**Characterization and evolution of the novel gene family *FAM90A* in primates originated by multiple duplication and rearrangement events**

Nina Bosch[1], Mario Cáceres[1], Maria Francesca Cardone[2], Anna Carreras[1], Ester Ballana[1], Mariano Rocchi[2], Lluís Armengol[1], Xavier Estivill[1, 3, *]

[1]Genes and Disease Program, Center for Genomic Regulation (CRG-UPF), Barcelona, Catalonia, Spain
[2]Department of Genetics and Microbiology, University of Bari, Bari, Italy
[3]Experimental and Health Sciences Department, Pompeu Fabra University, Barcelona, Catalonia, Spain

***Correspondence should be addressed to:**
Xavier Estivill MD, PhD, Genes and Disease Program, Centre for Genomic Regulation (CRG), Plaça Charles Darwin s/n, (Carrer Dr Aiguader, 88) PRBB Building, Room 521, 08003 Barcelona, Catalonia, Spain, Tel. +3493 316 0159, FAX +3493 316 0099, e-mail: xavier.estivill@crg.es

**Abstract**

Genomic plasticity of human chromosome 8p23.1 region is highly influenced by the presence of two groups of complex segmental duplications (SDs), termed REPD and REPP that mediate different kind of rearrangements. Part of the difficulty to explain the wide range of phenotypes associated to 8p23.1 rearrangements is that REPP and REPD are not yet well characterized, probably due to their polymorphic status. Here we describe a novel primate specific gene family, named *FAM90A* (family with sequence similarity 90), found within these SDs. According to the current human reference sequence assembly (Build 36), the *FAM90A* family includes 24 members distributed along 8p23.1 region and another one on chromosome 12p13.31, with variation in copy number (CNV) between individuals. The different members can be classified into two subfamilies, I and II, which differ in their upstream sequences and the first 5' UTR exon, but they mostly share the same ORF. Sequence analysis and comparative FISH studies showed that the *FAM90A* Subfamily II suffered a big expansion in the hominoid lineage, whereas Subfamily I members were likely generated some time around the divergence of orangutan and African great apes by a fusion process. We also show that *FAM90A* genes are ubiquitously expressed and the analysis of the Ka/Ks ratios provides evidence of functional constraint of some *FAM90A* genes in all species. The characterization of the *FAM90A* novel gene family contributes to a better understanding of the structural polymorphism of human chromosome 8p23.1, and constitutes a good example of how SDs, CNVs and rearrangements within themselves can promote the formation of new gene sequences that could have potential functional consequences.

## Introduction

Genome duplication is widely accepted as one of the main mechanisms for the birth of new genes and the expansion of gene families, as well as an opportunity to adopt new functions (1). Many genes are related to the formation of highly identical duplicated regions, such as the homeobox (2), globins (3) or primate-specific morpheus gene families (4). These tandem gene clusters are likely the result of misaligned homologous recombination between homologous sequences (5). Moreover, it has been postulated that segmental duplications (SDs - i.e. duplicated segments of genomic DNA of size >1 kb and that share >90% of sequence identity) would mediate genomic rearrangements via non-allelic homologous recombination (NAHR) (6). So far, a significant number of genomic disorders, such as the Williams-Beuren, Prader-Willi or DiGeorge syndromes, are known to arise from NAHR between SDs (7-9). In addition, genomic copy number gains or losses, also known as copy number variants (CNVs), are frequently the result of this kind of recombination events. Recent studies correlate the presence of SDs with the localization of CNVs(10) and evolutionary breakpoints in primates and other mammals (11-13).

Besides genomic gains and losses, depending on the relative position and orientation of SDs, NAHR can also lead to the inversion of the genomic intervening sequence. Occasionally, these inversions are involved in human disease, like in Hunter syndrome (14) or hemophilia A (15). Several other inversion variants do not have any evident phenotypic effects for the carriers, but result in an increased risk of transmission of rearrangements to the offspring or confer a reproductive advantage (16, 17, Stefansson, 2005 #524). The polymorphic inversion affecting the 8p23.1 region, found in 26% of the European and Japanese general population (18, 19), has no apparent phenotypic consequences for the carriers, but incorrect pairing between normal and

inverted chromosomes during meiosis leads to different types of rearrangements that have been associated with a wide range of phenotypes in the offspring (20-24). Furthermore, duplications affecting the 8p23.1 segment can also be related to the presence of benign euchromatic variants, which have been repeatedly reported in the literature (25-28).

The underlying basis for all these rearrangements is likely the presence of two sets of complex SDs, REPP and REPD, on the proximal and distal portions of 8p23.1. Although several studies have lately improved the characterization of the REPP and REPD genomic architecture, the polymorphic status of different components within these SDs complicates the analysis (19, 29). Even on the last effort to obtain an accurate sequence of human chromosome 8 by the International Human Genome Sequence Consortium, two of the four gaps remaining in this chromosome are still located within REPP and REPD, and they seem to be refractory to current cloning and mapping technologies (30). Incomplete and incorrectly assembled sequences are common to regions containing SDs, and hence these genomic fragments deserve special attention in order to obtain a proper characterization (31). Another phenomenon that increases the difficulty to depict the 8p23.1 region is the presence of euchromatic variants corresponding to copy number variants (CNVs) of alpha as well as beta defensin gene clusters (29, 32-35). Frequently, polymorphic genomic regions are related with hotspots in evolution, and they often contain genes related to adaptation to the environment (36). In the case of the 8p23.1 region, we find two gene families that have expanded in mammals, such as olfactory receptor genes and defensins . Evidences for the possible role that defensin genes have played in evolution is supported by the high ratio of non synonymous to synonymous substitutions they harbor, suggesting that positive selection has been acting in this region. Furthermore, the 8p23.1 region also shows a strikingly

high polymorphism rate in the human population, that is just exceeded by some regions of the Y chromosome (30).

In our effort to better understand the evolution and dynamics of these complex regions, we have identified a novel gene family (*FAM90A*) that contains at least 24 members per haploid genome (according to the hg18 UCSC reference assembly) distributed along the REPP and REPD plus one single copy on chromosome 12p13.31. The clustered *FAM90A* members represent a new CNV that brings in additional complexity to the SDs on 8p23.1. In the current work, we examine the sequence of the different copies, identify the functional ones and describe two different subfamilies of these genes. We also report the presence of a variable number of *FAM90A* members at the genomic level in individuals of the general population and in non-human primates, as well as the presence of transcripts corresponding to some of the family members. A mechanism by which these copies could have evolved and expanded through the primate lineage is hypothesized.

## Results

### *In-silico* analysis of the low copy repeats in the 8p23.1 region

The complex structure and variation of REPD and REPP SDs compromises the reliability of the assembly of the 8p23.1 region. Our first approach to obtain a better characterization of the 8p23.1 SDs was an *in-silico* analysis of the REPD and REPP sequences in the current reference human genome sequence (hg18). As previously described, these SDs are formed by several segments that are duplicated in many other parts of the genome, including olfactory receptors (OR), the largest gene super family in the human genome (37), and genes related to the immune response, the alpha- and beta-defensin genes, which are clustered around the 100-kb gap placed within REPD (34)

(Figure 1). In addition to the existing defensin gene clusters, the alignment of the REPD and REPP sequences against themselves using the Pipmaker algorithm (38) allowed us to detect a 7.6-kb fragment that was tandemly repeated at multiple locations. Along REPD, four different clusters, A, B, C and D, were found, which contain 6, 5, 8 and 3 copies of the 7.6 kb module (HsaCopy1-22), respectively (Figure 1). Furthermore, within REPP there are two additional individual sequences that share high identity with 6.4 kb of the 7.6-kb module (HsaCopy23-24, Figure 1), totaling the current 24 copies of this sequence present in the latest assembly of the 8p23.1 region. Interestingly, when we focused on the genomic architecture of the REPD clusters, cluster A and B are in opposite orientation and located at each side of a group of beta-defensin genes, and the same happens with cluster C and D (Figure 1). Therefore, the whole genomic region appears to be duplicated and distributed as specular images on both sides of the gap.

**Characterization of the novel *FAM90A* gene family**

A more detailed analysis of the 7.6-kb module revealed that it is composed of a unique sequence of 6.3 kb that includes a LINE repetitive element (L1MB3) and a complete copy of a LTR5A at the 3' end, corresponding to the ERVK family of endogenous retrovirus. In order to screen if this module resembled any other known human sequence, identity searches were performed against the human genome sequence and the RefSeq database. Besides the copies in the 8p23.1 region, the best blast hit corresponded to the family with sequence similarity 90 member A1 (*FAM90A1*) gene (*GenBank:* NM_018088), located as a single copy on chromosome 12p13.31. The transcribed portion of this gene (6,342 bp) shares 96% nucleotide identity with the two copies found in REPP and more than 93% identity with the 7.6-kb module repeated in REPD. Thus, these sequences constitute a novel gene family (*FAM90A*) in the human

genome, and we have named the different members in 8p23.1 as HsaCopy1-24 (distal to proximal) to avoid possible confusion with the current nomenclature (Supplementary Table 1).

