



Immunoglobulin genes in Primates

David N. Olivieri^{a,*}, Francisco Gambón Deza^b

^a School of Computer Science, University of Vigo, Ourense 32004, Spain

^b Immunology Unit, Hospital of Meixoiro, Vigo, Spain

ARTICLE INFO

Keywords:

Immunoglobulin genes

Machine learning

CH domains

Primate IG evolution

ABSTRACT

Five classes of immunoglobulins are known to exist in mammals. The number of isotypes of classes G, E and A varies among species for unknown reasons. Here, a study of the presence of immunoglobulin genes in Primates was carried out from the genomes and transcriptomes deposited in the NCBI repository. For this, a machine learning application based upon neural networks was implemented that scans the genomes and identifies the exon sequences that encode the immunoglobulin CH domains. From these exons, the immunoglobulins that each species possess can be inferred. Also, the presence of sequences outside the IGHC locus was found which were produced by retrotranscription of RNA that are probably not viable. From this study, the distribution of immunoglobulin genes across primate orders is described in detail. In Prosimians, IgD genes are not found; in Platyrrhines, a gene is identified for each of the immunoglobulin classes but the IgD gene does not have the CH2 exon; in the Cercopithecidae family, a gene is detected for each class in the Colobinae family, while in Cercopithecidae the genes for IgG have been duplicated several times. In hominids, a greater number of duplications that include the genes that code for IgA and IgE are observed. These results indicate that from the appearance of the Cercopithecidae, there is an evolutionary instability in the Ig locus.

1. Introduction

In mammals, five classes of immunoglobulins are known to exist: immunoglobulins M, D, G, E and A (IgM, IgD, IgG, IgE and IgA) (Murphy and Weaver, 2016). Mammalian IgM is similar to that found in the first jawed vertebrates; namely, it contains four exons for each of the four CH domains of its protein sequence. Unlike what happened in reptiles, this gene in mammals has remained unique and no viable duplications have been observed in any mammalian species. The gene has been extensively studied since it generates the antibodies of the primary immune response (Ohta and Flajnik, 2006).

Immunoglobulin D (IgD) is a mysterious antibody in mammals. At present, it is still unknown why this gene has been maintained in mammals since its expression is slight compared to the corresponding eleven domain gene of reptiles. In humans, the gene has three CH domains that correspond to the CH1–CH7–CH8 domains of reptiles (Gambón-Deza and Espinel, 2008). Some mammalian species have lost this gene entirely and others, such as the murids (Muridae), have only two CH (Preud'homme et al., 2000) domains.

Immunoglobulin E (IgE) and G (IgG) evolved from the duplication of an immunoglobulin called Y (IgY) that is currently found in birds and reptiles. IgY has 4 CH domains that were inherited from mammalian

IgE, whereas IgG has only three domains (having lost the CH2 domain from IgY) (Gambon-Deza et al., 2009). This latter antibody is the most abundant in mammals and is the most effective for responding to infections. IgE is an antibody that is detected in very low concentrations, typically involved in the response to parasites, but is also responsible for allergic responses (Gould and Sutton, 2008).

Mammalian immunoglobulin A (IgA) is an antibody with three CH domains. It is the principal secreted antibody and very effective against mucosal infections. Its origin however is not completely understood. IgAs are detected in amphibians in which two lineages have been described. One of the genes is located in the interior of the immunoglobulin heavy constant (IGHC) locus, in an area similar to that found in the IGHA of birds and crocodiles, and the other gene is located at the end of the locus at the same location found in mammals. These amphibian IgAs have four domains. Recently, studies of amphibian genomes have shown that the IgA located within the interior of the locus must have given rise to the IgAs present in birds and crocodiles, while the IgA located at the end of the locus gave way to the mammalian IgA with a loss of the CH2 exon (Estevez et al., 2016; Deza et al., 2007).

The IGHC locus has been studied intensively in humans and mice. In humans it is composed of seven functional genes (i.e., the genes for IgM,

* Corresponding author.

E-mail address: davidolivieri@yahoo.es (D.N. Olivieri).

IgD, IgG3, IgG1, IgA1, IgG2, IgG4, IgE and IgA2) and two pseudogenes (pseudogene-IgE and the pseudogene-IgG, IGHGP) (Rabbani et al., 1996). In this work, we studied the presence of these genes in representative species across the entire Primate order in order to better understand the IG gene evolution and configuration in humans.

In recent years, genome sequences of many primate species have become publicly available in the form of assembled WGS (Whole Genome Shotgun sequencing) datasets. These assemblies consist of relatively large contigs containing most of the genomic sequences. Despite the existence of gaps and the limitation of the segmented sequences, a general view of the IGHC locus can be obtained in each primate family. For this, we developed a software analysis program, we call *CHfinder*, that identifies the sequences of the exons that code for each of the CH domains of all the immunoglobulin isotypes.

2. Materials and methods

2.1. Datasets

The WGS assembly datasets of 30 primate species were obtained from the NCBI in the form of FASTA files consisting of assembled contigs, or for more mature projects, scaffolds and/or fully constructed chromosomes. The average genome coverage in these datasets is $> 15\text{--}20\times$ with contig assembly N50 > 15 kbs. A detailed summary of the accession numbers and relevant assembly parameters can be found in the Supplementary Materials (Table 6). The Primates studied in this work are listed in the phylogenetic tree of Fig. 1.

2.2. Software

While there are seven viable immunoglobulin genes in *Homo sapiens*, the number of genes varies across mammal species. Nonetheless, the Ig exon architecture is nearly universal across all mammals. A diagram of the *H. sapiens* IGHC exon structure and corresponding protein domains has been described (Lefranc et al., 2005) (see also sequence repositories of the IMGT). The relevant genomic signals (i.e., the exons, introns, 5'-URT and 3'-URT) of IG genes can be identified to a high degree of accuracy using homology criteria with a supervised machine learning classifier, producing a listing of viable exons (i.e., those exons that could form a functionally expressed Ig molecule). Therefore, to uncover IG exons from WGS genomes of non-human

primates, we developed an iterative gene discovery algorithm implemented in Python, called *CHfinder*.

In particular, the *CHfinder* program extracts exons for CH domains. Although other exons constitute the full Ig gene, these exons are of particular interest for characterizing and comparing Ig genes within and amongst species. Thus, *CHfinder* ignores the peptide leader (L) as well as all peptides corresponding to the transmembrane and cytoplasmic tail. While the exon/intron structure of Ig is thought to be universal across jawed vertebrates, the specific intron spacing between the exon sequences varies considerably. As such, the algorithm of *CHfinder* only imposes a simple structural requirement that the CH domains are found in a tandem arrangement along the DNA sequence, but places no hard restrictions on the intron separation.

Fig. 2 summarizes the principal steps of the *CHfinder* algorithm. This program was implemented as a multi-threaded application in the Python programming language, with the biopython library (Cock et al., 2009) for low-level sequence analysis, and the Tensorflow library (Abadi et al., 2016) for machine learning tasks (using deep neural networks). First, a *tblastn* query (Tatusova and Madden, 1999) from a consensus protein sequences from known immunoglobulins (Igs) is made against all available primate WGS datasets. The search result is a listing of candidate WGS contigs likely to contain exons, together with the position of the matching nucleotide sequence and similarity scores; this listing is referred to as a *hit table*. The algorithm processes each line of the *hit table*, analyzing a nucleotide region larger than the nucleotide positions in the hit, so as to determine the precise start/stop positions of the exon reading-frame (defined by AG and GT motifs, respectively). Once the possible exons are identified, they are translated into amino acid sequences by checking all valid reading frames. Those sequences between 260 and 360 nucleotides, divisible by 3 and not containing stop-codons in the reading frame, are saved and converted into numerical feature vectors (i.e., a unique array of numbers, that uniquely characterizes the string of amino acids). A simple transformation of AA to feature vector was used (based upon the frequency of pairs of AA), because it was found to perform better at discriminating sequences when compared to other more sophisticated transformation procedures (e.g., those based upon positional physicochemical properties of each AA within the sequences).

