes with HLS CNDs have important cell functions not directly related to the nervous system. ABCB10 is a mitochondrial inner membrane erythroid transporter involved in heme biosynthesis and multidrug resistance, and PAIP1 acts as a co activator in the regulation of translation initiation of poly(A) containing mRNAs.

If we focus on cytoband locations, two gene clusters can be identified for HLS genes: one at 1q21 and another at 5q13.2 (Table 1). It is well known that a pericentric inversion at chromosome 1 occurred after the separation of the Old World monkeys from the lineage leading to the great apes. Then, during human evolution, two duplications-transpositions (from 1q32 to 1q21 and to 1p11.2) followed by a second pericentric inversion (between 1p11.2 and 1q21) occurred specifically in the human lineage [26]. This lead to 91 Kb duplications that constitute HLS SD that originated by duplicative transposition [27]. Interestingly, the cDNAs corresponding to a transcribed sequence with strong similarity to DRD5, an mRNA similar to FAM72A, *PDE4DIP*, and *FCGR1A* are located into one of these 91 Kb-spanning duplicated regions. Another interesting cluster was at 5q13.2 (the predisposition locus of the neurodegenerative disorder SMA), where four HLS genes (two transcribed sequences, *NAIP* and *OCLN*) are located. Moreover, several copies of *CDH12*-derived sequences are present on different loci in the human chromosome 5, including 5q13.2. Interestingly, a 500-Kb inverted duplication has been previously described at the SMA locus involving these and other genes, such as *Serf1A* and *SMN1* [28]. Our results showed copy number increase for both genes, while Fortna et al. [10] did not observe a gain of *Serf1A* and their array did not include the *SMN1* gene. These two clusters at chromosome 1q21 and 5q13 with HLS duplications have been identified as well by others [9-12].

Overall, many of the genes reported here as HLS belong to gene families, which are prone to expansion events [29]. For example, Fc fragment of IgG receptors, ATP-binding cassette proteins, olfactory receptors, cadherins, or aquaporins.

Two of the HLS CNDs found in the present study (*NEK2*, and *PMP2*) were located in regions not previously described as SD (Table 1). Many others have confirmed the association between structural divergence and SD [6, 9-11]. Since most of the SD contain genes, it is expected that they constitute a major source of LS gene copy number increases.

2.3. Confirmation of some HLS CNDs by comparing with nine simian species

We chose seven out of the 11 selected genes with HLS duplication (*ABCB10*, *E2F6, CDH12, PMP2, TDG, USP10,* and *PAIP*) given that the other four were *H. sapiens* transcribed sequences with unknown function. Copy numbers of these seven genes were tested by qPCR in chimpanzee, gorilla and orangutan, as well as in other simians (*Hylobates lar*, *Cercopithecus aethiops*, *Papio hamadryas*, *Macaca mulatta*, *Lagothrix lagothricha*, and *Saimiri sciureus*) in order to confirm if they were actual HLS. All genes except *PMP2* showed qPCR data consistent with array CGH values (correlation coefficient values >0.75). In *ABCB10*, *E2F6*, *CDH12*, and *TDG* genes the specificity of human duplications was confirmed because all the other tested simian species presented lower number of copies (Figure 2A-D). Moreover, there was a positive correlation between qPCR data of these genes and the evolutionary distance measured in million years for each species, with correlation coefficient values above 0.80. In case of *USP10* gene, *C. aethiops* showed a copy number similar to human (Supplementary Figure 2A). We suggest that extra copies for this gene might be acquired during speciation of *C. aethiops* since *M. mulatta* and *P. hamadryas* did not show such duplication, even though they belong to the same Old World Monkeys family (Cercopithecidae). In case of *PAIP1* gene, the two New World Monkeys tested had copy number gene levels similar to those of humans (Supplementary Figure 2B). Therefore, we hypothesize that these duplications might have happened during Platyrrhini evolution. We tried to assess the evolutionary step at which each of the gene amplifications identified took place and the results are shown in Figure 3. Only the statistically significant differences between species were included in the graphic. Interestingly, it seems that most of the gains occurred in the *H. sapiens* ancestors, whereas the other lineages harbor scarce gene amplifications. According to these data, we suggest that gene amplifications did not arise in a single step but duplications would have taken place throughout the primate evolution with an amplification burst in *H. sapiens* speciation. However, this phylogenetic distribution of CNDs is preliminary and more samples of each species would be necessary to confirm it.

2.4. Copy number variations among individuals

Despite the use of DNA pools and combination with Fortna et al. [10] results, there is the possibility that polymorphisms between individuals of each species have impacted our analysis. Fourteen of the 23 HLS CNDs recorded in this study corresponded to regions identified as putative CNP according to the UCSC Genome Browser, March 2006 assembly (Table 1). However, only one of the HLS CNDs (*MSTP9* at 1p36.13) is included in a region considered as potential hotspot of CNV in humans and chimpanzees [30]. Finally, as Fortna et al. [10] carried out individual hybridizations, we checked from their results whether all the individuals of a species had $\log_2$ ratios indicative of CND for each of the genes with HLS CNDs found in the present study. Only one out of the 23 genes (*H. sapiens* transcribed sequence with strong similarity to DRD5) presented one of the individuals of a species with no CND. Overall, these results would reduce the possibility that the CNDs detected in the present study correspond to polymorphisms, at least in human and chimpanzee genomes. This is important as only HLS gains and losses common to all humans are supposed to have promoted human evolution and contributed to the development of HLS phenotypes. If this is true, then only those genes located in regions not associated with CNP would be among the key genes in human-specific evolution. Finally, there is the possibility that even if there is CNP among individuals for a given gene, the lowest human copy number could be higher than the highest non-human primate copy number and therefore, it would still represent a global increase in copy number of that particular gene in human genome *vs.* other great apes genomes.