Taking as reference the annotated copy *FAM90A1* on chromosome 12p13.31, which contains 6 exons and results in a 2,342 bp-long mRNA, we predicted the gene structure of the 24 copies on 8p23.1 SDs (Figure 2). Based on the gene structure and sequence similarity, the different members of this family could be divided into two subfamilies (I and II). Subfamily I includes *FAM90A1* and both single copies on REPP (HsaCopy23-24), and Subfamily II is formed by the rest of the members on REPD SDs (HsaCopy1-22). The main difference between the two groups is that the initial 1,036 bp of the *FAM90A1* gene, which include the first untranslated exon (exon 1), are exclusive of the three Subfamily I copies (Figure 2). This 1-kb sequence, which contains 302 bp homologous to an AluSx element and 98 bp related to a MIRb repetitive element, is found at multiple locations on different chromosomes, including an intron of the *ALG1* gene on chromosome 16, and constitutes part of a SD with at least 8 additional copies in the human genome. Members of Subfamily II have instead a 1,244 bp sequence that is only found associated with *FAM90A* members (Figure 2).

From the predicted mRNA sequence of each *FAM90A* member, we have also determined the coding sequence conservation in the different 8p23.1 copies using the SIXFRAME tool. The coding sequence of the *FAM90A1* gene consists of four exons (exons 3 to 6) and has an ORF with the potential to encode for a protein of 464 amino acids [Swiss-Prot Q9NVZ6]. This ORF is conserved in most of the 8p23.1 copies, indicating that they also have protein coding capacity (Supplementary Table 1 and Figure 2).

**Variation in *FAM90A* clusters in humans**

After the identification of the *FAM90A* clusters in the human genome reference sequence, we pursued the experimental characterization of the novel multiple-copy gene family in 20 unrelated individuals from the general population using pulsed-field gel electrophoresis (PFGE) analysis and Southern blotting (Figure 3). Genomic DNA was digested with the *Acc*65I restriction enzyme that allowed us to isolate the different clusters on REPD, plus the two single copies on REPP and *FAM90A1* on 12p13.31. As a probe, we used a 700-bp DNA fragment corresponding to the second intron of *FAM90A1*. The expected restriction pattern from the reference assembly included seven different fragments, ranging from 77 kb to 14 kb. Interestingly, the PFGE patterns showed high variability between individuals (Figure 3), indicating that there are differences in the structure of this region and that *FAM90A* members could be polymorphic in copy number in the human population.

To confirm these results we examined the available information from the complete sequences of other human BAC clones not included in the genome assembly. By BLAT analysis of the 7.6 kb module against non-redundant database we found three BAC sequences with entire *FAM90A* Subfamily II clusters that have different number of copies than in the reference human genome sequence (Supplementary Table 2). These clones likely belong to a different allele to the human reference sequence, although we can not discard the possibility that they represent additional copies of the clusters on the 8p23.1 region or part of the non-assembled genomic material located on the REPD and REPP gaps. Therefore, these results independently stress the existence of wide variability in the number of *FAM90A* copies in human chromosomes.

**Expression of *FAM90A* gene family members**

In order to confirm that *FAM90A* members are transcribed, we tested the tissue expression of these genes by RT-PCR from total adult RNA of 12 different tissues with primers binding to exons 3 and 4 of most *FAM90A* copies. A fragment of the expected size of the mRNA (351 bp) was amplified in all tested tissues (Figure 4). We also repeated the same RT-PCR with RNA from lymphoblastoid cell lines from seven individuals from the general population. The expected fragment from the *FAM90A* genes mRNA was detected in all individuals, suggesting that these genes are widely expressed in humans.

To investigate the expression of the two different *FAM90A* subfamilies, we searched for ESTs in public databases supporting the expression of particular *FAM90A* copies. Besides the reference mRNA of *FAM90A1* (12p13.31), there were several additional full-length mRNA and ESTs matching the exons of this gene. The main difference with the genic structure was that approximately half of these sequences include an alternatively-spliced exon between exons 2 and 3 (Figure 2). Moreover, there was evidence of expression of exons 2-6 of HsaCopy23 and 24 on 8p23.1 proximal duplicons from two additional ESTs (*GenBank:* AL832996 and CX762572), and one of them was fully sequenced by us. These ESTs confirmed the predicted exon-intron structure of the REPP copies, with the exception of the use of a new splicing donor site in exon 2. Conversely, there were many ESTs corresponding to different parts of the repeated module in the REPD clusters, including exons, introns and LTR, and ESTs that span the beginning and end of adjacent modules.

The tissue expression of the Subfamily I and Subfamily II members was further studied by RT-PCR with three different primer pairs which are specific, respectively, for genes *FAM90A1*, HsaCopy23-24 on REPP, and most Subfamily II copies on REPD (Figure 4). Consistent with the previous results, *FAM90A1* was expressed in all 13

tissues tested, whereas we could not detect HsaCopy23-24 expression in heart, testis, placenta and prostate. Regarding expression of Subfamily II members, we detected a clear band of the expected size on all tissues with the exception of kidney and lung, whereas a larger size band, corresponding perhaps to an alternative transcript, was observed in prostate.

### *FAM90A* genes in primates

To investigate the presence of members of this family throughout mammals, we performed exhaustive similarity searches of the *FAM90A* nucleotide and protein sequences against the available genome assemblies and non-redundant sequences in the databases. No significant similarity to *FAM90A* was found in non-primate species, but complete or partial copies of this gene were identified in chimpanzees, rhesus macaque and baboon, indicating that this family is exclusive to the primate lineage.

Due to the complexity of the REPD and REPP SDs, the syntenic region to 8p23.1 in the chimpanzee genome is not well resolved yet and most of the sequences homologous to *FAM90A* Subfamily II members are only partial and include sequencing gaps. However, there are two chimpanzee BACs, CH251-740L16 (*GenBank*: AC183981), which corresponds to human clusters B or C, and CH251-647K5 (*GenBank*: AC184710), which appears to be formed by the fusion of two highly-rearranged clusters in opposite orientation, that include respectively 7 (PtrCopy1-7) and 6 (PtrCopy8-13) full-length copies of Subfamily II. In addition, we found homologous copies to *FAM90A* Subfamily I members in chromosome 12 (PtrCopy14), chromosome 8 (PtrCopy15) and chromosome 11 (PtrCopy16) of the current chimpanzee genome assembly. Thus, organization of this gene family in chimpanzees is very similar to that in humans.

A different scenario is found in rhesus macaque and baboon. In the rhesus macaque genome assembly there are only two ~4 kb fragments with sequence similarity to *FAM90A* Subfamily II, one of which is represented in two overlapping baboon BACs (*GenBank*: AC116559 and AC116558). In addition, a rhesus contig sequence of 14.5 kb (*GenBank*: NW_001158155) includes a Subfamily II copy (MmuCopy1) flanked by smaller fragments that match the end and the beginning of another copy, respectively. The arrangement of all these sequences resembles the Subfamily II *FAM90A* clusters in humans and chimpanzees. However, in none of the rhesus or baboon sequences there is an LTR inserted at the end of the gene. Moreover, no Subfamily I copies have been identified in these species. The only position in the rhesus genome with similarity to the initial 1028-bp fragment of this family that is duplicated in humans is located in the *ALG1* gene intron at chromosome 20 (syntenic to human chromosome 16). Therefore, although Subfamily II members precede the divergence of Old World Monkeys (OWM) and hominoids, the expansion of these genes and the generation of Subfamily I members could have occurred in hominoids (see Discussion). In fact, preliminary BLAST searches with the available whole genome shotgun sequence traces of orangutan suggests that there are no Subfamily I copies in this species either, and that this gene arrangement is more likely specific of African great apes.

**Genomic distribution of *FAM90A* genes in primates**

Four human BACs and a 4.8 kb PCR-amplified *FAM90A* fragment were used as probes to investigate the chromosomal distribution and the expansion of *FAM90A* gene family along the primate lineage by FISH on human, chimpanzee, gorilla, orangutan and rhesus macaque metaphases. Representative results of comparative FISH experiments from primate chromosome spreads are shown in Supplementary Figures 1 and 2, and a

summary of the chromosome location obtained by all probes is reported in Table 2. The BAC clones corresponding to the *FAM90A1* region (12p13.31) and the REPP SDs (8p23.1) produced very similar results in human, chimpanzee and gorilla, with signals on several chromosomes that match the known locations of the SDs that span the region in the human and chimpanzee genome assemblies. Interestingly, in orangutan and rhesus macaque, these BACs did not hybridize to the homologue of human chromosome 12, and instead there were signals in the homologues of chromosome 8 and 16. A quite different FISH pattern was observed with the BACs of the *FAM90A* REPD clusters and with the 4.8 kb *FAM90A* fragment on interphase nuclei, where several different signals could be distinguished on human and chimpanzee chromosome 8, while single signals appeared on gorilla and orangutan. In the case of rhesus macaque we did not find any signal probably due to the sequence divergence and the resolution limit of the FISH technique. These results suggest that humans and chimpanzees have suffered a greater expansion of the gene than the rest of the species.

To better elucidate the architecture of the *FAM90A* clusters in non-human primates, we performed PFGE Southern experiments in chimpanzee, gorilla and orangutan using the same probe as in the human experiments (Supplementary Figure 3). In all three great ape species fragments with homology to *FAM90A* can be seen, but their sizes and numbers are different compared to humans. Gorilla and orangutan has only two fragments of high molecular weight, while chimpanzee had also some smaller bands that could be shared with those found on human individuals.