Next, supervised machine learning was used to classify the exon sequences into one of 17 different exon types that constitute the IG genes (4 for IgM, 3 for IgD, 3 for IgG, 4 for IgE and 3 for IgA). For this,

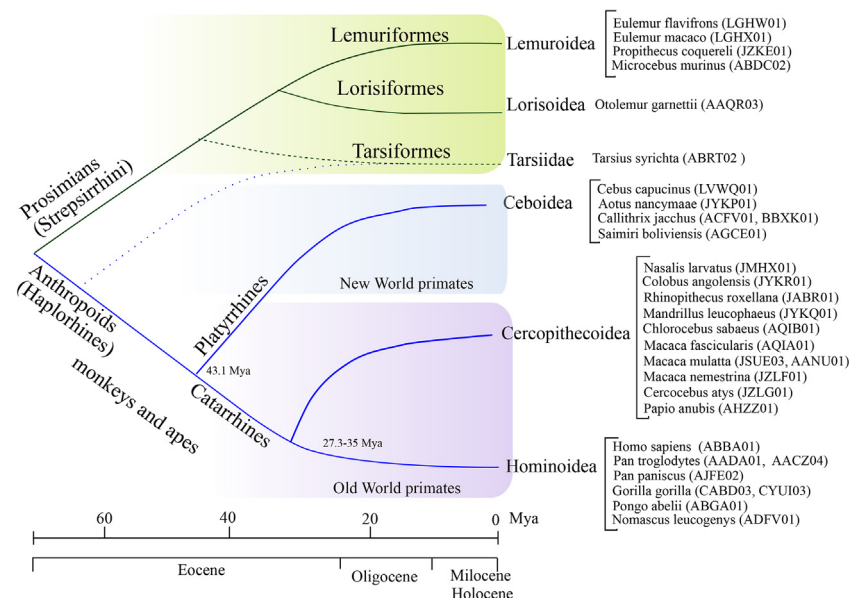


Fig. 1. The phylogenetic tree of the Primates species studied in this work. The tree is based upon divergence times obtained from Hedges et al. (2006), and previous molecular phylogenetic studies (Perelman et al., 2011; Rogers and Gibbs, 2014).

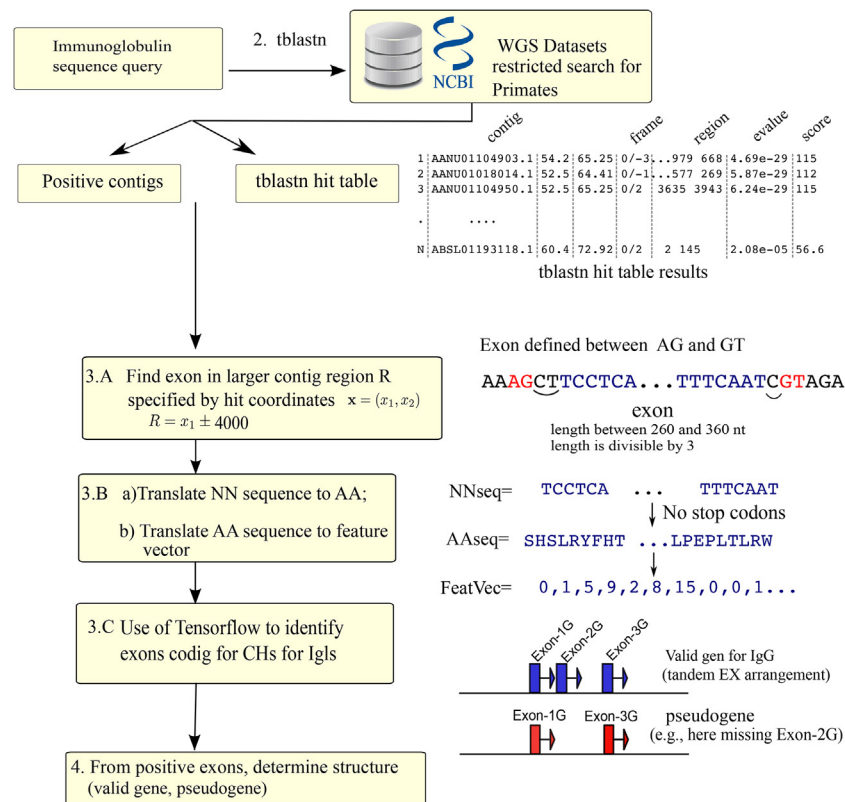


Fig. 2. The steps in the IG exon prediction algorithm. The selection of valid CH exons is based upon a *tblastn* pre-selection, an exon reading-frame identification procedure, and identification/classification with a supervised machine learning method using the Tensorflow library.

the sequences that have been converted into numerical feature vectors are analyzed with a deep neural network (DNN) model using the Tensorflow library (Abadi et al., 2016). The DNN model for all the classes is trained from a known set of annotated exons and specifying a large random genomic background (typically $\approx 3 - 7 \times$ the number of labeled genes).

In our gene finding algorithm, *CHfinder*, a probable functionally expressed Ig gene is one that must contain a tandem arrangement of the three or four viable exons (i.e. having no stop codons in the reading frame) along the germline sequence. Nonetheless, viable exons (i.e., homologous to Ig, and constituent exons) are found throughout the IG genome regions that are not arranged in a tandem configuration (i.e., they are isolated or an exon is missing); as such, these lone exons do not express Ig molecules.

Tree construction. To study the phylogenetic relationships from the immunoglobulin heavy constant (IGHC) exons, we constructed a large phylogenetic tree by aligning sequences with ClustalO (Sievers and Higgins, 2014) and then used phyML (Guindon et al., 2010) with the WAG matrix (part of the Fasttree software, Price et al., 2010). In all cases, 500 bootstrapped samples were made. For producing the graphics, Figtree (Rambaut, 2007) was used.

3. Results

The complete sequencing of the genomes of most primate species is still unfinished, but partially assembled genomes are provided as Whole Genome Shotgun sequencing (WGS) data files (see the Supplementary Table 6 for contig statistics of the WGS files analyzed). Despite the presence of gaps and possible gene segmentations in some areas (i.e., a gene can be split into two different contigs) the IGHG locus could be identified for each of the primate families. For this, an online NCBI *tblastn* query was made using Ig sequences against WGS data from Primates. The query result is a listing of homologous segments,

indicating the contig and location of homologous sequences. As described, from this listing *CHfinder* extracts candidate exons that are analyzed by the DNN to determine those that correspond to the CH domains of immunoglobulins.

3.1. Prosimians

In six Prosimian genomes, we found 70 exons that code for the CH regions of IgM, IgG, IgE and IgA. The results are shown in Table 1. The few exons that are found for *D. madagascariensis* can be explained by low sequencing coverage. In the other species, exons are detected for an IgA and an IgE in each species. For IgG, the results show that most species have at least one gene, while *P. coquereli* and *T. syrichta* possess two genes. The exons obtained for IgM demonstrate the existence of the gene in the Tarsiiformes lineage (*T. syrichta*) and Lorisiformes (*O. garretti*). In the differentiation line of Lemuriformes, the complete gene is detected in *M. murinus* and *E. flavifrons*. However, in the other two species no complete gene is detected; only the first two CHs are found in *E. macaco*, while no exon for CH is found in *P. coquereli*. We studied the region where *CHfinder* predicts the location of exons for CH1 and CH2 of IgM in *E. macaque*. These sequences are found, but at the locations where the CH3 and CH4 exons should be present; instead this region is occupied by short repeated sequences (microsatellites) and the start of the CH3 exon is interrupted by these sequences.

To clarify whether some lemur species have lost the IgM gene, the presence of RNAs for immunoglobulins in RNAseq files was studied. Table 2 shows the antibodies detected from the RNAseq files of a primate liver transcriptome study (Brawand et al., 2011) available from the public NCBI repository. The results show that IgM sequences are detected in some transcriptomes. For example, in the transcriptome SRR361336 from *P. coquereli*, an IgM sequence was found in RNAseq even though for this species the gene for IgM in the WGS genome was not detected. This suggests, for this particular case, that the gene was

Table 1

The number of exons that code for immunoglobulin CH domains in Prosimians.

	IgM					IgD				IgG				IgE				IgA				
	CH1	CH2	CH3	CH4	IgM ^a	CH1	CH2	CH3	IgD ^a	CH1	CH2	CH3	IgG ^a	CH1	CH2	CH3	CH4	IgE ^a	CH1	CH2	CH3	IgA ^a
<i>Daubentonia madagascariensis</i> - AGTM01	–	–	1	1	0	–	–	–	0	1	–	1	0	–	–	–	–	0	–	–	1	0
<i>Eulemur flavifrons</i> -LGHW01	1	1	1	1	0	–	–	–	0	1	1	1	1	1	1	1	1	1	1	1	1	1
<i>Eulemur macaco</i> -LGHX01	1	1	–	–	0	–	–	–	0	1	1	1	1	1	1	1	1	1	–	1	1	0
<i>Microcebus murinus</i> -ABDC03	1	1	1	1	1	–	–	–	0	1	1	1	1	1	1	1	–	0	1	1	1	1
<i>Otolemur garnettii</i> -AAQR03	1	1	1	1	0	–	–	–	0	1	1	1	1	1	1	1	1	1	1	1	1	1
<i>Propithecus coquereli</i> -JZKE01	–	–	–	–	0	–	–	–	0	2	2	2	2	1	1	1	1	1	1	1	1	1
<i>Tarsius syrichta</i> -ABRT02	1	1	1	1	1	–	–	–	0	–	2	2	0	2	1	1	1	1	1	1	1	1

^a Number of molecules deduced by CHfinder when finding the corresponding consecutive exons in the same contig.**Table 2**

The presence of the immunoglobulin classes in transcriptomes of Prosimians.