2.5. Pseudogenes or functional copies?

The gene copies that we detected as LS could represent functional gene copies or pseudogenes. It is noteworthy that pseudogenes for many of these HLS genes have been reported at www.pseudogene.org Human Build 36 [31] (Table 1): *OR2AP9*, *MSTP9, PDE4DIP, FCGR1A, NEK2, ABCB10, NAIP, PMP2, TDG, USP10*, and *FGF7*. Moreover, according to the NCBI database, pseudogenes have been described also for *CDH12* and *AQP7* (Table 1). Overall, 14 out of the 23 HLS genes have known pseudogenes. Chimpanzee pseudogenes have been identified in few of these genes (*PDE4DIP, NEK2*, and *PMP2*) and the number of pseudogenes is lower than the number of pseudogenes in the human genome in

all cases (www.pseudogene.org). What is more, we performed a BLAT (BLAST-like alignment tool) search (http://genome.ucsc.edu) with the full insert sequences obtained at www.ncbi.nlm.nih.gov from the accession number of each cDNA. Only BLAT hits with a score higher than 200 and with more than 90% of identity were selected as positive. All the HLS cDNAs showed one or several closely related copies in the human genome assembly (March, 2006), with the exception of two *H. sapiens* transcribed sequences (H57306 and AI088089) (Supplementary Table 6). In four of the genes (*MSTP9, FCGR1A, PAIP1* and *FGF7*), some of the copies contained internal sequences not present in the cDNA, which could be interpreted as introns, suggesting that the copies were due to gene duplication. According to www.pseudogene.org database three out of these four genes corresponded to duplicated pseudogenes and not processed pseudogenes. Notably, in these four genes the number of BLAT hits was higher than the number of reported pseudogenes, suggesting that some of the duplicated copies might be functional. The other HLS cDNAs (intronless) would correspond most likely to processed copies due to retrotransposition or to single exons (maybe from a functional gene copy). The BLAT alignment was performed also against the March 2006 assembly for the chimpanzee genome and fewer copies of all genes were found compared to the human genome, except in AI089407 with the same number of copies (Supplementary Table 7). However the comparison of our cDNAs with the whole genome assembly (both human and chimpanzee) is limited due to the difficulties entailed by the SD, which are hard to correctly identify and localize in a genome, and therefore are underrepresented in the current assemblies [32, 33]. Thus, the above-mentioned data are likely to be an underestimation.

It is important to distinguish between inactivated non-functional duplicates and fully functional copies, which may have played a key role in the evolution of HLS traits, especially the neuronal function-related genes. Even though pseudogenes are considered to be non-functional, it is interesting to keep in mind that expressed pseudogenes can have regulatory functions, e.g. by increasing mRNA stability of their homologous coding genes [34]. Moreover, the duplication events, besides leading to an increased expression, may lead to tissue specialization and expression diversity, i.e. the new copies would express in tissues not known to express the counterpart gene, as it has been observed for some pseudogenes [35]. Interestingly, a large proportion of human duplicate genes have been observed to diverge rapidly in their spatial expression [36]. Finally, there could be other non-pseudogene copies

that have retained gene function and/or that would have skipped the BLAT detection due to the underrepresentation of SD.

In conclusion, we were able to successfully identify 23 CNDs that occurred specifically in the human lineage and were present in fewer copies in the other great ape genomes examined. Eleven of these 23 genes have only been reported to have undergone HLS duplication in Fortna et al. [10] and in the present study; we confirmed six of them by qPCR. Four out of these six were still HLS after analyzing nine primate species. Most surely, many other genes may have an impact in LS copy number variation between human and great ape species, but the genes identified here are very likely to be *bona fide* LS changes, given the restrictive analysis used. We propose that these gene duplications have a potential to contribute to specific human phenotypes. Finally, it is important to remark that this study was focused mainly in HLS CNDs but further analysis will be done to investigate on CNDs differentiating other primates.

3. MATERIALS AND METHODS

3.1. Array CGH

Three great ape species were examined in the array CGH analysis: common chimpanzee (*Pan troglodytes*) (n = 2), gorilla (*Gorilla gorilla*) (n = 2), and orangutan (*Pongo pygmaeus*) (n = 1). They were pair wise compared to human (*H. sapiens*) (n=3). DNAs were extracted from blood samples, according to standard procedures. Pooled DNA samples were used for chimpanzee, gorilla, and human, in order to ensure the detection of fixed differences between the species, and to eliminate potential interindividual variability.