Finally, a more accurate estimate of the number of *FAM90A* copies in each primate species was obtained by real-time quantitative PCR analysis. Samples included genomic DNA from human, chimpanzee, bonobo, gorilla and orangutan. The primers and probe were initially designed in regions conserved between the human *FAM90A*

copies, but the availability of chimpanzee sequence subsequently showed that one of the primers had mismatches with most chimpanzee copies. In addition, sequence divergence between humans and macaques made impossible to design proper common primers and probe to be used with these species. To control for differences in DNA concentration between samples, *FAM90A* copy number was estimated in each species in comparison to a single copy locus corresponding to an ultra conserved region from human chromosome 6p21(see Methods). Orangutan was the species presenting a lower number of *FAM90A* copies, and we detected, respectively, a 14.6, 4.2, 8.4 and 9.0-fold increase in human, chimpanzee, bonobo and gorilla with respect to the orangutan. Thus, consistent with the above sequence analysis and experimental results, it is likely that an expansion of the *FAM90A* gene family in African great apes has occurred, although the real number of copies in non-human primates could be underestimated due to sequence changes affecting the primers or probe.

**Evolutionary analysis of *FAM90A* family genes**

To assess the relationship between all the *FAM90A* members identified in primates, we carried out a phylogenetic analysis of different parts of the sequence of these genes. Figure 5 shows the Neighbor-Joining tree based on the common sequence to both *FAM90A* subfamilies, except the LTR, using the rhesus MmuCopy1 as outgroup, but similar results were obtained with other methods. According to this, Subfamily II members in humans and chimpanzees form two separate clades that share a common origin. In addition, *FAM90A* copies located in the same cluster tend to be more closely related, forming groups with high bootstrap values (*e.g.*, HsaCopy1-5 or HsaCopy13-18), although there are exceptions involving mostly the copies located at the ends of the clusters (*e.g.*, HsaCopy12, 19 or 20). Finally, most of the Subfamily I members do not

group together in the tree and instead appear to be located in independent branches. This is probably a result of the relatively high divergence levels between them (average identity of 96%) and indicates an old origin of these copies. The only exceptions are HsaCopy23 and 24 (99.5% identity) and the two copies located in chromosome 12 of each species, *FAM90A1* and PtrCopy14, which are likely syntenic (97.9% identity). When phylogenetic trees were built based only on the LTR sequences, we obtained approximately the same associations between the different *FAM90A* copies (data not shown). The main difference in this case was that HsaCopy6 and 20 consistently group with members of Subfamily I, suggesting that there has been gene conversion between the ends of Clusters A and D and Subfamily I sequences.

The rate of synonymous (Ks) and nonsynonymous (Ka) substitutions in the coding sequence of all *FAM90A* genes was estimated with the maximum likelihood analysis program PAML (40). In general, the Ka/Ks ratio for this family was considerably high, with an average of 0.91 over the whole tree, very similar to the expected value of 1 assuming no functional constraint on the coding sequence (neutral evolution). However, there was some variation in Ka/Ks ratios between different groups of *FAM90A* genes (Figure 5), although these differences were not statistically significant. Finally, we divided the coding sequence of the gene in three parts of equal length and calculated the Ka/Ks separately for each of those. Significant differences in the evolutionary rate of the different parts of the protein were detected ($2\Delta l = 201.33$, df =128, $P < 0.0001$), with the C-terminal region (average Ka/Ks = 0.77) showing lower Ka/Ks ratios than the N-terminal (average Ka/Ks = 0.99) and central (average Ka/Ks = 0.92) regions.

**Discussion**

Duplication is known to play a central role in genome evolution (1, 41). With the description of the novel gene family *FAM90A,* composed in humans of at least 25 different members, we present an extreme example of gene expansion by duplication and generation of new gene conformations with different upstream and UTR regions by rearrangement and intra and/or interchromosomal duplication events. In addition, we have observed considerable variation in the organization and the number of copies of these genes in the human population, consistent with copy number variants previously reported for this region (42). Variability in the copy number of *FAM90A* provides further information on the complexity of the intricate genomic architecture of the SDs encompassing the 8p23.1 region (29, 33). The mechanism underlying the generation of the complex structure of the SDs of this region is far from clear. However, the analysis of the available genome sequence information and the experimental results presented here allow us to make several inferences about the evolution of the 8p23.1 genomic region and the *FAM90A* gene family in primates.

A model for the generation of the known copies of *FAM90A* genes in human and chimpanzees genomes from the ancestral sequences present in the common ancestor of hominoids and Old World monkeys could be postulated (Figure 6). Based on the genome sequence assembly and FISH results, Subfamily I members are absent in rhesus macaque and thus we propose that Subfamily II preceded Subfamily I genes. Members of both subfamilies in humans and chimpanzees would have originated from *FAM90A* sequences similar to those found in macaque chromosome 8 which include all *FAM90A* exons except exon 1. Next, there was an insertion of a LTR belonging to the ERV-K family of primate endogenous retroviruses, with the generation of the target site duplications of 6 bp typical of the elements of this type. From there, the whole genomic region probably duplicated at least once and the two *FAM90A* subfamilies diverged. For

Subfamily I copies, some time around the divergence of orangutan and great apes (~15 Ma ago), there was a deletion that fused the Subfamily II *FAM90A* sequences with one of the SDs from the *ALG1* gene region (originally located as a single copy in the homologue of human chromosome 16) that was present in chromosome 8. This resulted in co-opting part of the *ALG1* intron as the upstream sequences and exon 1 of the new *FAM90A* members, forming the initial 1036-bp segment exclusive of this subfamily. Then, the whole region went through a series of several independent duplication and gene loss events, which gave rise to the different Subfamily I copies, found nowadays in the human and chimpanzee genomes.

The precursor of the Subfamily II clusters could have been originated by a tandem duplication of the 7.6 kb module, either by a mechanism similar to slippage during DNA replication or non-allelic homologous recombination (NAHR) between the repeated regions at the beginning and end of each module (Figure 6). This was followed by a deletion of unknown size that eliminated part of the cluster and resulted in a ~500 bp shorter LTR at one end. Then, a process of duplication of the original cluster, deletion of the other border and flanking region in one of the clusters, plus additional duplication of the region spanning the two clusters could have generated the actual distribution of the REPP *FAM90A* clusters in humans, with clusters A and B being roughly a mirror image of clusters C and D. These series of events are consistent with the presence of identical flanking sequences to all clusters on one side, whereas on the other side the flanking sequences of clusters A and D differ from those of B and C. FISH, Southern and quantitative PCR analysis indicate that the duplication of the clusters and the expansion of *FAM90A* occurred most likely in the common ancestor of humans, chimpanzees and gorillas.

The origin of the *FAM90A* gene family in primates involved a complex process of duplications and rearrangements. After that, the evolution of the different *FAM90A* members has probably been characterized by events of NAHR, which could result in the variation of the copy number of these genes observed between individuals, and other structural changes, such as the polymorphic inversion affecting the 8p23.1 region (22). In addition, according to the phylogenetic analysis, recombination between *FAM90A* copies likely produced a great degree of sequence homogenization by gene conversion at two levels. First, there was gene conversion between the copies located in the same cluster that tend to share high sequence identity (Figure 5). Second, there are evidences of gene conversion between copies located at the ends of clusters with the same flanking sequences, such as HsaCopy6 (Cluster A) and HsaCopy20 (Cluster D), HsaCopy12 and HsaCopy19 (Cluster C), or HsaCopy10-11 (Cluster B) and HsaCopy20-21 (Cluster D). In all these cases, the flanking sequences of the corresponding clusters showed almost complete sequence identity, like the flank 3 sequences of Clusters A and C and Clusters B and D.

Comparative genomic analysis revealed that the *FAM90A* family is exclusive of primates and that the most closely-related sequences are two hypothetical genes from cow (*LOC615167*) and dog (*LOC609215*), although these sequences don't share the minimum percentage of identity needed to consoder two genes as homologs. There are several described examples of novel primate genes that have been created by fusion processes, like the chimeric genes derived from the melanocortin-concentrating hormone (43), which are specific of the hominoid lineage. In addition, several gene expansion processes similar to that of *FAM90A* genes have also been reported across primate species (44, 45), including the kruppel-associated box zinc finger gene clusters (46), the neuroblastoma breakpoint gene family (*NBPF*) (44, 45), the *morpheus* genes

on human chromosome 16 (4), or the extreme amplification of the sequences encoding the DUF1220 protein domain in humans (47). Therefore, all these examples stress the importance of SDs and structural changes in the generation of genomic variation and new gene sequences during evolution (48).

