

## Comparative genomic analysis of human and chimpanzee proteases

Xose S. Puente<sup>a,\*</sup>, Ana Gutiérrez-Fernández<sup>a</sup>, Gonzalo R. Ordóñez<sup>a</sup>,  
LaDeana W. Hillier<sup>b</sup>, Carlos López-Otín<sup>a</sup>

<sup>a</sup> Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, 33006 Oviedo, Spain

<sup>b</sup> Genome Sequencing Center, Washington University School of Medicine, St. Louis, MO 63108, USA

Received 19 May 2005; accepted 28 July 2005

Available online 12 September 2005

### Abstract

Proteolytic enzymes are implicated in multiple physiological and pathological processes. The availability of the sequence of the chimpanzee genome has allowed us to determine that the chimpanzee degradome—the repertoire of protease genes from this organism—is composed of at least 559 protease and protease-like genes and is virtually identical to that of human, containing 561 genes. Despite the high degree of conservation between both genomes, we have identified important differences that vary from deletion of whole genes to small insertion/deletion events or single nucleotide changes that lead to the specific gene inactivation in one species, mostly affecting immune system genes. For example, the genes encoding PRSS33/EOS, a macrophage serine protease conserved in most mammals, and GGTLA1 are absent in chimpanzee, while the gene for metalloprotease MMP23A, located in chromosome 1p36, has been specifically duplicated in the human genome together with its neighbor gene *CDC2L1*. Other differences arise from single nucleotide changes in protease genes, such as *NAPSB* and *CASP12*, resulting in the presence of functional genes in chimpanzee and pseudogenes in human. Finally, we have confirmed that the *Trypanosoma* lytic factor HPR is inactive in chimpanzee, likely contributing to the susceptibility of chimpanzees to *T. brucei* infection. This study provides the first analysis of the chimpanzee degradome and might contribute to the understanding of the molecular bases underlying variations in host defense mechanisms between human and chimpanzee.

© 2005 Elsevier Inc. All rights reserved.

**Keywords:** Human; Chimpanzee; Proteases; Degradome; Immune system

The recent sequencing of whole mammalian genomes constitutes a fundamental advance toward understanding the genetic basis of physiological differences between species and will increase our knowledge of molecular aspects of human diseases [1,2]. In this regard, the availability of the chimpanzee (*Pan troglodytes*) genome sequence [3], which is about 99% identical to that of the human genome, represents an excellent opportunity to identify genetic changes that might be responsible for some of the differences between humans and chimpanzees [4]. The study of specific gene families has proven very useful to facilitate the analysis of this large amount of information and can contribute to the understanding of global mechanisms implicated in the

evolution of human and other mammalian organisms [5–11]. Proteolytic enzymes are an excellent model for this kind of analysis, because they constitute a very diverse group of proteins representing more than 1.7% of all human genes. In addition, there is a considerable amount of knowledge of the structure and function of a significant number of proteases (<http://www.merops.ac.uk>), as well as a growing understanding of their relevance in numerous biological and pathological processes [6,12]. This increasing complexity of proteolytic systems has prompted us to introduce novel concepts and experimental approaches into the global study of proteases [12], including the concept of the degradome to define the complete repertoire of protease genes present in one organism.

A characteristic of proteolytic enzymes is their participation in essential biological processes such as cell cycle progression, fertilization and development, migration and neuritogenesis, tissue remodeling, apoptosis, and host de-

\* Corresponding author. Fax: +34 985 103564.

E-mail address: [xspuente@uniovi.es](mailto:xspuente@uniovi.es) (X.S. Puente).

fense. The importance of proteolysis is further reflected by the existence of more than 60 human hereditary diseases due to mutations in protease genes [6] (<http://www.uniovi.es/degradome>). In addition, alterations in the spatiotemporal patterns of expression of proteases are associated with numerous human pathologies, including cardiovascular diseases, neurodegenerative diseases, rheumatoid arthritis, or cancer [13]. However, despite the existence of numerous expression studies of proteases in normal and pathological conditions, little is known about the biological function of many of these protease genes. In this regard, comparison of human and chimpanzee degradomes can provide new insights into the physiological and pathological relevance of proteolytic enzymes and contribute to identifying individual genes or general processes that might have been important for the evolution of these closely related organisms.