Customized human cDNA microarrays (The Finnish DNA Microarray Center, Turku, Finland; http://microarrays.btk.fi) were used for the hybridizations [40]. The arrays contained 16,000 annotated genes from the entire genome and procured from the Research Genetics clone library (http://www.resgen.com). The clones were printed in duplicate on poly-L-lysine coated glass slides (The Finnish DNA Microarray Center). Each slide contained a number of reference genes. Strict quality controls were applied to the slides. Array CGH was performed as previously described [41]. Briefly, genomic DNA from reference (human) and test samples

(chimpanzee, gorilla or orangutan) were labeled directly with Cy3 and Cy5 fluorochromes, respectively, by random priming. Labeled DNAs were mixed and hybridized in presence of Cot-1 DNA to array CGH slides. After hybridization, the slides were washed and scanned with an Agilent confocal scanner (Agilent Technologies, Palo Alto, CA). Microarray images were analyzed using Agilent's Feature Extraction software (version 7.1; Agilent Technologies) with the locally weighted linear-regression curve fit option. Log_2 ratios were calculated from Cy3/Cy5 channels and further normalized (lowess dye-normalization). Measurements flagged as unreliable by the Feature Extraction software were excluded from subsequent analysis. Genomic alignment information was retrieved from the University of California at Santa Cruz's Genome Browser database, December 2004 freeze, available at http://genome.ucsc.edu/. Sites with log_2 ratios ± 2 Standard Deviations of the middle 50% quantile of data were selected as representing putative gene CNDs between human and the great ape species tested, provided that both clone replicas were above or under these thresholds. After obtaining a list of HLS CNDs, a second analysis was performed in which array loci were scored as potential variants only if consistent CNDs were observed both in our experiments and in Fortna et al. [10] raw data. This approach ensured the detection of genes with *bona fide* HLS CNDs.

UCSC Genome Browser (March 2006 assembly) and NCBI databases were used to analyze gene functions and gene expression levels in different tissues. UCSC Genome Browser was used also to determine segmental duplications (SD) and copy number polymorphisms (CNP) for each HLS CND. CNP were determined by various methods, such as SNP microarrays, BACs microarrays, ROMA, and deletions from genotype analysis and from haploid hybridization analysis [42-49].

3.2. Quantitative PCR

DNAs used for qPCR were one sample from each species analyzed by array CGH and DNAs isolated from cell lines derived from *H. lar, C. aethiops, P. hamadryas, M. mulatta, L. lagothricha,* and *S. sciureus*, kindly donated by Dr. M. Ponsà (Universitat Autònoma de Barcelona, Spain).

For primer design, human, chimpanzee and macaque genome sequences were used and only identical sequences (after a blast analysis using the database of Reference Genomic

Sequences) were included in the search for primer sequences. Primers for *ABCB10*, *E2F6*, *CDH12*, *PMP2*, *TDG*, *USP10*, *PAIP1*, and *PPIA* (control) genes were designed by using Primer 3 software (http://frodo.wi.mit.edu/) (Supplementary Table 8). They were *in silico* and experimentally verified for specificity, by using *in silico* PCR (http://genome.ucsc.edu/) against human, chimpanzee, orangutan, rhesus and marmoset genomes, and by standard PCR, respectively. In *in silico* PCR, ABCB10 and PAIP1 primers did not amplify in orangutan genome, and any primer amplified with marmoset sequences. QPCR amplification reaction was performed on a Real Time 7300 PCR System using the Power Master Mix PCR SYBR Green (Applied Biosystems, CA). All samples were amplified in triplicates. PCR product amplification could not be obtained in all species with all primers, most probably due to differences *vs.* human genome in the sequences recognized by these primers. Relative quantification was performed by standard curve method for quantification against a control amplicon of the *Cyclophilin A* gene (*PPIA*), following manufacturer instructions (Applied Biosystems). Gene copy numbers of all species were relative to human, set to a value of 1. The t-test was used to carry out statistical comparisons between species by using the SPSS software package (SPSS, Inc.), in order to discard normal variation between samples.

4. ACKNOWLEDGEMENTS

The authors wish to thank Dr. A. Navarro for comments on an earlier version of the manuscript and helpful discussion, and J. Martínez for helping with PCR experiments

References

[1] R. Toder, Y. Xia, E. Bausch, Interspecies comparative genome hybridization and interspecies representational difference analysis reveal gross DNA differences between humans and great apes. Chromosome Res 6 (1998) 487-494.
[2] R.A. Gibbs, et al., Evolutionary and biomedical insights from the rhesus macaque genome. Science 316 (2007) 222-234.
[3] E.S. Lander, et al., Initial sequencing and analysis of the human genome. Nature 409 (2001) 860-921.
[4] T.S. Mikkelsen, et al., Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437 (2005) 69-87.
[5] G.H. Perry, et al., Copy number variation and evolution in humans and chimpanzees. Genome Res 18 (2008) 1698-1710.
[6] D.P. Locke, et al., Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res 13 (2003) 347-357.