In most cases the function of the genes expanded in the primate genome is not yet clear. In *FAM90A* proteins the only known feature is a 19 amino acid motif corresponding to a CCHC zinc-finger domain that could be involved in DNA or RNA binding. The degree of conservation of *FAM90A* protein sequences is compatible with an overall low functional constraint acting on *FAM90A* proteins, as inferred from the Ka/Ks analysis. At least one *FAM90A* member per species has Ka/Ks values clearly lower than 1, indicating that they could encode functional proteins. These include *FAM90A1* (Ka/Ks = 0.35) and HsaCopy10-11 and HsaCopy21-22 in humans (average Ka/Ks = 0.63), the available Subfamily II clusters in chimpanzees (average Ka/Ks = 0.48), and the MmuCopy1 in rhesus (Ka/Ks = 0.78). However, there are also several copies that have accumulated inactivating mutations in the coding sequence through a process of pseudogenization and have Ka/Ks ratios close to 1, such as ΨHsaCopy3, 7, 9, 23 and 24. In addition, there might be other copies with Ka/Ks ratios close to 1 that have lost the ability of being transcribed. Therefore, together with the existence of CNVs and gene conversion events in this region, it is very difficult to define the number of functional *FAM90A* genes in each individual.

Besides the possible changes in the *FAM90A* coding sequence, in this work we describe the generation of a novel gene conformation in the lineage of African great apes, Subfamily I, which has the same coding capacity, but a different first 5'UTR exon and upstream sequences. Interestingly, Subfamily I members, and in particular *FAM90A1*, are the *FAM90A* genes for which there is the largest number of ESTs to

support mRNA expression in humans. Our RT-PCR analysis has showed that both Subfamily I and Subfamily II, are ubiquitously expressed in diverse human tissues. It is tempting to speculate that the acquisition of the new regulatory sequences resulted in a different expression profiles that has been favored by natural selection. Preliminary analyses suggest that previously existing sequences in the intron of the *ALG1* gene were co-opted for the new function and no evidences of acceleration of nucleotide changes in the 5'UTR and upstream regions of Subfamily I members have been found (M.C., unpublished results). However, more exhaustive expression studies and regulation analysis in Subfamily I and Subfamily II genes are needed to assess the role of SDs and structural changes as generators of regulatory diversity during evolution.

## Material and methods

*Sequence analysis*

*In-silico* analysis of the 8p23.1 region sequence was performed on the March 2006 UCSC assembly (Version hg18 – NCBI build 36, http://genome.ucsc.edu/). The analyzed region was divided into four fragments: a) from 6.5 Mb to 7.4 Mb before the 8p23.1 distal gap; b) from 7.5 Mb to 8 Mb after the distal gap; c) from 11.8 Mb to 12 Mb before the proximal 8p23.1 gap; d) from 12.2 Mb to 12.5 Mb after the proximal gap. Repetitive elements in the four different fragments were masked using RepeatMasker (http://www.repeatmasker.org) and the remaining sequences were aligned with PipMaker (38) in order to identify duplicated segments between them. BLAST-based algorithms (49), Celera database, tools available at the Biology Workbench(50) and NCBI's Entrez Gene (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene) were used to obtain information on

the *FAM90A* gene. Functional domains in the FAM90A protein were identified with InterPro Scan software (http://www.ebi.ac.uk/InterProScan).

*Pulse-Field Gel Electrophoresis (PFGE) and Southern blotting*

High-quality genomic DNA was isolated in agarose plugs prepared from lymphoblastoid cell lines from blood samples of different human donors plus one chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*) and orangutan (*Pongo pygmaeus*) individual. DNA plugs were treated with *Acc*65I restriction enzyme, which cuts outside the *FAM90A* clusters but has no target site within them. The digestions were then electrophoresed by PFGE using a CHEF MAPPER system (Biorad) at 6V/cm for 19 h on a 1% agarose gel and blotted onto positively charged nylon membrane (Hybond-N+, Amersham). The filter was pre-hybridized at 42 ºC for four hours on 20X SSC and the hybridization was performed over night at 42 ºC using a 700 bp fragment from the second intron from Subfamily I as a probe. The probe was labeled with the PCR DIG probe Synthesis kit (Roche) and the detection was done with Anti-Digoxigenin-AP and CDP-STAR reagent (Roche).

*RT-PCR expression analysis*

Total RNA was extracted from lymphoblastoid cell lines of human individuals according to a standard protocol using Trizol reagent (Invitrogen). Isolated RNA (5µg) was incubated with DNAse I (Ambion) for 30 minutes at 37ºC and DNAse inactivation reagent was added afterwards. The DNase I-treated total RNA isolated from control individuals as well as commercial total adult RNA from ovary, liver, spleen, lung, placenta, kidney, thymus, heart, skeletal muscle, testes, and colon (Stratagene) and brain

(Ambion) were used for RT-PCR gene-expression analysis. cDNA was synthesized from 1 μg of total RNA by reverse transcription using the SuperScript First Strand Synthesis System (Invitrogen). PCRs were performed in a 12.5 μl reaction volume with 2 μl of cDNA using standard cycling program conditions. Primers to amplify *FAM90A* members were designed by Primer3 (51) and are listed in Supplementary Table 3. Information on ESTs matching particular *FAM90A* copies was obtained from the UCSC genome browser and several IMAGE clones corresponding to cDNAs of *FAM90A* members were directly sequenced to confirm their identity.

*Comparative FISH study*

Metaphase spreads were obtained from human and primate cell lines (lymphoblasts or fibroblasts), including common chimpanzee, gorilla, orangutan and rhesus monkey (*Macaca mulatta*). EBV transformed human lymphoblasts were grown in standard RPMI media containing 10% fetal calf serum and antibiotics. DNA extraction from BACs was done as reported previously (52). FISH experiments were performed essentially as described by Lichter *et al*. (53). Digital images were obtained using a Leica DMRXA2 epifluorescence microscope equipped with a cooled CCD camera (Princeton Instruments, Princenton, NJ, USA). Cy3-dCTP, FluorX-dCTP, DEAC, Cy5-dCTP and DAPI fluorescence signals, detected with specific filters, were recorded separately as gray scale images. Pseudocoloring and merging of images were performed using Adobe Photoshop™ software.

*Real-time PCR analysis*

For the quantitative real-time PCR amplification, two sets of universal probe library (UPL) probes and primer pairs were used, one targeting the last exon of *FAM90A* genes

and the other a single-copy ultra conserved region on human chromosome 6p12.31. The probes and primer sets were designed at the ProbeFinder Design assay center website (https://www.roche-applied-science.com/sis/rtpcr/upl/adc.jsp) and the selected primers targeted regions identical in the great majority of *FAM90A* sequences (Supplementary Table 3). Real-time PCR was performed in the LightCycler® 480 instrument (Roche Molecular Diagnostics), using the following program conditions for both amplicons: 10 minutes of pre-incubation at 95 °C followed by 45 cycles of 15 seconds at 95 ºC, 1 minute at 59 ºC and 30 seconds at 72 ºC. Individual reactions were carried out in triplicate per each sample in 10 µl volumes in a 384 multiwell plate (Roche Molecular Diagnostics) following manufacturer's instructions. Independent genomic DNA-based standard curves were used to determine the efficiencies of the *FAM90A* target amplification in each species. Estimates of *FAM90A* copies quantification were obtained in the form of crossing point (Cp) values based on the 'second derivative maximum' method as computed by the LightCycler 480 (Roche Molecular Diagnostics). Further data analysis was performed with the Cp raw data as described by Pfaffl (54).

*Evolutionary analyses*

To identify homologous sequences to *FAM90A* genes in other species, similarity searches against available genome assemblies and non-redundant databases were performed using BLAT and BLAST. In addition, BLASTP and TBLASTN searches with the *FAM90A* protein were also performed in the NCBI website (http://www.ncbi.nlm.nih.gov/blast). Multiple sequence alignments were carried out with the MUSCLE program with default parameters (55). Phylogenetic trees of the common sequence to all *FAM90A* copies were obtained by neighbor-joining from 1,000 bootstrap replicates using the PHYLIP software package (56). Similar results were

obtained using the UPGMA, DNA parsimony and maximum likelihood methods. Distances of the different branches were calculated using the BASEML module of the PAML program (40). Synonymous (Ks) and nonsynonymous (Ka) substitution rates along different branches were calculated by maximum likelihood under the codon substitution model implemented in PAML(40). In this analysis, codons including alignment gaps or stop signals were removed from all the sequences. To compare the Ka/Ks ratios of different parts of the tree, a likelihood ratio test was performed as previously described (57).

**Acknowledgements**

We are grateful to Susana de la Luna, Baldo Oliva and Arcadi Navarro for helpful discussions. We thank the "Centre de Transfusió i Banc de Teixits de l'Hospital Vall d'Hebrón" (Barcelona, Spain) for the material from blood donors. Financial support was received from Genome Spain, Genome Canada, Spanish Ministry of Science and Education (SAF 2002-00799), the Spanish Ministry of Health (G=/184) and CIBERESP (Instituto de Salud Carlos III). Nina Bosch is a recipient of a BEFI fellowship from "Instituto de Salud Carlos III FIS-ISCIII". Mario Cáceres was supported by the "Ramón y Cajal" Program (Spanish Ministry of Science and Education).