SRR file	Specie	IgM	IgD	IgG	IgE	IgA
SRR361352	<i>E. coronatus</i>	–	–	+	–	–
SRR357438	<i>E. coronatus</i>	+	–	+	–	–
SRR361343	<i>E. mongoz</i>	–	–	+	–	+
SRR361350	<i>P. coquereli</i>	–	–	+	–	+
SRR361336	<i>P. coquereli</i>	+	–	+	–	+
SRR361343	<i>O. garnettii</i>	+	–	+	–	+

not detected either because of a problem with the genome sequencing or assembly of the WGS.

Another general characteristic common in all prosimians is the absence of a gene for IgD. With the CHfinder program, we did not find any coding exons for any of the three IgD mammalian CHs. A schematic representation of the locus for *M. murinus* is shown in Fig. 3. The specific search for this gene with tblastn was negative.

3.2. Platyrrhines

From seven WGS genome files of four species of Platyrrhines (from *Callithrix jacchus*, four genomes were studied) 90 CH sequences were obtained (see Table 3). The data indicate that these primates have an isotype for each immunoglobulin class. In some genomes, extra exons are detected for the CH3 domain of IgE. These exons are found outside the locus and are apparently orphans, a phenomenon already described in human Vs and λ genes (Nagaoka et al., 1994). An analysis of the location of these extra exons was carried out. In the case of contig

NTIC01000002.1, a continuous reading frame is detected that generates a CH2–CH3 sequence of IgE, suggesting a process of retrotranscription of RNA of the IgE, as opposed to the duplication of a segment (see Fig. 4). We also studied an extra exon of *C. jacchus* detected in the contig NTIC0102223.1. In this case, there is a reading frame that generates the four CH domains of the constant region of IgE. From computational methods, three exons were detected that are separated by small intron spacings of 30 nucleotides. Once again, these appear to be due to RNA retrotranscription insertion in the DNA. The sequences corresponding to CH2 and CH3 have small deletions. In this same contig, the presence of an MHC class I gene is detected. A sequence was detected in the genome of *A. nancymae* in the contig JYKP02087504.1 that has a similar structure to that in contig NTIC01000002.1.

With the CHfinder program, the CH2 exon of the IgD genes was not detected. To determine the reason (i.e., whether this is because the gene has only two exons or it for some other unknown reason), we studied the contig JYKP02059748.1 of *A. nancymae* in great detail using an independent method. In addition to the IgD gene, this contig also contains genes for IgM, IgG and IgE. It is a large contiguous sequence without significant gaps. Despite an exhaustive search for exons in the region between the IgM and IgG genes, only the two exons detected. As further confirmation, the CH2 of IgD was not detected in the contig LVWQ01093284.1 of *Cebus capucinus* (Fig. 3).

3.3. Cercopithecoidea

The CHfinder program was used to search for Ig domains in species of the Cercopithecoidea suborder. The total number of CH exons obtained for each Cercopithecoidea species with CHfinder are given in

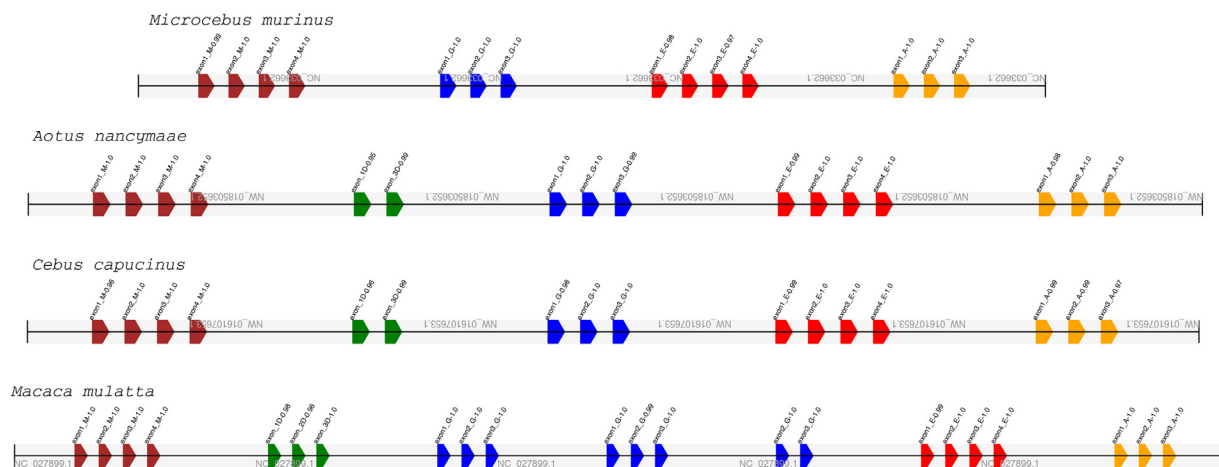


Fig. 3. The schematic representation and location of exons coding for immunoglobulin CHs in Prosimian *M. murinus*, Platyrrhines *A. nancymae* and *C. capucinus* and the Cercopithecoidea *M. mulatta*. Although the diagram was made from a listing of the exon locations and order, the distances were adjusted to fit on the same graph. Also, the true sequences are in the complementary chain (the diagram is rotated for illustration). Exons are colored as follows: IgM (brown), IgD (green), IgG (blue), IgA (orange), and IgE (red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Table 3
The number of exons coding for the immunoglobulin CH domains in Platyrrhines.

	IgM					IgD				IgG				IgE					IgA			
	CH1	CH2	CH3	CH4	IgM ^a	CH1	CH2	CH3	IgD ^a	CH1	CH2	CH3	IgG ^a	CH1	CH2	CH3	CH4	IgE ^a	CH1	CH2	CH3	IgA ^a
<i>Aotus nancymaae-JYKP02</i>	1	1	1	1	1	1	–	1	1 ^b	1	1	–	0	1	1	2	1	1	–	1	1	0
<i>Callithrix jacchus-ACFV01</i>	1	1	1	–	0	–	–	1	0	–	1	1	0	–	1	3	1	0	1	1	1	1
<i>Callithrix jacchus-BBXK01</i>	1	1	1	–	0	–	–	1	0	–	1	1	0	1	1	3	1	0	1	1	1	1
<i>Callithrix jacchus-JRUL01</i>	–	–	–	–	0	–	–	–	0	–	–	1	0	–	–	–	–	0	–	–	–	0
<i>Callithrix jacchus-NTIC01</i>	1	1	1	1	1	2	–	–	0	1	1	2	0	2	1	4	1	1	1	1	1	1
<i>Cebus capucinus-LVWQ01</i>	1	1	1	1	1	1	–	1	1 ^b	1	1	1	1	1	2	1	1	1	1	1	1	1
<i>Saimiri boliviensis-AGCE01</i>	1	1	1	1	1	–	–	1	0	–	1	1	0	1	1	1	1	1	1	1	1	1

^a Number of molecules deduced by CHfinder when finding the corresponding consecutive exons in the same contig.
^b Genes apparently viable but only with the domain CH1 and CH3.