[7] H. Kehrer-Sawatzki, D.N. Cooper, Structural divergence between the human and chimpanzee genomes. Hum Genet 120 (2007) 759-778.
[8] A. Kallioniemi, et al., Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258 (1992) 818-821.
[9] Z. Cheng, et al., A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437 (2005) 88-93.
[10] A. Fortna, et al., Lineage-specific gene duplication and loss in human and great ape evolution. PLoS Biol 2 (2004) E207.
[11] V. Goidts, et al., Identification of large-scale human-specific copy number differences by inter-species array comparative genomic hybridization. Hum Genet 119 (2006) 185-198.
[12] G.M. Wilson, et al., Identification by full-coverage array CGH of human DNA copy number increases relative to chimpanzee and gorilla. Genome Res 16 (2006) 173-181.
[13] L. Dumas, et al., Gene copy number variation spanning 60 million years of human and primate evolution. Genome Res 17 (2007) 1266-1277.
[14] T. Marques-Bonet, et al., A burst of segmental duplications in the genome of the African great ape ancestor. Nature 457 (2009) 877-881.
[15] F. Preuss, et al., Drosophila doubletime mutations which either shorten or lengthen the period of circadian rhythms decrease the protein kinase activity of casein kinase I. Mol Cell Biol 24 (2004) 886-898.
[16] Y. Gilad, et al., Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. Genome Res 15 (2005) 674-680.
[17] N. Kouprina, et al., The SPANX gene family of cancer/testis-specific antigens: rapid evolution and amplification in African great apes and hominids. Proc Natl Acad Sci U S A 101 (2004) 3077-3082.
[18] D.B. Zimonjic, et al., Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. Proc Natl Acad Sci U S A 94 (1997) 11461-11465.
[19] J. Amann, M. Valentine, V.J. Kidd, J.M. Lahti, Localization of chi1-related helicase genes to human chromosome regions 12p11 and 12p13: similarity between parts of these genes and conserved human telomeric-associated DNA. Genomics 32 (1996) 260-265.
[20] A. Ciccodicola, et al., Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. Hum Mol Genet 9 (2000) 395-401.
[21] M.C. Popesco, et al., Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. Science 313 (2006) 1304-1307.
[22] J.M. Sikela, The jewels of our genome: the search for the genomic changes underlying the evolutionarily unique capacities of the human brain. PLoS Genet 2 (2006) e80.
[23] P.J. Brooks, C. Marietta, D. Goldman, DNA mismatch repair and DNA methylation in adult brain neurons. J Neurosci 16 (1996) 939-945.
[24] C. SLI, Highly significant linkage to the SLI1 locus in an expanded sample of individuals affected by specific language impairment. Am J Hum Genet 74 (2004) 1225-1238.
[25] A.G. Clark, et al., Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302 (2003) 1960-1963.
[26] D.L. Maresco, et al., Localization of FCGR1 encoding Fcgamma receptor class I in primates: molecular evidence for two pericentric inversions during the evolution of human chromosome 1. Cytogenet Cell Genet 82 (1998) 71-74.

[27] J.M. Szamalek, et al., Characterization of the human lineage-specific pericentric inversion that distinguishes human chromosome 1 from the homologous chromosomes of the great apes. Hum Genet 120 (2006) 126-138.
[28] A. Courseaux, et al., Segmental duplications in euchromatic regions of human chromosome 5: a source of evolutionary instability and transcriptional innovation. Genome Res 13 (2003) 369-381.
[29] J.P. Demuth, et al., The evolution of mammalian gene families. PLoS ONE 1 (2006) e85.
[30] G.H. Perry, et al., Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci U S A 103 (2006) 8006-8011.
[31] J.E. Karro, et al., Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res 35 (2007) D55-60.
[32] J.A. Bailey, et al., Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11 (2001) 1005-1017.
[33] J. Cheung, et al., Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. Genome Biol 4 (2003) R25.
[34] S. Hirotsune, et al., An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423 (2003) 91-96.
[35] E.S. Balakirev, F.J. Ayala, Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet 37 (2003) 123-151.
[36] K.D. Makova, W.H. Li, Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res 13 (2003) 1638-1645.
[37] P. Khaitovich, W. Enard, M. Lachmann, S. Paabo, Evolution of primate gene expression. Nat Rev Genet 7 (2006) 693-702.
[38] M.V. Han, et al., Adaptive evolution of young gene duplicates in mammals. Genome Res 19 (2009) 859-867.
[39] G.H. Perry, et al., Diet and the evolution of human amylase gene copy number variation. Nat Genet 39 (2007) 1256-1260.
[40] H. Vauhkonen, et al., Characterizing genetically stable and unstable gastric cancers by microsatellites and array comparative genomic hybridization. Cancer Genet Cytogenet 170 (2006) 133-139.
[41] G. Armengol, et al., Genomic imbalances in Schistosoma-associated and non-Schistosoma-associated bladder carcinoma. An array comparative genomic hybridization analysis. Cancer Genet Cytogenet 177 (2007) 16-19.
[42] D.F. Conrad, et al., A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38 (2006) 75-81.
[43] D.A. Hinds, et al., Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38 (2006) 82-85.
[44] A.J. Iafrate, et al., Detection of large-scale variation in the human genome. Nat Genet 36 (2004) 949-951.
[45] D.P. Locke, et al., Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79 (2006) 275-290.
[46] S.A. McCarroll, et al., Common deletion polymorphisms in the human genome. Nat Genet 38 (2006) 86-92.
[47] R. Redon, et al., Global variation in copy number in the human genome. Nature 444 (2006) 444-454.
[48] J. Sebat, et al., Large-scale copy number polymorphism in the human genome. Science 305 (2004) 525-528.

[49] A.J. Sharp, et al., Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77 (2005) 78-88.

FIGURE LEGENDS

Figure 1. Venn diagram with the number of CNDs detected in each species genome *vs.* the human genome and the CNDs shared by two or three species (intersection spaces). A plus or minus sign before the number denotes whether those genes had an increase or decrease in copy number, respectively. A total of 156 copy number changes were detected at 150 loci. This difference is due to loci with increases in one species and decreases in other/s species.