*Conflict of Interest statement. None declared.*

**References**


1. Ohno, S. (1969) The spontaneous mutation rate revisited and the possible principle of polymorphism generating more polymorphism. *Can J Genet Cytol,* **11,** 457-67.
2. Holland, P.W. and Takahashi, T. (2005) The evolution of homeobox genes: Implications for the study of brain development. *Brain Res Bull,* **66,** 484-90.
3. Shen, S.H., Slightom, J.L. and Smithies, O. (1981) A history of the human fetal globin gene duplication. *Cell,* **26,** 191-203.
4. Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M. and Eichler, E.E. (2001) Positive selection of a gene family during the emergence of humans and African apes. *Nature,* **413,** 514-9.
5. Babushok, D.V., Ostertag, E.M. and Kazazian, H.H., Jr. (2006) Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cell Mol Life Sci*.
6. Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet,* **14,** 417-22.
7. Bayes, M., Magano, L.F., Rivera, N., Flores, R. and Perez Jurado, L.A. (2003) Mutational mechanisms of Williams-Beuren syndrome deletions. *Am J Hum Genet,* **73,** 131-51.
8. Emanuel, B.S. and Shaikh, T.H. (2001) Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat Rev Genet,* **2,** 791-800.
9. Lupski, J.R. and Stankiewicz, P. (2005) Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet,* **1,** e49.
10. Conrad, B. and Antonarakis, S.E. (2007) Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. *Annu Rev Genomics Hum Genet*.
11. Tuzun, E., Bailey, J.A. and Eichler, E.E. (2004) Recent segmental duplications in the working draft assembly of the brown Norway rat. *Genome Res,* **14,** 493-506.
12. Armengol, L., Pujana, M.A., Cheung, J., Scherer, S.W. and Estivill, X. (2003) Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum Mol Genet,* **12,** 2201-8.
13. Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L. *et al.* (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science,* **309,** 613-7.
14. Bondeson, M.L., Dahl, N., Malmgren, H., Kleijer, W.J., Tonnesen, T., Carlberg, B.M. and Pettersson, U. (1995) Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet,* **4,** 615-21.
15. Lakich, D., Kazazian, H.H., Jr., Antonarakis, S.E. and Gitschier, J. (1993) Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet,* **5,** 236-41.
16. Osborne, L.R., Li, M., Pober, B., Chitayat, D., Bodurtha, J., Mandel, A., Costa, T., Grebe, T., Cox, S., Tsui, L.C. *et al.* (2001) A 1.5 million-base pair inversion polymorphism in families with Williams-Beuren syndrome. *Nat Genet,* **29,** 321-5.

17. Gimelli, G., Pujana, M.A., Patricelli, M.G., Russo, S., Giardino, D., Larizza, L., Cheung, J., Armengol, L., Schinzel, A., Estivill, X. *et al.* (2003) Genomic inversions of human chromosome 15q11-q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum Mol Genet,* **12,** 849-58.
18. Giglio, S., Broman, K.W., Matsumoto, N., Calvari, V., Gimelli, G., Neumann, T., Ohashi, H., Voullaire, L., Larizza, D., Giorda, R. *et al.* (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am J Hum Genet,* **68,** 874-83.
19. Sugawara, H., Harada, N., Ida, T., Ishida, T., Ledbetter, D.H., Yoshiura, K., Ohta, T., Kishino, T., Niikawa, N. and Matsumoto, N. (2003) Complex low-copy repeats associated with a common polymorphic inversion at human chromosome 8p23. *Genomics,* **82,** 238-44.
20. Giorda, R., Ciccone, R., Gimelli, G., Pramparo, T., Beri, S., Bonaglia, M.C., Giglio, S., Genuardi, M., Argente, J., Rocchi, M. *et al.* (2007) Two classes of low-copy repeats comediate a new recurrent rearrangement consisting of duplication at 8p23.1 and triplication at 8p23.2. *Hum Mutat*.
21. Barber, J.C., Maloney, V., Hollox, E.J., Stuke-Sontheimer, A., du Bois, G., Daumiller, E., Klein-Vogler, U., Dufke, A., Armour, J.A. and Liehr, T. (2005) Duplications and copy number variants of 8p23.1 are cytogenetically indistinguishable but distinct at the molecular level. *Eur J Hum Genet*.
22. Giglio, S., Calvari, V., Gregato, G., Gimelli, G., Camanini, S., Giorda, R., Ragusa, A., Guerneri, S., Selicorni, A., Stumm, M. *et al.* (2002) Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am J Hum Genet,* **71,** 276-85.
23. Giglio, S., Graw, S.L., Gimelli, G., Pirola, B., Varone, P., Voullaire, L., Lerzo, F., Rossi, E., Dellavecchia, C., Bonaglia, M.C. *et al.* (2000) Deletion of a 5-cM region at chromosome 8p23 is associated with a spectrum of congenital heart defects. *Circulation,* **102,** 432-7.
24. Floridia, G., Piantanida, M., Minelli, A., Dellavecchia, C., Bonaglia, C., Rossi, E., Gimelli, G., Croci, G., Franchi, F., Gilgenkrantz, S. *et al.* (1996) The same molecular mechanism at the maternal meiosis I produces mono- and dicentric 8p duplications. *Am J Hum Genet,* **58,** 785-96.
25. Barber, J.C. (2005) Directly transmitted unbalanced chromosome abnormalities and euchromatic variants. *J Med Genet,* **42,** 609-29.
26. Tsai, C.H., Graw, S.L. and McGavran, L. (2002) 8p23 duplication reconsidered: is it a true euchromatic variant with no clinical manifestation? *J Med Genet,* **39,** 769-74.
27. Harada, N., Takano, J., Kondoh, T., Ohashi, H., Hasegawa, T., Sugawara, H., Ida, T., Yoshiura, K., Ohta, T., Kishino, T. *et al.* (2002) Duplication of 8p23.2: a benign cytogenetic variant? *Am J Med Genet,* **111,** 285-8.
28. Engelen, J.J., Moog, U., Evers, J.L., Dassen, H., Albrechts, J.C. and Hamers, A.J. (2000) Duplication of chromosome region 8p23.1-->p23.3: a benign variant? *Am J Med Genet,* **91,** 18-21.
29. Taudien, S., Galgoczy, P., Huse, K., Reichwald, K., Schilhabel, M., Szafranski, K., Shimizu, A., Asakawa, S., Frankish, A., Loncarevic, I.F. *et al.* (2004) Polymorphic segmental duplications at 8p23.1 challenge the determination of individual defensin gene repertoires and the assembly of a contiguous human reference sequence. *BMC Genomics,* **5,** 92.

30. Nusbaum, C., Mikkelsen, T.S., Zody, M.C., Asakawa, S., Taudien, S., Garber, M., Kodira, C.D., Schueler, M.G., Shimizu, A., Whittaker, C.A. *et al.* (2006) DNA sequence and analysis of human chromosome 8. *Nature,* **439,** 331-5.
31. Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.C. and Scherer, S.W. (2003) Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol,* **4,** R25.
32. Aldred, P.M., Hollox, E.J. and Armour, J.A. (2005) Copy number polymorphism and expression level variation of the human {alpha}-defensin genes DEFA1 and DEFA3. *Hum Mol Genet*.
33. Hollox, E.J., Armour, J.A. and Barber, J.C. (2003) Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *Am J Hum Genet,* **73,** 591-600.
34. Linzmeier, R.M. and Ganz, T. (2005) Human defensin gene copy number polymorphisms: Comprehensive analysis of independent variation in alpha- and beta-defensin regions at 8p22-p23. *Genomics*.
35. Ballana, E., Gonzalez, J.R., Bosch, N. and Estivill, X. (2007) Inter-population variability of DEFA3 gene absence: correlation with haplotype structure and population variability. *BMC Genomics,* **8,** 14.
36. Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W. and Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science,* **297,** 1003-7.
37. Glusman, G., Yanai, I., Rubin, I. and Lancet, D. (2001) The complete human olfactory subgenome. *Genome Res,* **11,** 685-702.
38. Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res,* **10,** 577-86.
39. Wu, Q. and Krainer, A.R. (1999) AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Mol Cell Biol,* **19,** 3225-36.
40. Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci,* **13,** 555-6.
41. Samonte, R.V. and Eichler, E.E. (2002) Segmental duplications and the evolution of the primate genome. *Nat Rev Genet,* **3,** 65-72.
42. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W. *et al.* (2006) Global variation in copy number in the human genome. *Nature,* **444,** 444-54.
43. Courseaux, A. and Nahon, J.L. (2001) Birth of two chimeric genes in the Hominidae lineage. *Science,* **291,** 1293-7.
44. Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. and van Roy, F. (2005) A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol Biol Evol,* **22,** 2265-74.
45. Fortna, A., Kim, Y., MacLaren, E., Marshall, K., Hahn, G., Meltesen, L., Brenton, M., Hink, R., Burgers, S., Hernandez-Boussard, T. *et al.* (2004) Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biol,* **2,** E207.
46. Eichler, E.E., Hoffman, S.M., Adamson, A.A., Gordon, L.A., McCready, P., Lamerdin, J.E. and Mohrenweiser, H.W. (1998) Complex beta-satellite repeat structures and the expansion of the zinc finger gene cluster in 19p12. *Genome Res,* **8,** 791-808.