In this work, we report the analysis of the complete set of protease and protease-like genes in the chimpanzee genome and its comparison to the human degradome. We show that the chimpanzee degradome is composed of 559 protease and protease-like genes with more than 99.1% identities with their human orthologs. We also describe the identification and detailed analysis of several protease genes that are different between human and chimpanzee and that might be responsible for some of the physiological differences between both species. On the basis of this analysis, we conclude that most protease-gene differences between both organisms are found in immune system genes, highlighting the importance of analyzing a limited set of genes to identify specific processes that might have been altered as a result of evolution.

## Results

### Comparison between human and chimpanzee protease genes

The chimpanzee degradome was identified and characterized by screening the chimpanzee genome with the complete set of human protease genes as described under Materials and methods. By following this strategy, we have concluded that the chimpanzee degradome is virtually identical to that of human, being composed of at least 559 protease genes grouped in 68 different families (Fig. 1 and Supplemental Table 1). The genome coverage for this set of genes is very high (95.6% at the nucleotide level), similar to the draft genome average [3]. In this sense, it cannot be ruled out that other protease-coding genes could be present in unfinished regions of the chimpanzee genome, although this possibility is very unlikely due to the high coverage of the chimpanzee genome and the fact that most gaps are smaller than the average size of a protease gene [3]. Despite the high coverage of the chimpanzee genome, we observed the presence of frameshifts in coding sequences in at least 30 genes (5.3%), thereby making it difficult to distinguish pseudogenized genes from sequencing errors. As most of these putative frameshifted chimpanzee genes are functional in mouse or rat, we resequenced those genes in which sequence problems were located within regions of high-quality nucleotide sequence in both ARACHNE and PCAP chimpanzee assemblies. This analysis revealed that all observed frameshifts were due to sequence problems in the assemblies (Supplemental Table 1). Finally, and similar to the situation with available assemblies for other genomes, the chimpanzee assemblies exhibit problems in regions where tandem duplication of genes has

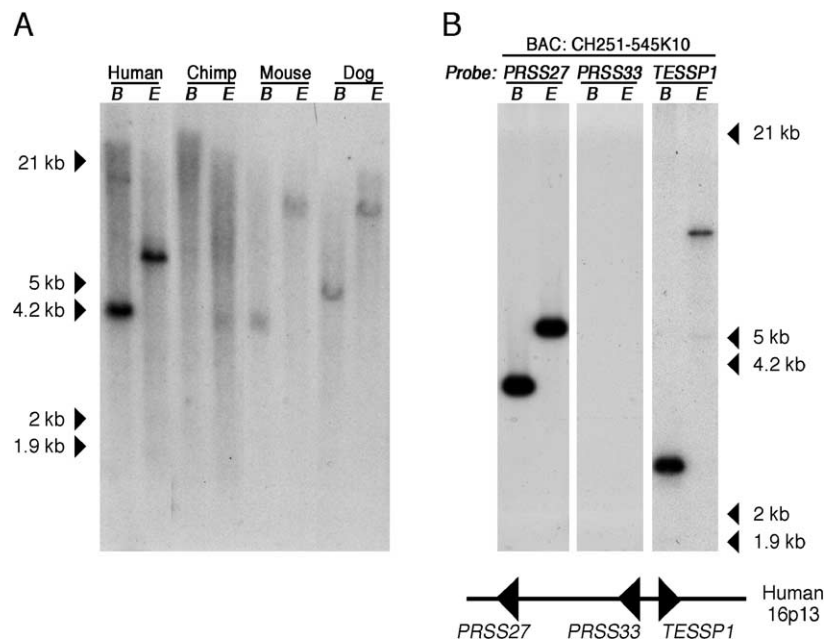


Fig. 1. *PRSS33* is absent from the chimpanzee genome. (A) Southern blot analysis of genomic DNA from the indicated species with a human *PRSS33* probe. (B) Chimpanzee BAC CH251-545K10, corresponding to the syntenic region of human chromosome 16p13, was analyzed by Southern blot with probes for human *PRSS27*, *PRSS33*, and *TESSP1* genes. The organization of this region in the human genome is shown below. Restriction enzymes are designated by *B* (*Bam*HI) and *E* (*Eco*RI).

occurred, which are usually artifactually collapsed during the assembly process. Nevertheless, and despite these limitations, we have completed and evaluated data on 6760 chimpanzee exons, and only 419 putative exons were not identified in the sequence.