Figure 2. Mean relative copy number measured by quantitative PCR of: A) *ABCB10*; B) *E2F6*; C) *CDH12*; and D) *TDG* genes, in 10 primate species. All species were relative to human, set to a value of 1. PCR product amplification could not be obtained in all species with all primers, most probably due to differences *vs.* human genome in the sequences recognized by these primers. *Columns*, mean values for three replicates; *bars*, standard errors.

Figure 3. Graphic showing the evolution steps in which most probably amplification of each of the following genes occurred *vs.* the ancestor, according to qPCR results: 1, *ABCB10*; 2, *E2F6*; 3, *CDH12*; 4, *TDG*; 5, *USP10*; 6, *PAIP1.* In case of uncertainty, the amplification event has been located at the most ancient ancestor with a question mark besides the number of the gene.

Figure 1

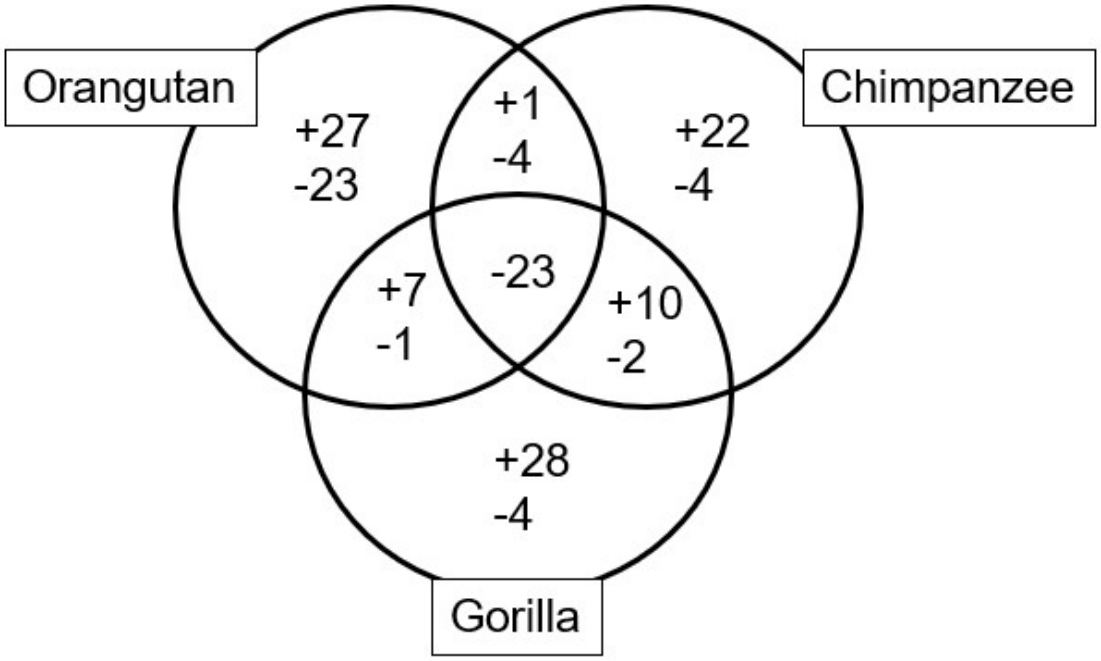
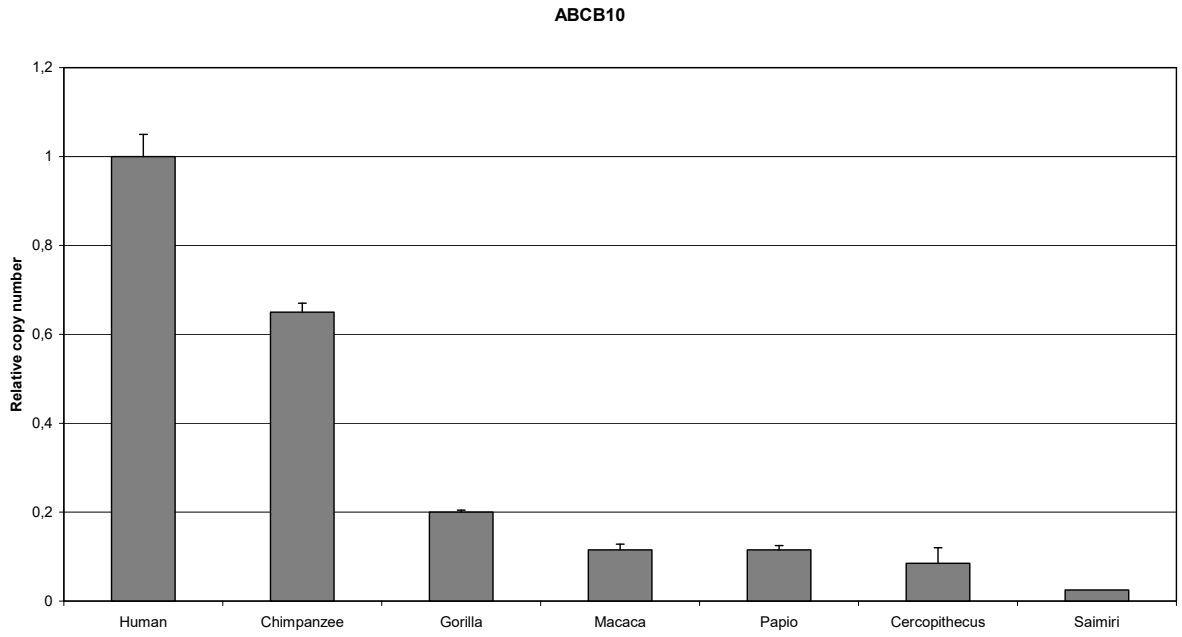
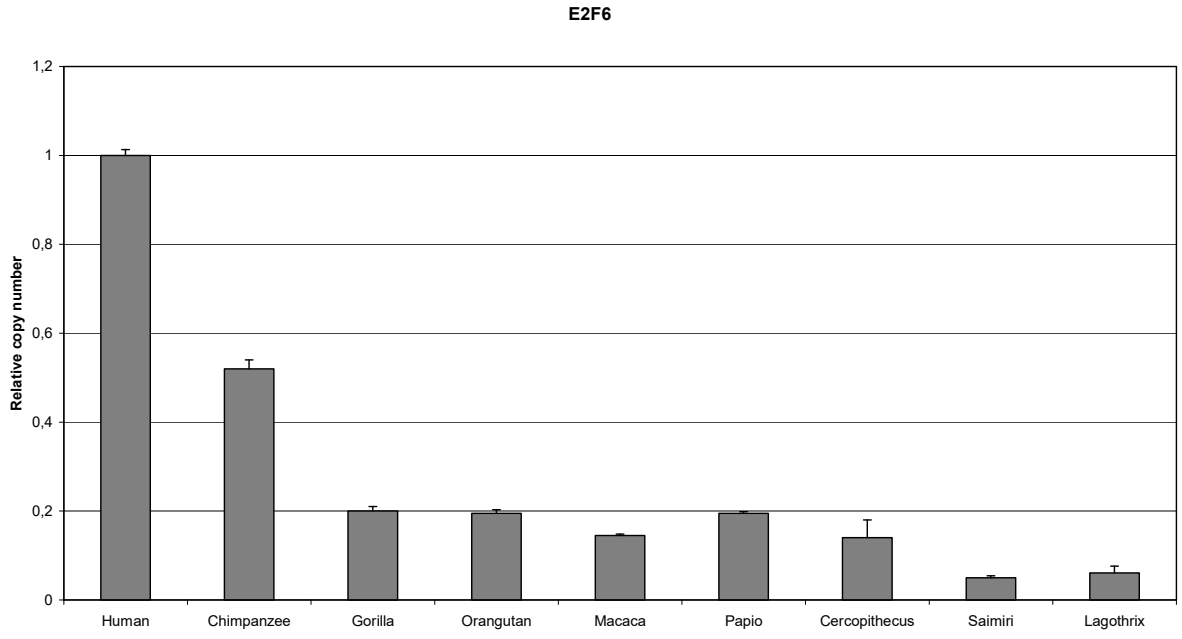