47. Popesco, M.C., Maclaren, E.J., Hopkins, J., Dumas, L., Cox, M., Meltesen, L., McGavran, L., Wyckoff, G.J. and Sikela, J.M. (2006) Human lineage-specific amplification, selection, and neuronal expression of DUF1220 domains. *Science,* **313,** 1304-7.
48. Bailey, J.A. and Eichler, E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet,* **7,** 552-64.
49. McGinnis, S. and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res,* **32,** W20-5.
50. Subramaniam, S. (1998) The Biology Workbench--a seamless database and analysis environment for the biologist. *Proteins,* **32,** 1-2.
51. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol,* **132,** 365-86.
52. Ventura, M., Archidiacono, N. and Rocchi, M. (2001) Centromere emergence in evolution. *Genome Res,* **11,** 595-9.
53. Lichter, P., Tang, C.J., Call, K., Hermanson, G., Evans, G.A., Housman, D. and Ward, D.C. (1990) High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science,* **247,** 64-9.
54. Pfaffl, M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res,* **29,** e45.
55. Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* **5,** 113.
56. Nilsson, R.H., Larsson, K.H. and Ursing, B.M. (2004) galaxie--CGI scripts for sequence identification through automated phylogenetic analysis. *Bioinformatics,* **20,** 1447-52.
57. Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol,* **15,** 568-73.
58. Wienberg, J., Jauch, A., Stanyon, R. and Cremer, T. (1990) Molecular cytotaxonomy of primates by chromosomal in situ suppression hybridization. *Genomics,* **8,** 347-50.

**Figure Legends**

**Figure 1.** Schematic representation of the 8p23.1 region. *(A)* Ideogram of human chromosome 8 showing magnification of the 8p23.1 region. *(B)* Wide colored arrows indicate the orientation of the *FAM90A* clusters (named A, B, C and D) and other single *FAM90A* copies. Distance between clusters D and HsaCopy23 is 3.4 Mb and coincides with the region where the polymorphic inversion affecting 8p23.1 occurs (18). *(C)* Diagram of *FAM90A* clusters, represented with the same colors as in *B*, and their positions with regard to alpha and beta-defensin gene clusters and olfactory receptors (OR). Thin arrows show the directions of transcription and the numbers underneath are the sizes of the clusters in kb according to hg18. Besides *FAM90A* clusters, REPD is composed of an alpha-defensin cluster (DEFA6, DEFA4, DEFA3, DEFT1, DEFA3, DEFT1, DEFA3 and DEFA5), two copies of OR7E125P and OR7E154P pseudogenes, two copies of a beta-defensin cluster (DEFB4, DEFB103A, DEFB104, DEFB106, DEFB105, and DEFB107), an estimated 100 kb gap, and olfactory receptor OR7E96P. REPP contains three different olfactory receptor pseudogenes OR7E158P, OR7E161P and OR7E160P, followed by HsaCopy23 of *FAM90A* family, an estimated 100 kb gap, and finally *FAM90A* HsaCopy24.

**Figure 2.** Exon-intron structure for *FAM90A* Subfamily I and Subfamily II genes. Filled boxes correspond to non-translated exons and open boxes to the coding sequence. The alternatively spliced exon is represented with a dotted line. Repetitive elements are symbolized as gray and black rectangles. Nonsense substitutions corresponding to HsaCopy23 and HsaCopy24 from Subfamily I, and HsaCopy3 from Subfamily II are represented by asterisks. Frameshifts mutations are depicted as downward triangles for the 1 bp insertions on HsaCopy3 and HsaCopy9 or as upward triangles for the 1 bp deletions on HsaCopy7. The black arrow indicates the alternative GC donor splice site

present in nine members of Subfamily II. 5' UTR and 3' UTR are represented as black lines and the coding sequence (CDS) as a white rectangle below the diagrams.

**Figure 3.** Southern blot hybridization of pulse-field gel electrophoresis (PFGE) of *Acc*65I digested genomic DNA from 20 individuals with a probe corresponding to the second intron of *FAM90A1*. Size of marker fragments is indicated on both sides. Black bars between the panels correspond to the approximate location of the 7 expected digestion fragments based on the reference assembly (cluster C, 77 kb; cluster A, 68 kb; cluster B, 54 kb; *FAM90A1* fragment, 49 kb; cluster D, 45 kb; and HsaCopy23 and HsaCopy24, two 13.6 kb fragments).

**Figure 4.** RT-PCR analysis of the expression pattern of *FAM90A* gene family. (*A*) RT-PCR of *FAM90A* members from mRNA of 12 different human tissues. (*B*) RT-PCR of *FAM90A* members from mRNA of lymphoblastoid cell lines from 7 individuals of the general population. (*C*) RT-PCR with primers specific to *FAM90A* members from Subfamily I (*FAM90A1* on chromosome 12 or HsaCopy23/24 on chromosome 8) and for multiple members of Subfamily II on 13 human tissues. 1-kb molecular size ladder is shown on the left side. Wells on the right correspond to genomic DNA.

**Figure 5.** Evolutionary analysis of *FAM90A* copies in primates. Phylogenetic tree was obtained by Neighbor-joining using the sequence common to all the available full-length *FAM90A* members (5081 bp). Bootstrap values of 1000 replicates are indicated in boldface for the main nodes and only branches with more than 70% bootstrap support are represented in the tree. Branch lengths were calculated with the BASEML module of the PAML program (40) and correspond to the number of nucleotide substitutions per

position. Ka/Ks ratios obtained using the free-ratio model of the PAML CODEML module (40) are indicated in italics above the main branches, with dashes representing branches with Ks = 0. Brackets indicate the average Ka/Ks values of copies forming different clades within the tree. Gene-inactivating mutations are represented as asterisks.

**Figure 6.** Schematic diagram of the most parsimonious evolutionary model for the origin of *FAM90A* genes in humans and chimpanzees. Main steps of the process of generation of the two *FAM90A* subfamilies are shown, with the approximate time period when they occurred indicated on the left using as reference the origin of different phylogenetic groups in the human and chimpanzee lineage. The different sequences included in *FAM90A* copies are represented as orange and red rectangles and coding and non-coding exons are represented in black. Sequences homologous to L1MB3 and LTR5A elements are shown as dashed grey boxes. Purple, green, yellow and blue rectangles symbolize the flanking sequences and the ones found in current *FAM90A* copies are numbered. Arrows in top of the diagrams correspond to the Subfamily I and II modules shown in Figure 2. Part of the intronic sequence of the *ALG1* gene included in Subfamily members is also shown. The ancestral ape chromosomal nomenclature of roman numerals is used to indicate the chromosomal location. The *FAM90A* copies currently present in the human and chimpanzee reference genomes are included within rectangles. The Subfamily II modules in Clusters A-D drawn are just approximate and do not represent the actual number of copies. Organization of the Subfamily II clusters in chimpanzees is not well resolved due to problems with the genome sequence assembly. See text for details.

Supplementary **Table 1**. Nomenclature of *FAM90A* members in the human genome reference sequence.

| FAM90A members[a] | HUGO name[b] | Location[c] | Coordinates (hg18) | Subfamily | Exons | ORF length[d] |
|---|---|---|---|---|---|---|
| HsaCopy1 | *LOC645353* | REPD (Cluster A) | 7100096-7105408 | II | 5 | 464 |
| HsaCopy2 | *FAM90A3* | REPD (Cluster A) | 7107719-7113030 | II | 5 | 464 |
| HsaCopy3 | *FAM90A4* | REPD (Cluster A) | 7115341-7120652 | II | 5 | 345 |
| HsaCopy4 | *LOC441314* | REPD (Cluster A) | 7122963-7128274 | II | 5 | 464 |
| HsaCopy5 | *FAM90A5* | REPD (Cluster A) | 7130585-7135896 | II | 5 | 464 |
| HsaCopy6 | *LOC645392* | REPD (Cluster A) | 7138207-7143516 | II | 5 | 464 |
| HsaCopy7 | *FAM90A6P* | REPD (Cluster B) | 7393425-7398741 | II | 5 | 233 |
| HsaCopy8 | *FAM90A7* | REPD (Cluster B) | 7401070-7406391 | II | 5 | 464 |
| HsaCopy9 | *FAM90A2P* | REPD (Cluster B) | 7408720-7414037 | II | 5 | 345 |
| HsaCopy10 | *LOC645558* | REPD (Cluster B) | 7416365-7421687 | II | 5 | 464 |
| HsaCopy11 | *LOC645572* | REPD (Cluster B) | 7424014-7429331 | II | 5 | 464 |
| HsaCopy12 | *LOC645651* | REPD (Cluster B) | 7608636-7613947 | II | 5 | 464 |
| HsaCopy13 | *LOC645709* | REPD (Cluster C) | 7616275-7621595 | II | 5 | 464 |
| HsaCopy14 | *LOC441323* | REPD (Cluster C) | 7623923-7629243 | II | 5 | 464 |
| HsaCopy15 | *FAM90A8* | REPD (Cluster C) | 7631571-7636890 | II | 5 | 464 |
| HsaCopy16 | *LOC441325* | REPD (Cluster C) | 7639218-7644538 | II | 5 | 464 |
| HsaCopy17 | *LOC441326* | REPD (Cluster C) | 7646866-7652186 | II | 5 | 464 |
| HsaCopy18 | *FAM90A9* | REPD (Cluster C) | 7654514-7659834 | II | 5 | 464 |
| HsaCopy19 | *FAM90A10* | REPD (Cluster C) | 7661166-7667482 | II | 5 | 464 |
| HsaCopy20 | *FAM90A11P* | REPD (Cluster D) | 7906714-7912030 | II | 5 | 464 |
| HsaCopy21 | *FAM90A12* | REPD (Cluster D) | 7914358-7919676 | II | 5 | 464 |
| HsaCopy22 | *FAM90A12* | REPD (Cluster D) | 7922006-7927323 | II | 5 | 464 |
| HsaCopy23 | *FAM90A2P* | REPP | 12067117-12073497 | I | 6 | 266 |
| HsaCopy24 | *LOC389633* | REPP | 12316398-12322780 | I | 6 | 266 |
| *FAM90A1* | *FAM90A1* | Chr12p13.31 | 8265123-8271464 | I | 6 | 464 |

[a]Working nomenclature used for this study (distal to proximal criteria).