This analysis allowed us to conclude that there are 557 protease orthologous genes between human and chimpanzee, showing an average of 99.06 and 99.11% identities at the amino acid and nucleotide level (median 99.31 and 99.22%), respectively, compared to the genome average of 98.77% [3]. Interestingly, there are at least 75 protease genes (13.5%) that code for proteins completely identical to their human counterparts (Supplemental Table 1). This group of highly conserved genes includes several housekeeping protease genes, such as those encoding the different components of the proteasome, but there are also numerous genes highly expressed in the nervous system and implicated in neurological disorders, as is the case of *BACE1*, *PSEN1*, *UCHL1*, or *IMMP2L* [14–17]. This observation emphasizes the idea that most differences between both species in this regard are not present in the coding sequences but in regulatory elements, as previously shown for genes expressed in the CNS [18,19]. In addition, our analysis also revealed a number of chimpanzee protease genes that differ significantly (by more than 2 standard deviations) from their human orthologs, with identities at the protein level below 97% (Table 1). In all cases the amino acid substitutions were located within the protease domain, suggesting that they might influence the functional activity of the protease or its substrate specificity. The observation that for some of these genes the ratio of nonsynonymous ( $K_a$ ) to synonymous ( $K_s$ ) substitutions was greater than 1 indicates that these genes have been subjected to positive selection during the evolution. This group contains many protease genes implicated in immunological processes, including proteases from neutrophilic granules as *PRTN3*, the natural killer granzymes *GZMH* and *GZMM*, mast

cell granule proteases as *TPS1* or *TPSB2*, or the complement factor D. Other genes showing a high degree of divergence include the pancreatic elastase A (*ELA3A*) or the apoptosis-related protease *CASP5*. These data are in agreement with previous studies showing major differences in host defense proteases between human and rodents [6,20] and suggest that host defense has been one of the main driving forces in the evolution of chimpanzee and human proteases.

We next examined the human and chimpanzee genomes for the presence of specific gene duplications, as recent studies suggest that at least one protease gene (*USP10*) has been duplicated in the human genome compared to chimpanzee [21]. This analysis led us to find a protease gene, *MMP23*, encoding a matrix metalloproteinase predominantly expressed in reproductive tissues [22,23] and duplicated in the human lineage (*MMP23A* and *MMP23B*). By contrast, we did not find evidence of *USP10* duplication in the human lineage. Interestingly, we found that the human genome contains two processed *USP10* pseudogenes on chromosomes 14 and 22, with numerous frameshifts that disrupt the reading frame and result in premature stop codons. However, the chimpanzee genome contains only one copy of the *MMP23* gene and does not contain *USP10* pseudogenes and shows only a functional *USP10* gene at chromosome PTR18 syntenic to the human *USP10* gene located at 16q24. These findings could contribute to explaining the differences observed in this regard when comparative genomic hybridization is used to identify duplicated genes.

Finally, the detailed comparative analysis of human and chimpanzee degradomes led us to identify two additional human genes—*PRSS33* and *GGTLA1*—for which we could not find an ortholog in the chimpanzee genome. Likewise, we found three chimpanzee protease genes—*NAPSB*, *CASP12*, and *HPP*—that are inactive or absent in human, as well as one human gene—*HPR*—that has been specifically inactivated in