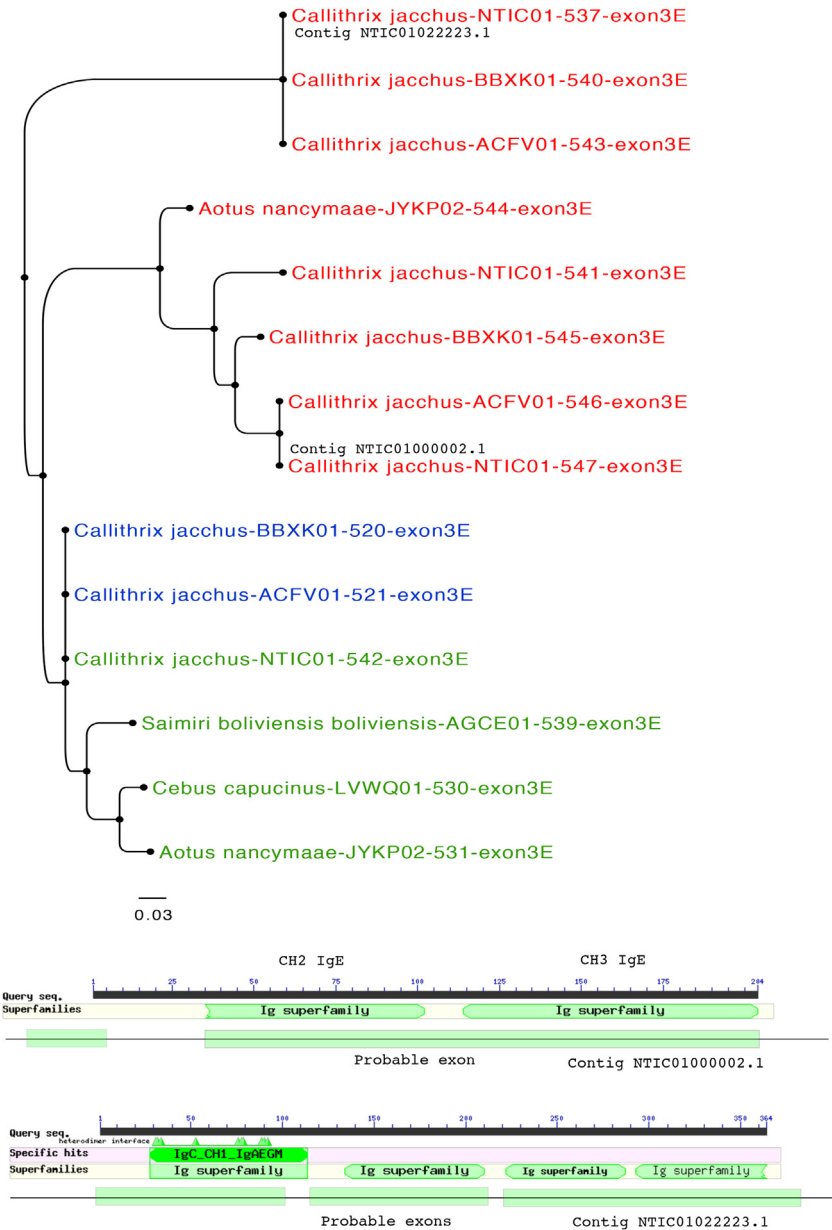


Fig. 4. (Top) Phylogenetic tree made with the exons that encode the CH3 domain of IgE in Platyrrhines. Those colored in green are exons within a complete gene. The exons in blue correspond to incomplete sequences, but are suggestive of belonging to the gene IGHE. The exons colored red are the CH3 of IgE and encoded outside the IGHE. (Bottom) Shown are the sequences that are found in contigs outside the IGHC. These sequences are represented by the IG domains found from a tblastn search with the deduced amino acid sequences of the exons from their respective contigs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

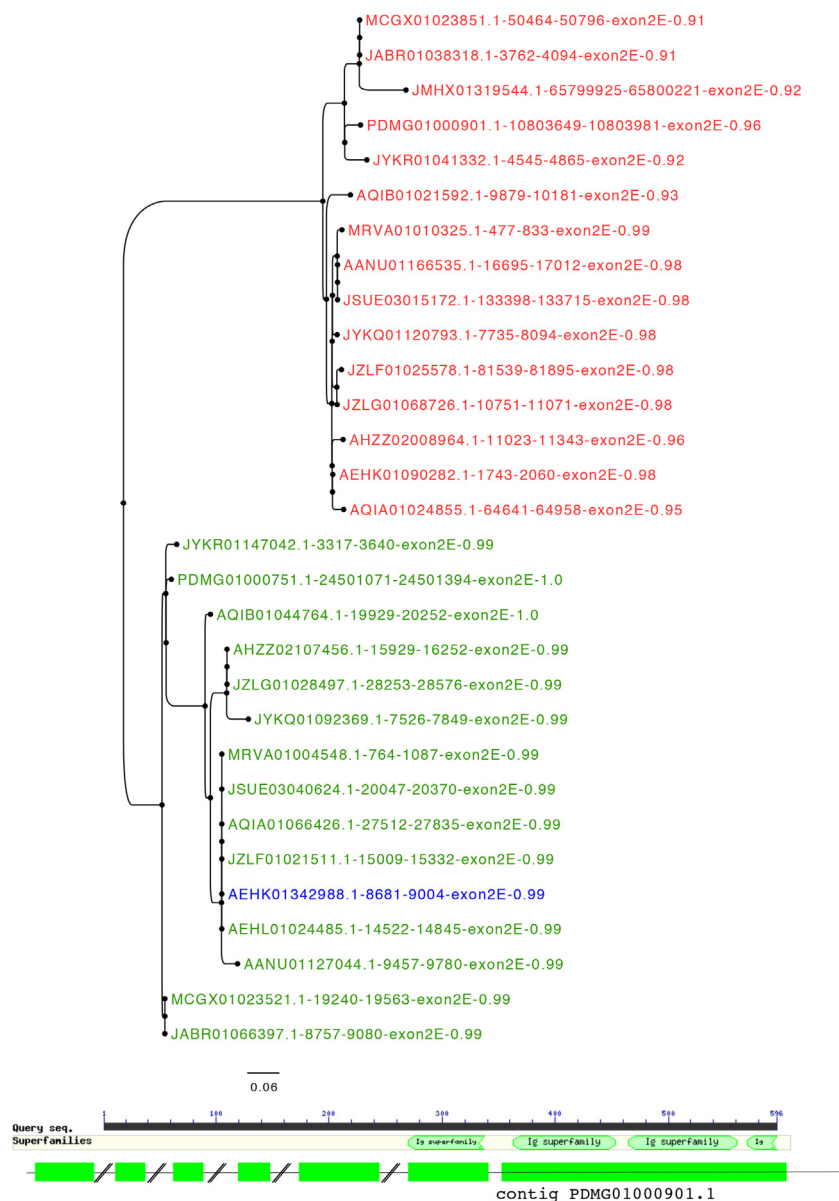
Table 4. These results show evidence of additional IgE CH2 exons and duplications of IgG in some species. In this suborder, IgD having the three mammalian domains are detected in most species (Fig. 5).
We studied the excess of exon-2 of IgE. In all the genomes of the Cercopithecoidea suborder, this exon is duplicated; however, only one

of the copies is located in the IGHE locus, while the other copy is found in remote contigs. The duplication of this exon must have occurred in the founding species and the sequence has been maintained. To confirm these findings, the sequence PDMG01000901.1 of the genome of the Cercopithecidae *P. tephrosceles* was analyzed (Fig. 6). Again, an almost

Table 4

The number of exons that code for immunoglobulin CH domains in Cercopithecoidea.

	IgM					IgD				IgG				IgE				IgA				
	CH1	CH2	CH3	CH4	IgM ^a	CH1	CH2	CH3	IgD ^a	CH1	CH2	CH3	IgG ^a	CH1	CH2	CH3	CH4	IgE ^a	CH1	CH2	CH3	IgA ^a
<i>Cercocebus atys</i> -JZLG01	1	1	1	1	1	1	1	1	1	3	3	3	3	1	2	1	1	1	1	1	1	1
<i>Chlorocebus sabaeus</i> -AQIB01	1	1	1	1	1	1	1	1	1	–	1	1	0	1	2	2	1	1	1	1	1	1
<i>Colobus angolensis</i> -JYKR01	1	1	1	1	1	1	–	1	0	2	1	1	1	2	3	1	1	1	–	–	–	0
<i>Macaca fascicularis</i> -AEHL01	1	1	1	1	1	1	1	1	0	1	–	1	0	1	1	1	1	1	1	1	1	1
<i>Macaca fascicularis</i> -AQIA01	1	1	1	1	1	1	1	1	1	2	2	2	1	1	2	1	1	1	1	1	1	1
<i>Macaca fascicularis</i> -CAEC01	–	–	–	1	0	–	1	–	0	2	1	–	0	–	–	–	–	0	–	–	–	0
<i>Macaca mulatta</i> -AANU01	–	1	1	1	0	–	–	–	0	5	5	5	3	1	2	1	1	1	1	1	1	1
<i>Macaca mulatta</i> -AEHK01	1	1	1	1	1	1	1	1	1	–	1	2	0	1	2	–	1	1	1	1	1	1
<i>Macaca mulatta</i> -JSUE03	1	1	1	1	1	1	1	1	1	3	4	4	2	1	2	1	1	1	1	1	1	1
<i>Macaca nemestrina</i> -JZLF01	1	1	1	1	1	–	–	–	0	4	4	3	3	1	2	1	1	1	1	1	1	1
<i>Mandrillus leucophaeus</i> -JYKQ01	1	1	1	1	1	1	1	1	0	3	3	2	1	1	2	1	1	1	1	1	1	1
<i>Nasalis larvatus</i> -JMXX01	–	–	–	1	0	1	–	1	0	1	1	1	1	2	1	–	–	0	1	1	1	1
<i>Papio anubis</i> -AHZZ02	1	1	1	1	1	1	1	1	1	2	2	2	2	1	2	1	1	1	–	1	1	0
<i>Ptilocolobus tephrosceles</i> -PDMG01	1	1	1	1	1	1	1	1	1	–	1	1	0	2	2	1	1	1	1	1	1	1
<i>Rhinopithecus bieti</i> -MCGX01	1	1	1	1	1	1	1	–	0	1	2	1	1	2	2	1	1	1	1	1	1	1
<i>Rhinopithecus roxellana</i> -JABR01	1	1	1	1	1	1	1	1	0	2	1	1	1	2	2	2	1	1	1	1	1	1