Figure 2

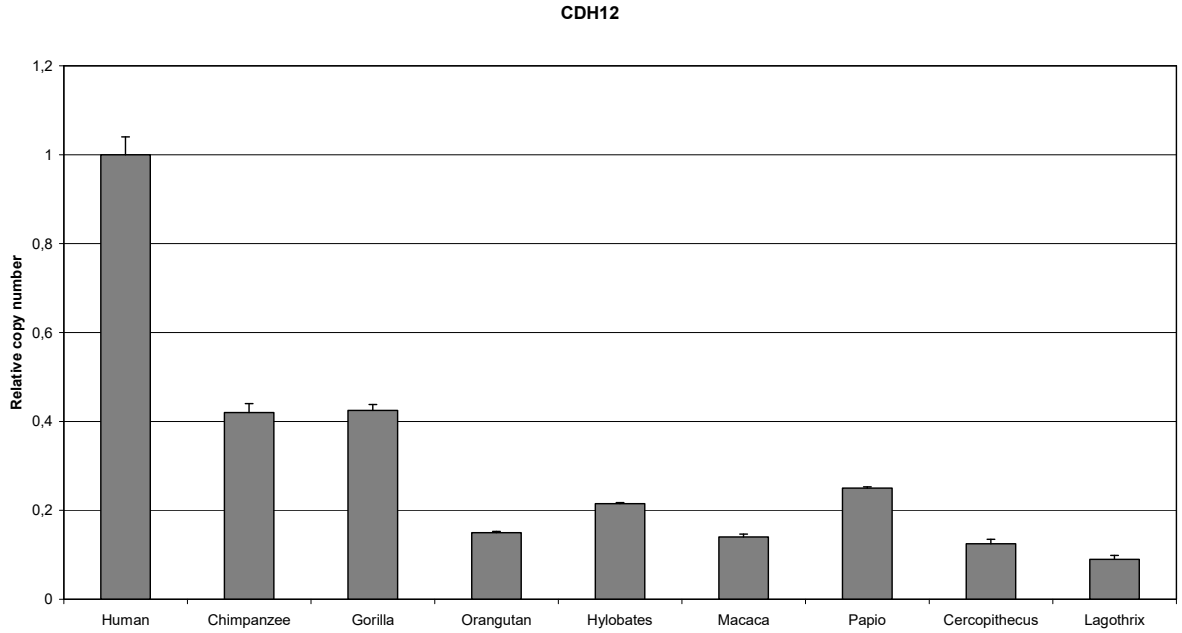
A


ABCB10
Relative copy number
1,2
1
0,8
0,6
0,4
0,2
0
Human
Chimpanzee
Gorilla
Macaca
Papio
Cercopithecus
Saimiri