[b]Name given by the Human Genome Organization.

[c]Segmental duplications on 8p23.1, REPD refers to the distal set and REPP to the proximal duplicons.

[d]Length of the open reading frame expressed in number of amino acid residues.

**Table 2.** Summary of FISH experiments performed on chromosome spreads from primate species with four HSA clones and a PCR fragment from *FAM90A* gene.

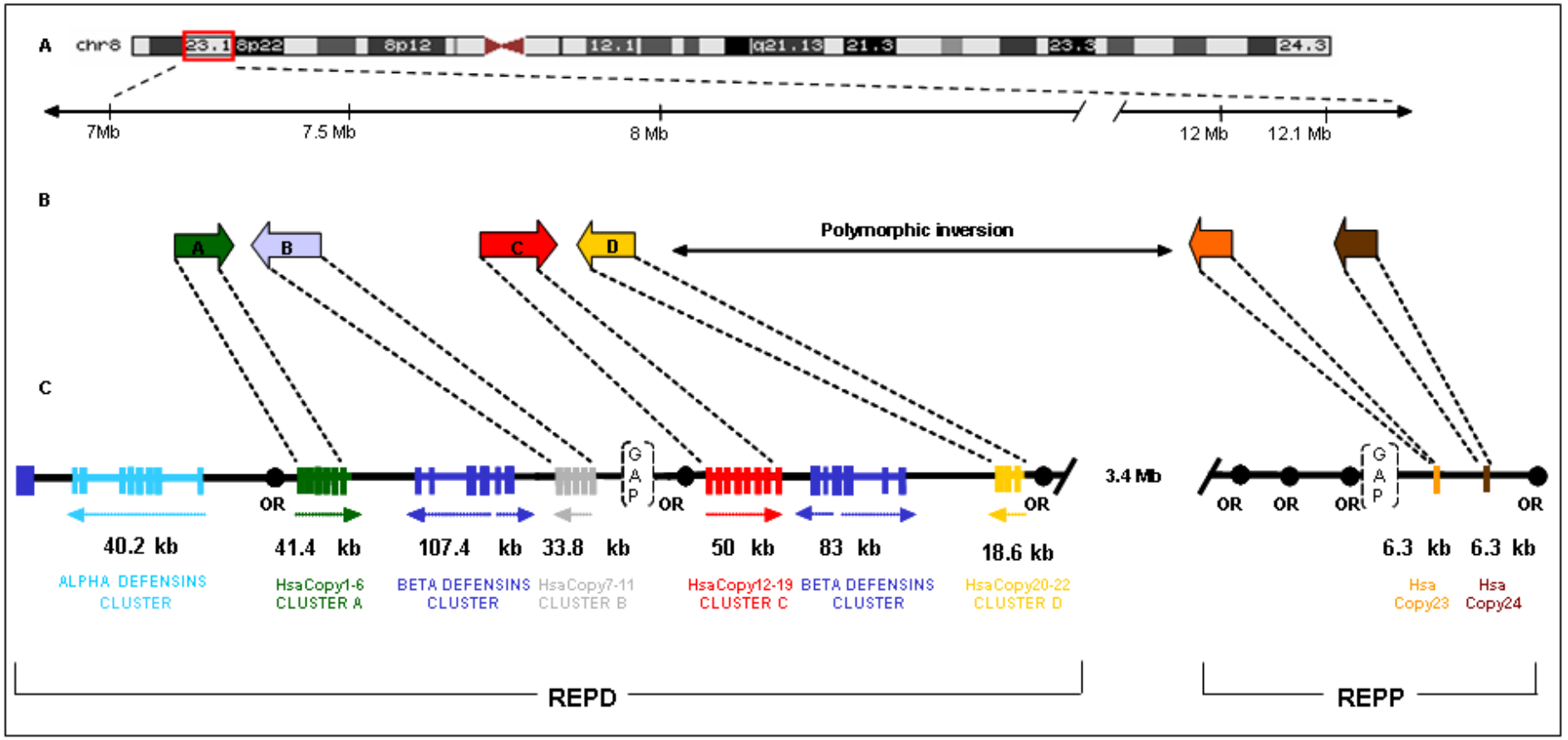
| FISH probes | Human | Chimpanzee | Gorilla | Orangutan | Rhesus |
|---|---|---|---|---|---|
| RP11-585J10 | III - IV - VII | III - IV - VII | III - IV - VIII | VIII | MMU2 (VII/XXI) MMU8 (VIII) |
| | VIII - XI - XII - XVI | VIII - XI - XII - XVI | XI - XII - XVI | XVI | MMU20 (XVI) |
| RP11-351I21 | III - IV - VII | III - IV - VII | III - IV - VII | IV | MMU4 (IV) |
| | VIII - XI - XII - XVI | VIII - XI - XII - XVI | VIII - XI - XII - XVI | VIII - XVI | MMU8 (VIII) MMU16 (XIII) |
| RP11-115E11 RP11-1067L18 | VIII * | VIII * | VIII | VIII | MMU2 (VII/XXI) MMU8 (VIII) |
| *FAM90A* GENE | VIII * | VIII * | VIII | VIII | – |

*Symbol indicates the presence of multiple signals on chromosome 8.

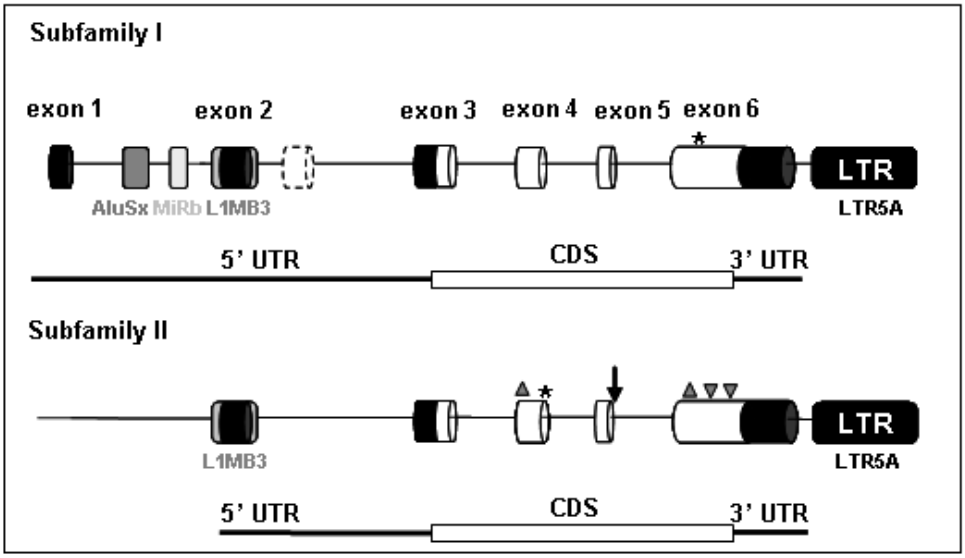
Ape chromosomes are indicated by the phylogenetic nomenclature (roman numerals) with the exception of rhesus macaque where nomenclature reported by Wienberg et al. (58) is used and their homologous chromosomes in apes are shown in parenthesis.

**Supplementary Table 2. Human BAC clones containing *FAM90A* members.**

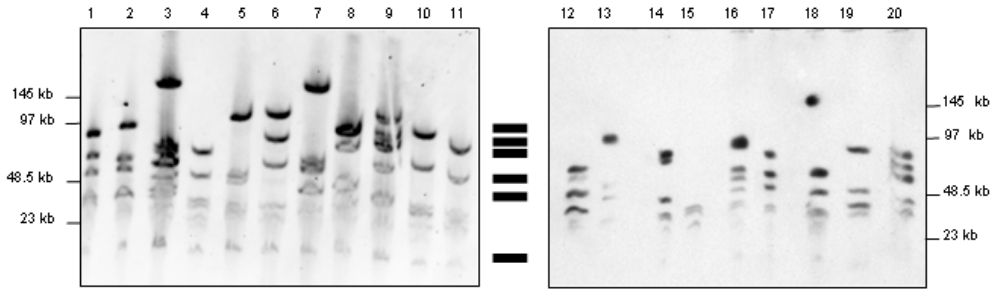
| *GeneBank* | BAC clone | *FAM90A* Copy number |
|---|---|---|
| AC193090 | WI2-1590L12 | 1 |
| AC148106 | RP11-65P161 | 13 |
| AC144950 | RP11-191L23 | 6 |