^a Number of molecules deduced by CHfinder when finding the corresponding consecutive exons in the same contig.**Fig. 5.** (Top) Phylogenetic tree made with the exons that encode the CH2 domain of IgE from the Cercopithecoidea family. Exons that form complete genes are colored in green; those exons that do not constitute a complete gene but are suggestive of belonging to the IGHE are colored in blue; those exons identified as IgE CH2 and encoded outside the IGHE are indicated in red. Together with the taxon label, the contig accession and the exon location are indicated; the last number is the sequence prediction probability from CHfinder that CH2 sequence of IgE. (Bottom) Shown are the sequences that are found in contigs outside the IGHC. These sequences are represented by the IG domains found from a tblastn search with the deduced amino acid sequences of the exons from their respective contigs. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

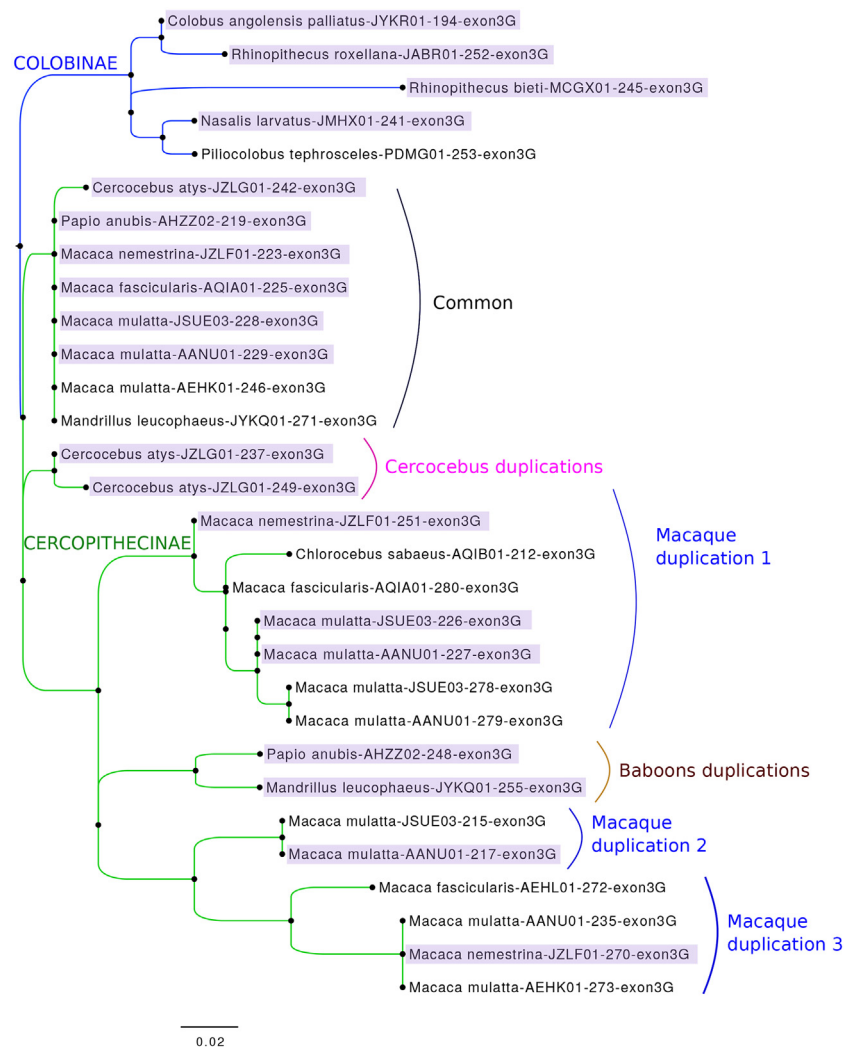


Fig. 6. Phylogenetic tree made with the exons that encode the IgG CH3 domain from the Cercopithecoidea family. The branches in the two main clades (blue) are formed from Colobinae species and Cercopithecoidea species (green). Subclades are indicated where duplication processes occurred. The exons from taxa having probable functional genes determined by CHfinder are indicated (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

complete chain of the constant region of IgE exists with an open reading frame. This occurs in two exons separated by a short intron spacing, indicative of a retrotranscription. The genomic structure of this region is very similar to that described above in Platyrrhines, thereby suggesting that the insertion process of this sequence occurred prior to the diversification of these species with the Cercopithecoidea suborder.

As can be deduced from Table 4, duplications of the IgG gene have occurred. Within the Cercopithecoidea (Old World monkeys), there are two evolutionary families: one that gave rise to the Cercopithecoidea (macaques, mandrills, baboons and cercocebus), and the other to the Colobinae (Columbus, Nasalis, Rhinopithecus and Ptilocolobus). As shown in Table 4 and Fig. 6, these duplications did not occur in the Colobinae family since there is only one gene of IgG per species. Duplications can be seen throughout all species of the Cercopithecoidea. However, such duplications were recent because each species has uniquely duplicated sequences (e.g., the duplicated sequences are different between the *Cercocebus* and *Macaca* species).

Results from the different genomes of *Macaca mulatta* indicate differences in the number of IgGs between them. It should be noted that the sequencing of this area of the genome is complicated. The switch regions are composed of repeated palindromic sequences and span a considerable distance. In addition, the presence of microsatellites is common. This conditions the presence of gaps or contigs not assigned to

any chromosome. To attempt to identify the functional IgGs of *Macaca mulatta*, the presence of messenger RNAs in the SRR1778441 and SRR357438 files of a spleen and liver samples were studied. The reconstruction of transcriptome from the spleen generated sequences of transcripts for one IgM, two IgG and one IgA. No IgD and IgE sequences were detected, probably due to their low expression. From liver transcriptomes, the following Ig classes are detected: one IgM, one IgG (equal a one of the found in spleen) and one IgA. In this genome more exons are detected than those necessary for the probable functional genes, indicating the presence of pseudogenes. In this species, there is a chromosome assembly (Mmul-8.0.1). The scaffold NC-027899.1 (part of chromosome 7) containing the IGHC locus, was studied with CHfinder. The program identifies, in addition to the gene for IgM, IgD, IgA and IgE, two (probable) viable genes for IgG and a pseudogene (Fig. 3).

3.4. Hominids

The results of applying CHfinder to Hominid WGS files are provided in Table 5. Evidence shows that duplications of IgG genes occurred in these species; duplications also occurred in both IgA and IgE genes. Because of the proximity of these Hominid species to humans, the sequences can be studied further by performing alignment against those

Table 5

The number of exons that code for immunoglobulin CH domains in Hominids.

	IgM					IgD				IgG				IgE				IgA				
	CH1	CH2	CH3	CH4	IgM ^a	CH1	CH2	CH3	IgD ^a	CH1	CH2	CH3	IgG ^a	CH1	CH2	CH3	CH4	IgE ^a	CH1	CH2	CH3	IgA ^a
<i>Gorilla_gorilla-CYUI03</i>	1	1	1	1	1	1	1	1	1	5	6	6	2	2	3	2	2	0	1	1	1	1
<i>Gorilla_gorilla-CABD03</i>	1	1	1	1	1	1	1	1	1	4	3	4	2	2	2	2	2	1	1	1	1	1
<i>Nomascus_leucogenys-ADFO1</i>	–	–	1	1	0	1	1	–	0	3	2	2	2	–	1	–	1	0	1	1	2	1
<i>Pan_paniscus-AJFE02</i>	1	1	1	1	1	1	1	1	1	3	6	6	0	1	2	1	2	1	1	1	–	0
<i>Pan_troglodytes-AACZ04</i>	–	–	–	1	0	1	1	1	1	7	5	5	3	1	1	1	2	1	–	1	2	0
<i>Pan_troglodytes-AADA01</i>	1	1	1	1	1	1	1	1	0	3	2	1	1	–	–	–	1	0	–	–	–	0
<i>Pan_troglodytes-NBAG03</i>	2	2	1	2	1	1	1	1	1	7	8	8	7	1	1	1	2	1	2	2	2	2
<i>Pongo_abelii-ABGA01</i>	–	1	1	1	0	2	1	1	0	5	4	4	3	3	3	1	2	1	–	1	1	0
<i>Pongo_abelii-NDHI03</i>	1	1	1	1	1	1	1	1	1	4	4	5	4	4	4	2	3	2	3	3	3	3

^a Number of molecules deduced by CHfinder when finding the corresponding consecutive exons in the same contig.

from humans.