B


E2F6
Relative copy number
1,2
1
0,8
0,6
0,4
0,2
0
Human
Chimpanzee
Gorilla
Orangutan
Macaca
Papio
Cercopithecus
Saimiri
Lagothrix

C


CDH12
Relative copy number
1,2
1
0,8
0,6
0,4
0,2
0
Human
Chimpanzee
Gorilla
Orangutan
Hylobates
Macaca
Papio
Cercopithecus
Lagothrix

D

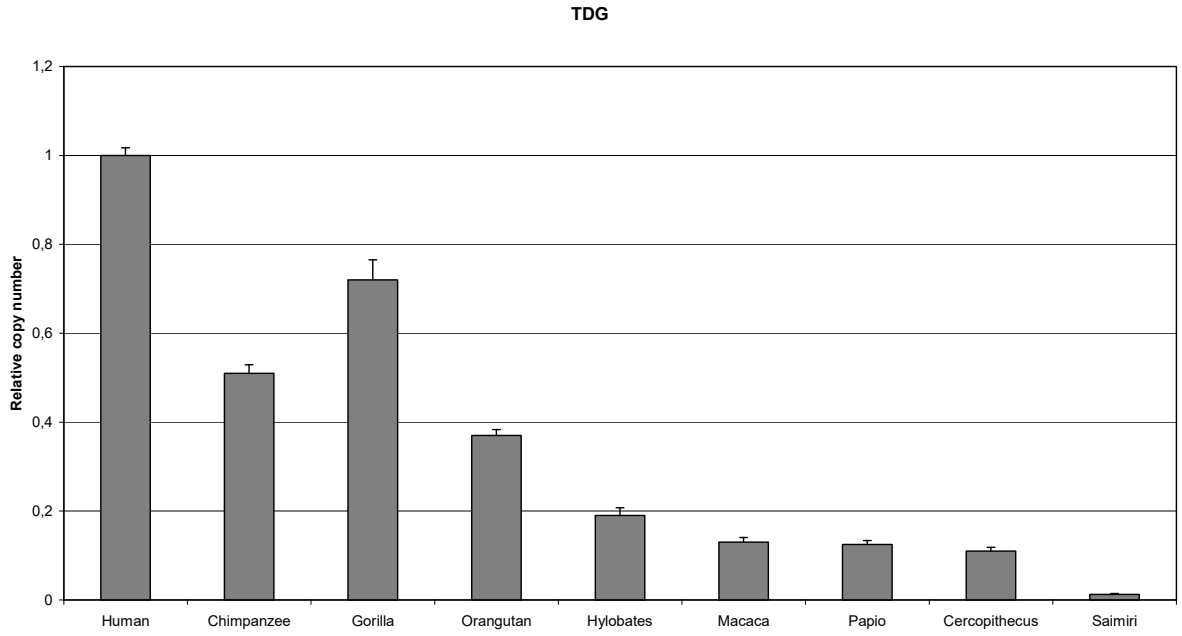

TDG
Relative copy number
1,2
1
0,8
0,6
0,4
0,2
0
Human
Chimpanzee
Gorilla
Orangutan
Hylobates
Macaca
Papio
Cercopithecus
Saimiri

Figure 3

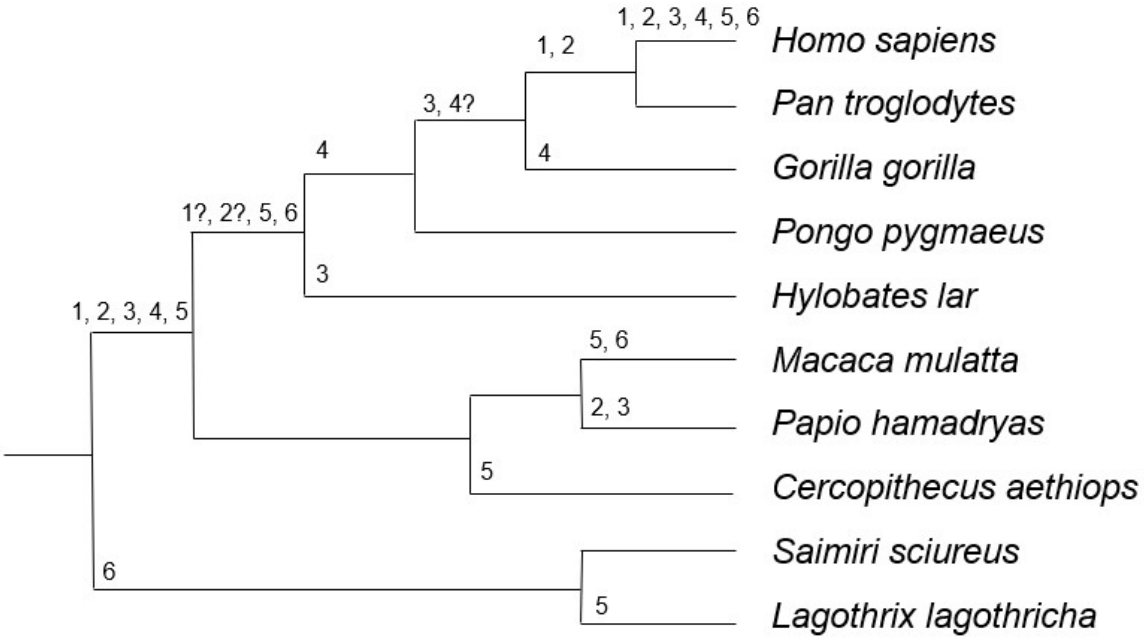

1, 2, 3, 4, 5, 6
Homo sapiens
1, 2
Pan troglodytes
3, 4?
4
Gorilla gorilla
4
1?, 2?, 5, 6
Pongo pygmaeus
3
Hylobates lar
1, 2, 3, 4, 5
5, 6
Macaca mulatta
2, 3
Papio hamadryas
5
Cercopithecus aethiops
Saimiri sciureus
6
5
Lagothrix lagothricha