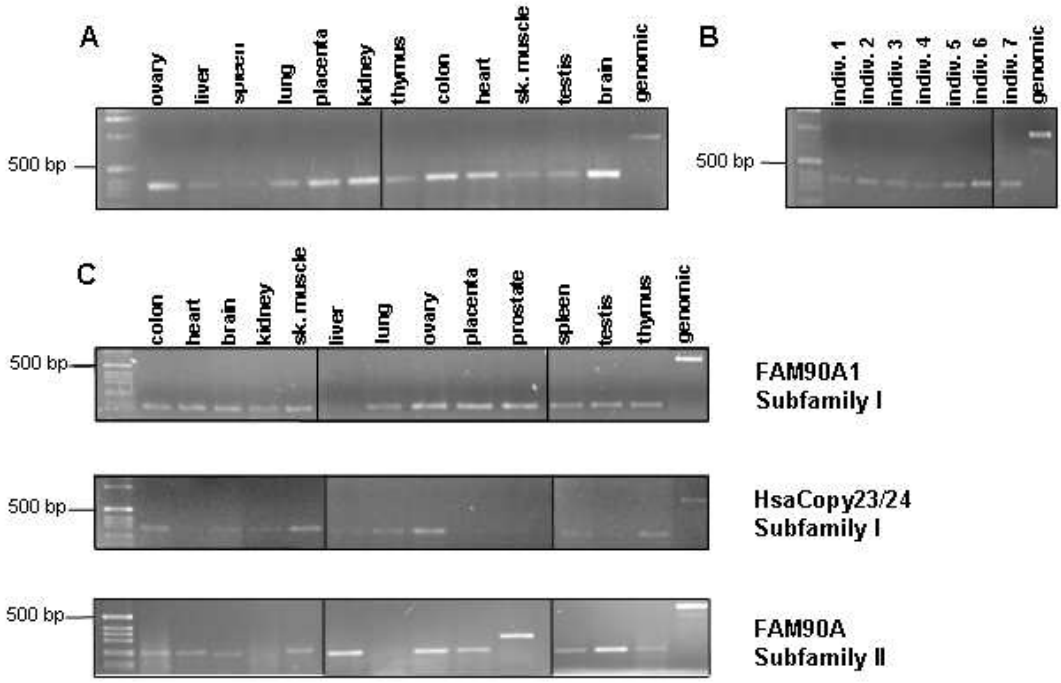

**BoschN Figure 1**



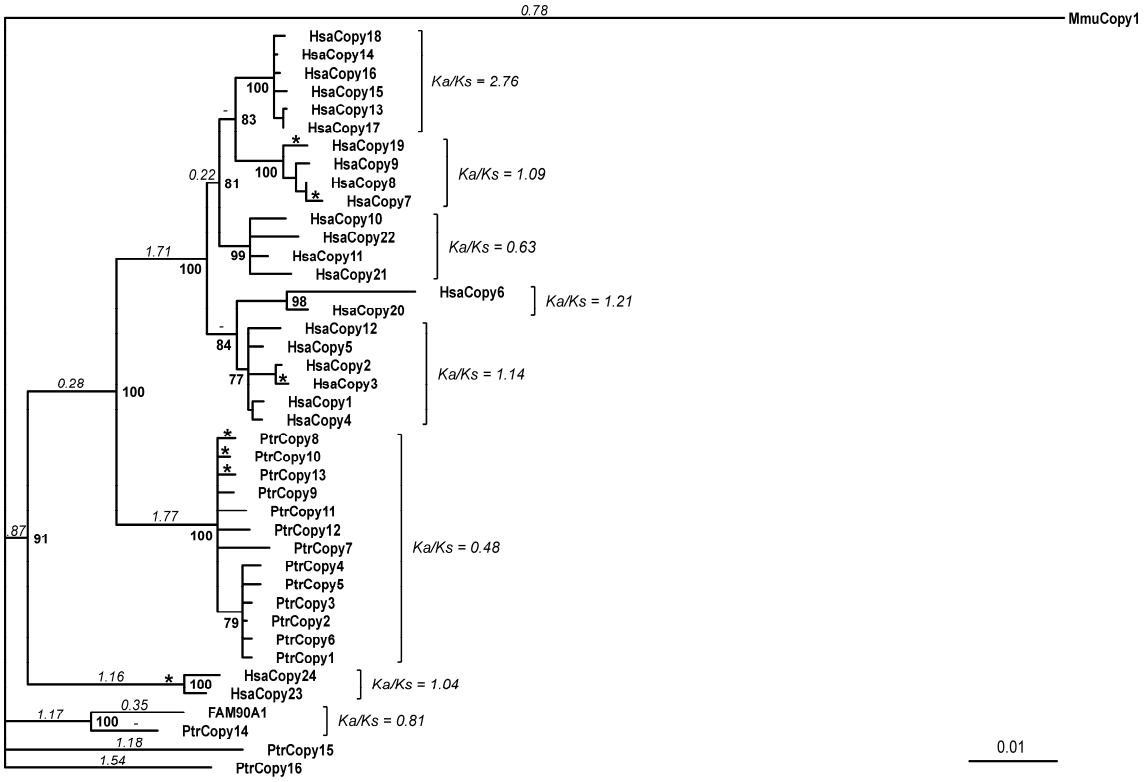

**BoschN Figure 2**



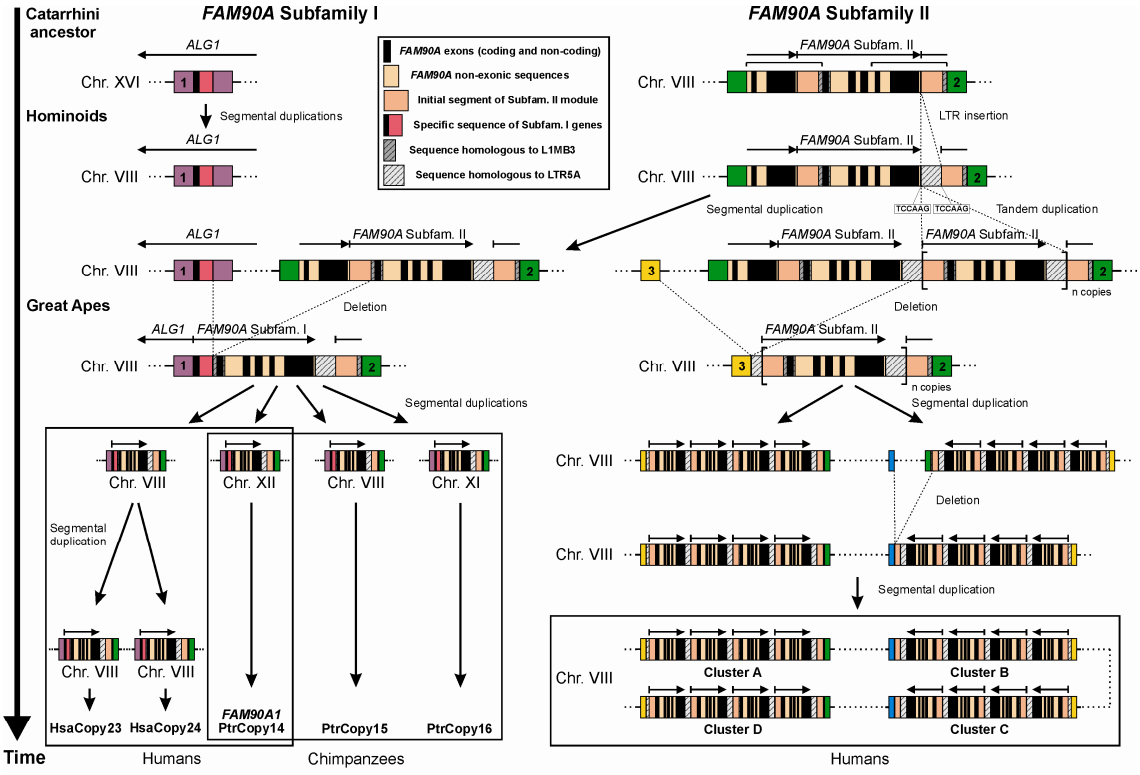

**BoschN Figure 3**




**BoschN Figure 4**




BoschN Figure 5




BoschN Figure 6

**Abbreviations**

**SDs:** Segmental duplications

**REPD:** distal duplicons on 8p23.1

**REPP:** proximal duplicons on 8p23.1

**FAM90A:** family with sequence similarity 90

**CNV:** copy number variant

**NAHR:** non-allelic homologous recombination

**OR:** olfactory receptor

**LINE**: long interspersed element

**LTR:** long terminal repeat

**ERVK:** endogenous retrovirus family K

**ALG1:** beta-1, 4-manosyltransferase

**OWM:** Old World monkeys

**Ka:** synonymous substitutions rate

**Ks:** nonsynonymous substitutions rate

Figure 2

The only exceptions are HsaCopy3, 7, 9, and 23-24, which have a total of 6 potential gene-inactivating mutations resulting probably in non-functional proteins (Figure 2). These mutations are two nonsense mutations in HsaCopy3 (ORF position 307) and in HsaCopy23-24 (ORF position 800), and four frameshift mutations in HsaCopy7 (1 bp insertion in ORF position 698 and 1 bp deletions in ORF positions 915 and 921) and HsaCopy9 (1 bp insertion in ORF position 290). In addition, there are 10 *FAM90A* Subfamily II copies (HsaCopy1-5, 10-11, and 20-22), which present a T to C mutation in the donor splice site of exon 5 that changes the common GT to the alternative donor splice site GC (Figure 2), although that should not affect the encoded protein (39).