In the WGS genome of *P. troglodytes* (NBAG03), CHfinder detects two exons for the CH1, CH2, and CH4 domains of the IgM. We studied the gene for this immunoglobulin in detail to determine if there is a second IgM gene in this genome. In addition to containing the IgM domains, the contig NBAG03000176.1 also contains exons for IgD and IgG; these results indicate that it is part of the main IGHC locus. The contig NBAG03003939.1, contains three additional exons and was used to compare sequences of the exons of the IgM gene of the main locus. From these studies, exact duplications of the CH1, CH2, and CH4 domains were detected. The exon for the CH3 domain has alterations at the start of its sequence, having lost the AG of exon initiation. Taking this into account, it was not detected in other chimpanzee genomes, suggesting that it is a particular duplication and that an orphan pseudogene was formed.

As mentioned above, duplications of the IgG gene are evident. Fig. 7 shows the tree constructed from the complete IgG sequences in Hominids obtained with the CHfinder and another tree with the exons encoding the CH3 domain of the IgG. There are more CH3 exons than sequences of IgG because either segments of the locus are partially sequenced, or the exons are in different contigs and the software cannot associate these exons to the same Ig coding gene. The CHfinder program also detects exons that form pseudogenes. The trees show that both gorillas and chimpanzees share sequences similar to the four IgGs described in humans. One point of interest is the divergence of the orangutan from the rest of the hominids. In the assembly NDI03 of *P. abelii*, the program CHfinder extracts four complete sequences of IgG. All these are in the same clade as the sequence of human IgG1. This indicates that all orangutan IgGs have a common ancestor with human IgG1. These duplications of the gene in *P. abelii* occurred after the point of divergence of chimpanzees, gorillas, and humans. In addition, evolutionary instability in the locus is evident, since duplications have continued to occur especially in the clade of IgG3 where gorillas and chimpanzees have created an additional IgG. The tree with the coding domains for the CH3 of the IgG confirms the previous results. Additionally, the human pseudogene IGHGP has been added and shown to be a duplication of IgG2 immunoglobulin and is also present in chimpanzees and bonobos, but not in gorillas.

In the genome of *P. abelii* (NDHI03), the CHfinder program found the sequence of two IgEs and three IgAs. The presence of these extra genes is suggestive of duplications in the locus that encompass the gene segment for IgG-IgE-IgA. Such a duplication has already been suggested in humans (Flanagan and Rabbitts, 1982). The instability of the locus also demonstrates the existence of a large number of other Ig homologous exons that are non-functional, and are most likely pseudogenes (Fig. 8).

The locus structure of the immunoglobulins constant regions (IGHC) can be determined with the JGI Genomes OnLine Database (GOLD) (Mukherjee et al., 2017) sequences available at the Refseq-genomes of the NCBI. From this data, we studied the DNA segment of the IGHC

locus in detail. In humans, CHfinder finds 32 exons represented schematically in Fig. 8. These exons make up the following genes: one IgM, one IgD, five IgGs (one of which is a pseudogene, IGHGP), two IgA and two IgEs (one of which is another pseudogene, since only two exons are identified). Other hominids are represented in the same figure.

In chimpanzees, the IgE pseudogene is not detected, however there are exons that make up three additional IgGs (IGHG3 duplications) together with a pseudogene of this same class. The most important modifications are found in gorillas. In this species, the IgE pseudogene persists, isolated exons of IgG (pseudogenes) are detected, internal IgA is not found in the locus, and interestingly there is an IgG gene inserted between the exons for CH1 and CH2 of another IgG gene (Fig. 8). The study of a liver transcriptome suggests the functionality of the unique gene and probably the inserted gene behaves as a pseudogene (see Fig. 9). In the orangutan, four IgGs and one pseudogene are detected. The IgE that in the previous species was a pseudogene here is the viable one and vice versa; the IgE located at the end of the locus is not viable.

4. Discussion

The locus of the constant regions of immunoglobulins has changed significantly in the evolutionary lines of vertebrates. The constant regions of the antibodies carry effector functions that are differentiated across each immunoglobulin class. Hence, detailed knowledge of the evolutionary changes, i.e., either generation or loss, of Ig classes is of great interest in comparative immunology.

Mammals have five classes of immunoglobulins. With the exception of IgD, all species have at least one gene for each of the Ig classes. The gene structure of Ig has been maintained throughout evolution since the origin in mammals more than 250 MYA. The IgD class has been lost in some mammal species, such as the opossum, having substituted it for short repeat sequences of ERV and LINE1 (Wang et al., 2009). In our study, we found that the Prosimians also lack the IgD class, which is particularly relevant since it is a species close to humans in evolution. In this case, neither the CHfinder program search nor tblastn queries against the WGS genome assemblies provide evidence for the existence of the three CH exons of the IgD antibody. Interestingly, in the Tarsiiformes species studied (*T. syrichta*), the absence of IgD exons supports classifying it within the Prosimians.

IgD in humans is an antibody that is found in very low concentrations in serum. Despite this low in serum concentration, IgD is widely expressed as B cell receptors in many species. In amphibians and reptiles, the ancestor gene for these can be found (Estevez et al., 2016; Gambón-Deza and Espinel, 2008). In these animals, they typically have eleven CH domains and differential splicing can produce several forms of this immunoglobulin. In mammals, domains have been lost in evolution, resulting in only three domains (CH1, CH7, and CH8 from the original ancestor) (Gambón-Deza and Espinel, 2008). In humans, the maintenance of this gene throughout evolution suggests some primary function, however, in Prosimians this function is apparently not

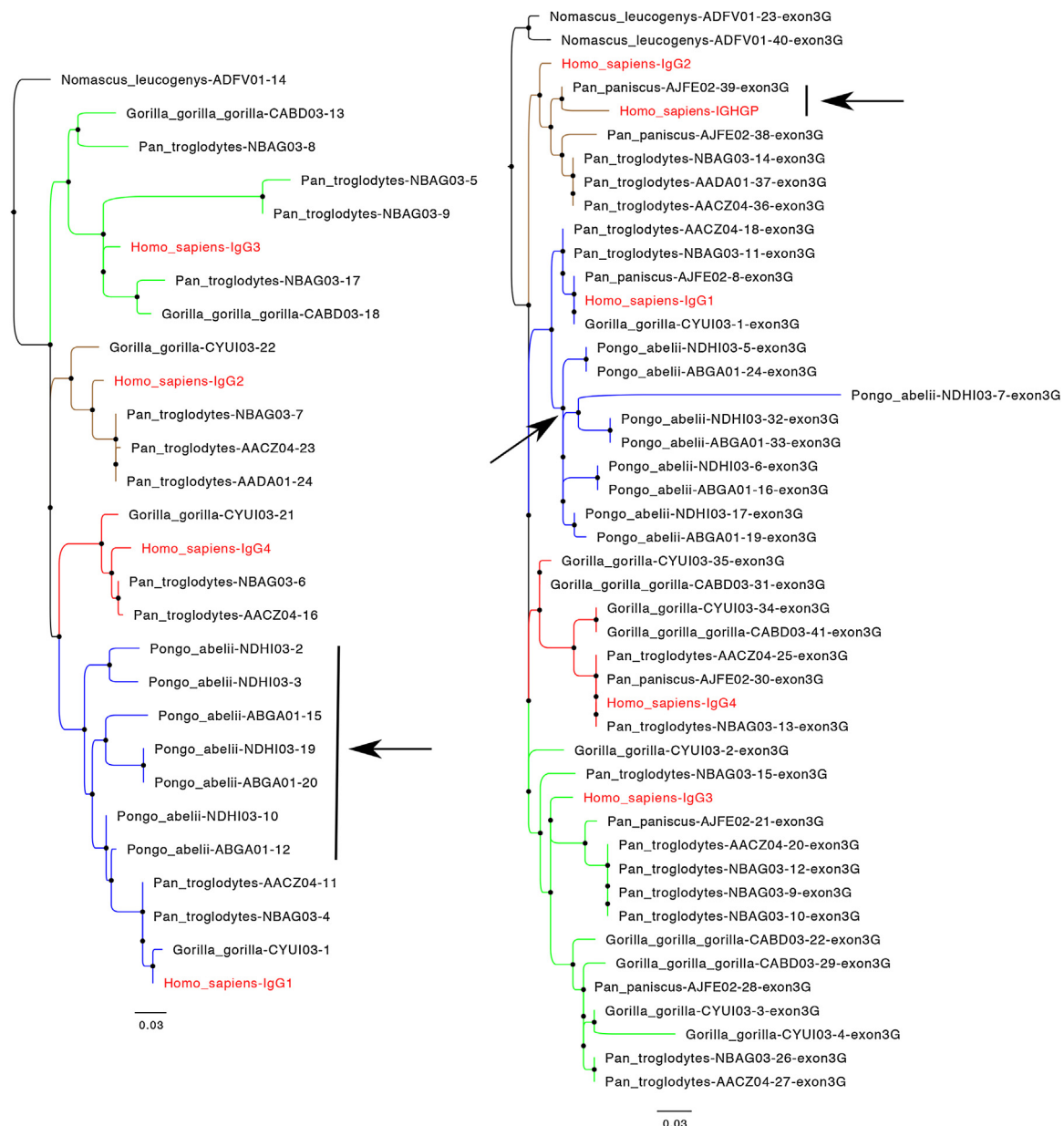


Fig. 7. The tree constructed by aligning the complete IgG sequences (left) and CH3 sequences (right) detected with the *CHfinder* program. To act as markers, the sequences of *H. sapiens* (red) were aligned with the other primate sequences. In each tree, the sequences of *P. abelii* are located in the same clade as the human IgG1 sequences. The human IGHGP sequence is also indicated to show its origin. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

necessary since they have been able to survive without this gene.