Table 1. Genes showing HLS variation in copy number. All of them were present in more copies in humans than in the non-human primates studied. Genes marked in bold are those not previously described as human-specific.

Gene name	Gene symbol	Accession[a]	Cytoband
Macrophage stimulating, pseudogene 9	*MSTP9*	AA707273	1p36.13
Homo sapiens transcribed sequence with strong similarity to dopamine receptor D5 (DRD5)		AI148329	1q21.1
H. sapiens mRNA similar to family with sequence similarity 72, member A (FAM72A)		R22949 AA628867 AI271431	1q21.1
Phosphodiesterase 4D interacting protein (myomegalin)	*PDE4DIP*	AA443157	1q21.1
Fc fragment of IgG, high affinity Ia, receptor (CD64)	*FCGR1A*	AA453258	1q21.2
NIMA (never in mitosis gene a)-related kinase 2	*NEK2*	W93379	1q32.3
ATP-binding cassette, sub-family B (MDR/TAP), member 10	***ABCB10***	**R83876**	**1q42.13**
E2F transcription factor 6	***E2F6***	**AA935533**	**2p25.1**
Cadherin 12, type 2 (N-cadherin 2)	***CDH12***	**N74018 AA418564**	**5p14.3**
***H. sapiens* transcribed sequences**		**H40480**	**5q13.2**
Neuronal apoptosis inhibitory protein	*NAIP*	AA621150	5q13.2
***H. sapiens* transcribed sequences**		**H57306**	**5q13.2**
Occludin	*OCLN*	AI291184	5q13.2
Olfactory receptor family 2, subfamily A member 9 pseudogene	*OR2A9P*	AA001222	7q35
Peripheral myelin protein 2	***PMP2***	**AI089407**	**8q21.13**
***H. sapiens* transcribed sequences**		**AI088089**	**9p12[b]**
***H. sapiens* transcribed sequences**		**H15704**	**9p12[b]**
Aquaporin 7	*AQP7*	H27752	9p13.3
H. sapiens similar to DEAD/H box polypeptide 11 (CHL1-related helicase gene-1) (DDX11)		AA402879	12p13.31
Thymine-DNA glycosylase	***TDG***	**AA496947**	**12q23.3**
Fibroblast growth factor 7 (keratinocyte growth	*FGF7*	AA009609	15q21.2[c]

factor)			
Ubiquitin specific protease 10	***USP10***	**AA455233**	**16q24.1**
Poly(A) binding protein interacting protein 1	***PAIP1***	**AA598533** **H92758**	**17p11.2**

[a] In some cases, redundant cDNAs corresponded to a single gene

[b] cytoband is 9p11.2 according to March 2006 Assembly

[c] cytoband is 15q21.1 according to march 2006 Assembly

Table 2. Characteristics of the genes with HLS CNDs [a].

Gene	Function/ Expression	Gene clusters	Pseudo-genes[b]	SD	CNP	CNDs previously reported[c]
MSTP9	Neuronal		yes	3	yes	[12]
Similar to DRD5	Neuronal	inv(1)	yes	2	yes	[9, 27]
Similar to FAM72A		inv(1)	nk	2	yes	[27]
PDE4DIP	Neuronal	inv(1)	yes	5	yes	[9, 13, 27]
FCGR1A	Neuronal	inv(1)	yes	2	yes	[9, 11, 14, 27]
NEK2	Reproduction Neuronal		yes	0	no	[13]
ABCB10			yes	4	no	
E2F6	Neuronal		nk	3	no	
CDH12	Neuronal	SMA region	yes	5	yes	
Ts (H40480)		SMA region	nk	2	yes	
NAIP	Neuronal	SMA region	yes	2	yes	[9, 12, 14]
Ts (H57306)		SMA region	nk	2	yes	
OCLN	Neuronal	SMA region	nk	2	yes	[9, 13]
OR2A9P			yes	1	yes	[11]
PMP2	Neuronal		yes	0	no	
Ts (AI088089)	Parathyroid		nk	13	yes	
Ts (H15704)	Cochlea		nk	7	yes	
AQP7	Reproduction		yes	4	no	[9, 11, 13]
Similar to DDX11			nk	2	yes	[9, 19]
TDG	Neuronal		yes	2	no	
FGF7	Neuronal		yes	8	no	[13, 18]
USP10	Neuronal		yes	3	no	
PAIP1			nk	2	no	

[a] Ts, transcribed sequence; nk, not known; SD, segmental duplications reported in UCSC Genome Browser; CNP, copy number polymorphisms reported in UCSC Genome Browser; CND, copy number differences

[b] according to www.pseudogene.org and to NCBI

[c] apart from [10]