In the next family in evolution, the Platyrrhines, *CHfinder* detects CHs of IgD, but only the CH1 and CH3 domains of mammals (equivalent to CH1 and CH8 of reptiles and amphibians). The loss of exon CH2 is already described in murids (Muridae) (Giudicelli et al., 2005). The fact that domains are associated with function suggests that the persistence of the CH1 and CH3 domains has greater relevance. Thus, the CH1 is the domain that receives the VDJ segment and the CH3 is the one that connects with the transmembrane exon (or is secreted); this suggests basic functions maintained in evolution, namely that associated with recognition and the presence on the surface of the cell membrane.

The Cercopithecoidea have five classes of immunoglobulins. Apart from the instability of the IgD just described, the locus is very stable with a gene for each class. This family has two branches: the Colobinae family, characterized with stability of the locus, and the Cercopithecoidea branch, with evident change. For this later family,

evolutionary instability appears in the area of the locus where the IgG is located and duplications are detected in the species studied. This instability is maintained in evolution, and increased in hominids. These results are of interest since it has been considered that simians should have a similar structure in the heavy chain locus and it has been assumed that they should have the same immunoglobulin classes and even subclasses of IgG. The subclasses of IgG in humans appeared very recently and orthology cannot be inferred in *Cercopithecus*. *Macaca mulatta* is often used to study immune response in preclinical studies. The results presented in this work may explain recent contradictory data (Boesch et al., 2016).

The evolutionary instability of the locus in hominids is common. The point of divergence of the orangutan from other Hominids is interesting. The orangutan has several IgGs that have a common origin with human IgG1, but there is no evidence of a relationship with the other three IgG types that humans possess. Gorillas and chimpanzees

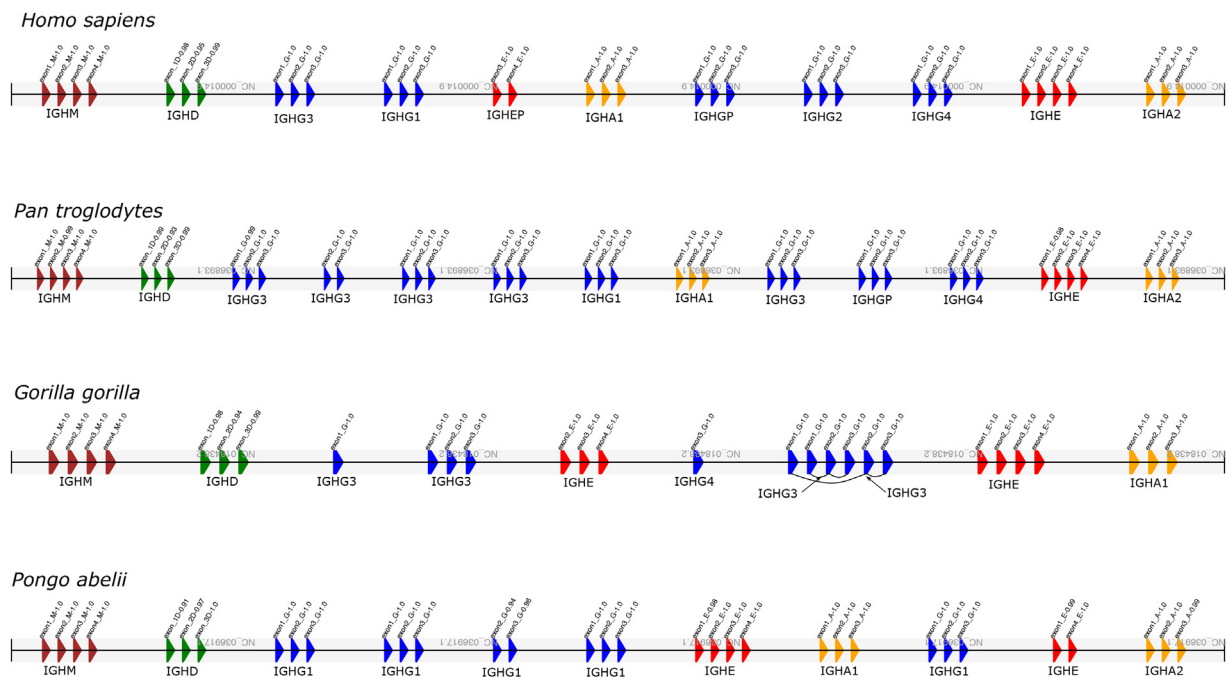


Fig. 8. The schematic representation and location of exons coding for immunoglobulin CHs in Hominids in sequences from the RefSeq Genome Database (refseq_genomes). Although the diagram was made from a listing of the exon locations and order, the distances were adjusted to fit on the same graph. Also, the true sequences are in the complementary chain (the diagram is rotated for illustration). Exons are colored as follows: IgM (brown), IgD (green), IgG (blue), IgA (orange), and IgE (red). In humans, the positions of the various IgG are indicated. In the gorilla, an insertion of a gene inside another is shown (the lines indicate the detected splicing). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

are shown to have orthologous genes to human IgGs. The data also shows evidence of duplications of genes for IgE and IgA, and the presence of other Ig-like exons also suggests the existence of pseudogenes. Such evolutionary instability of the locus may indicate testing the environment through processes of gene duplication. From the results, it appears that the addition of IgG genes seems to give an evolutionary advantage, while the addition of IgEs generates pseudogenes, thereby implying that a single IgE gene is sufficient. An intermediate situation seems to be inferred from IgA.

The instability in the location of genes for IgG suggests a recombination processes or unequal gene exchange between chromatids during meiosis. The most probable hypothesis to explain this instability

would be the generation of a recombinational hotspots with unequal cross-links due to the presence of paralogous genes. The cross-linking of the entire chromatid can explain the elongation of the locus, with the appearance of duplicated sequences coding for IgE and IgA.

By scanning whole genomes, we detected possible exons that code for CH regions located outside the IGH locus. For example, in Platyrrhines and Cercopithecidae, exons that code for CHs are detected for IgE outside the locus. These sequences in some species encompass the complete sequence of the heavy chain constant region of IgE in two exons separated by an intron of less than 50 nucleotides. The results suggest a retrotranscription of a messenger RNA with insertion in DNA in the common ancestor of these species. This most probably does not

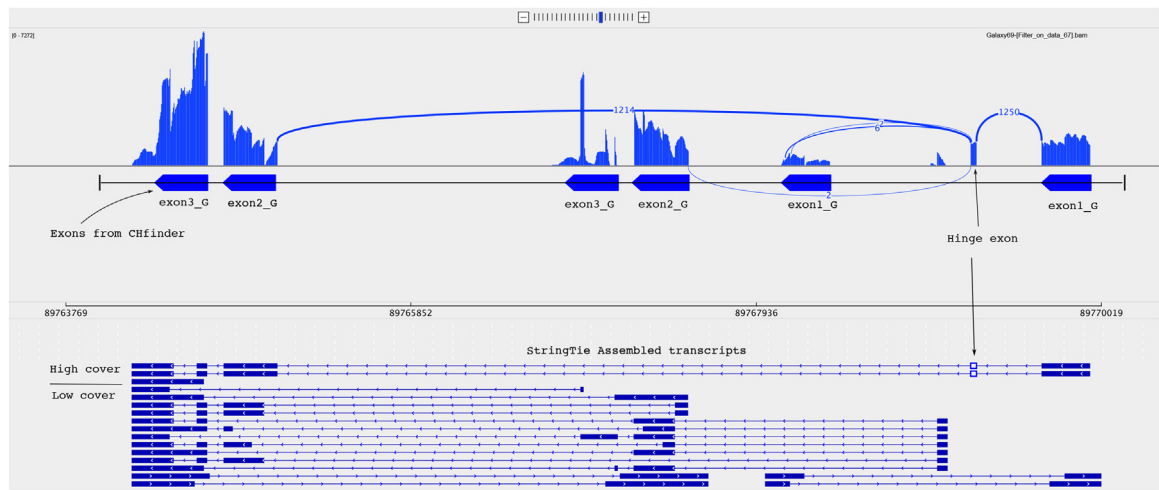


Fig. 9. Sashimi plots of a gorilla liver transcriptome in the IGH locus, showing the splice junctions by alignment of reads to the genome coordinates. From the exon locations found by CHfinder, an IgG gene can be seen inserted in the middle of another IgG. Shown in the lower part of the graph, are the splicing junctions deduced by the StringTie program (Pertea et al., 2015). In the Sashimi plot, only the splicing of the exon hinge is shown; there is a jump sequence track to the second CH2 exon, leaving three exons that code for a second IgG embedded in the middle of the other IgG.

generate a functional gene, although more studies would be necessary to confirm this hypothesis.

In conclusion, we show the immunoglobulin heavy chain locus in Primates possesses several features: the absence of IgD genes in Prosimians, the absence of the exon for CH2 of IgD in Platyrrhines, the onset of inability in the locus in Cercopithecidae and the presence of duplications in hominids with tolerance to the addition of genes for IgG and partially for IgA, while there is evolutionary pressure so that only one gene exist for a viable IgE.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.molimm.2018.07.020>.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: a system for large-scale machine learning. Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16. USENIX Association, Berkeley, CA, USA, pp. 265–283.
- Boesch, A.W., Osei-Owusu, N.Y., Crowley, A.R., Chu, T.H., Chan, Y.N., Weiner, J.A., Bharadwaj, P., Hards, R., Adamo, M.E., Gerber, S.A., et al., 2016. Biophysical and functional characterization of rhesus macaque IgG subclasses. *Front. Immunol.* 7, 589.
- Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al., 2011. The evolution of gene expression levels in mammalian organs. *Nature* 478, 343.
- Cock, P., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., de Hoon, M., 2009. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>.
- Deza, F.G., Espinel, C.S., Beneitez, J.V., 2007. A novel IgA-like immunoglobulin in the reptile *Eublepharis macularius*. *Dev. Comp. Immunol.* 31, 596–605.
- Estevez, O., Garet, E., Olivieri, D., Gambón-Deza, F., 2016. Amphibians have immunoglobulins similar to ancestral IgD and IgA from amniotes. *Mol. Immunol.* 69, 52–61.
- Flanagan, J., Rabbitts, T., 1982. Arrangement of human immunoglobulin heavy chain constant region genes implies evolutionary duplication of a segment containing γ , ϵ and α genes. *Nature* 300, 709.
- Gambón-Deza, F., Espinel, C.S., 2008. IgD in the reptile leopard gecko. *Mol. Immunol.* 45, 3470–3476.
- Gambón-Deza, F., Sánchez-Espinel, C., Magadan-Mompo, S., 2009. The immunoglobulin heavy chain locus in the platypus (*Ornithorhynchus anatinus*). *Mol. Immunol.* 46, 2515–2523.
- Giudicelli, V., Chaume, D., Lefranc, M.-P., 2005. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33, D256–D261.
- Gould, H.J., Sutton, B.J., 2008. IgE in allergy and asthma today. *Nat. Rev. Immunol.* 8, 205.
- Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. <https://doi.org/10.1093/sysbio/syq010>.
- Hedges, S., Dudley, J., Kumar, S., 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* 22, 2971–2972.
- Lefranc, M., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A., Lefranc, G., 2005. IMGT unique numbering for MHC groove G-domain and MHC superfamily (MhcSF) G-like-domain. *Dev. Comp. Immunol.* 29, 917–938. <https://doi.org/10.1016/j.dci.2005.03.003>.
- Mukherjee, S., Stamatis, D., Bertsch, J., Ovchinnikova, G., Verezhenska, O., Isbandi, M., Thomas, A.D., Ali, R., Sharma, K., Kypides, N.C., Reddy, T.B.K., 2017. Genomes online database (gold) v.6: data updates and feature enhancements. *Nucleic Acids Res.* 45, D446–D456. <https://doi.org/10.1093/nar/gkw992>.
- Murphy, K., Weaver, C., 2016. Janeway's Immunobiology. Garland Science.
- Nagaoka, H., Ozawa, K., Matsuda, F., Hayashida, H., Matsumura, R., Haino, M., Shin, E.K., Fukita, Y., Imai, T., Anand, R., et al., 1994. Recent translocation of variable and diversity segments of the human immunoglobulin heavy chain from chromosome 14 to chromosomes 15 and 16. *Genomics* 22, 189–197.
- Ohta, Y., Flajnik, M., 2006. IgD, like IgM, is a primordial immunoglobulin class perpetuated in most jawed vertebrates. *Proc. Natl. Acad. Sci. U. S. A.* 103, 10723–10728.
- Perelman, P., Johnson, W., Roos, C., Seuánez, H., Horvath, J., Moreira, M., Kessing, B., Pontius, J., Roelke, M., Rumpler, Y., Schneider, M., Silva, A., O'Brien, S., Pecon-Slatery, J., 2011. A molecular phylogeny of living primates. *PLoS Genet.* 7, e1001342. <https://doi.org/10.1371/journal.pgen.1001342>.
- Pertea, M., Pertea, G., Antonescu, C., Chang, T.-C., Mendell, J., Salzberg, S., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. <https://doi.org/10.1038/nbt.3122>.
- Preud'homme, J.-L., Petit, I., Barra, A., Morel, F., Lecron, J.-C., Lelievre, E., 2000. Structural and functional properties of membrane and secreted IgD. *Mol. Immunol.* 37, 871–887.
- Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5, e9490.
- Rabbani, H., Pan, Q., Kondo, N., Smith, C.E., Hammarström, L., 1996. Duplications and deletions of the human IGHC locus: evolutionary implications. *Immunogenetics* 45, 136–141.
- Rambaut, A., 2007. Figtree, a Graphical Viewer of Phylogenetic Trees. See <http://tree.bio.ed.ac.uk/software/figtree>.
- Rogers, J., Gibbs, R., 2014. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nat. Rev. Genet.* 15, 347–359. <https://doi.org/10.1038/nrg3707>.
- Sievers, F., Higgins, D., 2014. Clustal Omega, accurate alignment of very large numbers of sequences. *Multiple Sequence Alignment Methods*. Springer, pp. 105–116.
- Tatusova, T.A., Madden, T.L., 1999. BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* 174, 247–250.
- Wang, X., Olp, J.J., Miller, R.D., 2009. On the genomics of immunoglobulins in the gray, short-tailed opossum *Monodelphis domestica*. *Immunogenetics* 61, 581–596.

328 **5. Supplementary information**

Cuadro 6: Characteristics of the WGS of Primates. All values for the contig N50 were > 15kbp, except *N. larvatus* (for which no exons were found). For some assemblies, a reliable number for the coverage could not be ascertained, and are indicated by N/A (not available).

SPECIES	WGS	N50(KBP)	Cov.
LEMURIFORMES			
<i>E. flavifrons</i>	LGHW01	27.3	52x
<i>E. macaco</i>	LGHX01	20.0	21x
<i>P. coquereli</i>	JZKE01	28.1	104.7x
<i>M. murinus</i>	ABDC02	182.9	221.6x
LORISIFORMES			
<i>O. garnettii</i>	AAQR03	27.1	137x
TARSIFORMES			
<i>T. syrichta</i>	ABRT02	38.2	48x
PLATYRRHINI			
<i>C. capucinus</i>	LVWQ01	41.2	81x
<i>A. nancymae</i>	JYKP01	28.5	113.4x
<i>C. jacchus</i>	ACFV01	29.3	6.6x
<i>C. jacchus</i>	BBXK01	61	N/A
<i>S. boliviensis</i>	AGCE01	38.8	80x
CERCOPITHECOIDEA			
<i>N. larvatus</i>	JMHX01	13.3	290
<i>C. angolensis</i>	JYKR01	38.4	86.8x
<i>R. roxellana</i>	JABR01	77.2	53.7x
<i>M. leucophaeus</i>	JYKQ01	31.3	117.2x
<i>C. sabaeus</i>	AQIB01	90.4	95x
<i>M. fascicularis</i>	AQIA01	86	68x
<i>M. mulatta</i>	JSUE03	107	47.4x
<i>M. mulatta</i>	AANU01	25.7	N/A
<i>M. nemestrina</i>	JZLF01	107	113.1x
<i>C. atys</i>	JZLG01	113	192.0x
<i>P. anubis</i>	AHZZ01	40.3	92x
APES			
<i>H. sapiens</i>	ABBA01	100	N/A
<i>P. troglodytes</i>	AADA01	108.4	N/A
<i>P. troglodytes</i>	AACZ04	384.8	70x
<i>P. paniscus</i>	AJFE02	67	26x
<i>G. gorilla</i>	CABD03	53	N/A
<i>G. gorilla</i>	CYUI03	~1000	NA/
<i>P. abelii</i>	ABGA01	15.6	6x
<i>N. leucogenys</i>	ADFV01	35	5.6x