Table 1  
Differential and divergent protease genes between human and chimpanzee

Protease	Gene	Human locus	Status/identity (%)	$K_a/K_s$ ratio	Function
Caspase 12	<i>CASP12</i>	11q22	Pseudogene in human	–	Host defense, Alzheimer disease
EOS serine protease	<i>PRSS33</i>	16p13	Absent in chimp	–	Host defense (?)
$\gamma$ -Glutamyltransferase-like A1	<i>GGTLA1</i>	22q11	Absent in chimp	–	Host defense(?)
Haptoglobin primate	<i>HPP</i>	–	Absent in human	–	Unknown
Haptoglobin-related protein	<i>HPR</i>	16q22	Pseudogene in chimp	–	Putative <i>Trypanosoma</i> lytic factor, host defense
Matrix metalloproteinase 23A	<i>MMP23A</i>	1p36	Absent in chimp	–	Unknown, expressed in lymphoid tissues
Napsin B	<i>NAPSB</i>	19q13	Pseudogene in human	–	Unknown, expressed in lymphoid tissues
Pancreatic endopeptidase E (A)	<i>ELA3A</i>	1p36	94.44	0.3851	Digestion and cholesterol metabolism
Proteinase 3	<i>PRTN3</i>	19p13	94.53	1.0598	Neutrophilic granules protein, host defense
Plasma-kallikrein-like 1	<i>KLKBL1</i>	4q35	95.01	0.8885	Unknown
$\gamma$ -Glutamyltransferase-like 4	<i>GGTL4</i>	22q11	95.11	0.5575	Unknown
Tryptase $\alpha$ /tryptase $\beta$ 1	<i>TPS1/TPSB1</i>	16p13	95.27	0.3237	Mast cell granules protein, host defense
Granzyme H	<i>GZMH</i>	14q11	95.53	2.8784	Neutrophilic granules protein, host defense
Tryptase $\beta$ 2	<i>TPSB2</i>	16p13	96.00	0.2665	Mast cell granules protein, host defense
Caspase-5	<i>CASP5</i>	11q22	96.17	1.7679	Processing of proinflammatory cytokines, host defense
Complement factor D	<i>DF</i>	19p13	96.44	0.8051	Complement factor D, host defense
Cathepsin G	<i>CTSG</i>	14q11	96.86	0.4663	Neutrophilic granules protein, host defense
Granzyme M	<i>GZMM</i>	19p13	96.86	0.2507	Neutrophilic granules protein, host defense
Plasma-kallikrein-like 2	<i>KLKBL2</i>	8p23	96.88	1.3827	Unknown

The status for differential genes is indicated, while for divergent genes the percentage of identity at the protein level is shown.

chimpanzee (Table 1). Because these genes might contribute to explaining some of the physiological differences between human and chimpanzee, we performed a detailed analysis of all of them in the chimpanzee genome.

#### Analysis of protease genes absent from the chimpanzee genome

Analysis of the chimpanzee degradome revealed that the gene encoding the serine protease EOS was absent from the chimpanzee genome. This gene has been conserved throughout evolution, being present in human, mouse, rat, dog, cow, and opossum genomes [20,24] (and data not shown). As this gene is located in tandem with other serine protease genes in human chromosome 16p13, its absence in the chimpanzee genome could be the result of a sequence being artifactually collapsed during the assembly process. To clarify this question, we performed Southern blot analysis with genomic DNA from different mammalian species using as a probe a 409-bp fragment corresponding to human *PRSS33*. We detected a hybridization band of the expected molecular weight in human, mouse, and dog DNA (Fig. 1A), while no band was observed in chimpanzee DNA, suggesting that *PRSS33* is absent from this primate. To investigate this finding further, we analyzed two chimpanzee BAC clones spanning this region (CH251-545K10 and CH251-350O2) for the presence of *PRSS33* and the immediately upstream and downstream genes in the human genome—*PRSS27/MPN* and the pseudogene *TESSP1*, respectively (Fig. 1B). As can be seen in Fig. 1B, *PRSS27* and *TESSP1* were present in BAC CH251-545K10, while *PRSS33* could not be detected in the same BAC nor in other BACs from the same region (CH251-350O2, RP43-59F8, RP43-40A16) (data not shown).

On the other hand, analysis of the chimpanzee genome revealed the presence of a single gene for the matrix metalloprotease *MMP23*, while the human genome contains two closely related genes (99.9% identity at the nucleotide level) encoding *MMP23A* and *MMP23B* and located in a recently duplicated region of chromosome 1p36 [23]. A detailed analysis of these genes revealed that chimpanzee *MMP23* was more closely related to human *MMP23B*, suggesting the absence of *MMP23A* in chimpanzee. However, the high sequence identity between both genes could result in their artifactual collapse in the chimpanzee genome assembly, as still is the case in the more recent human genome assembly (NCBI build 35, May 2004). To investigate the absence of *MMP23A* from the chimpanzee genome, we performed Southern blot analysis using genomic DNA from human and chimpanzee, BAC clones CH251-252K11, CH251-541E6, and CH251-696F3 corresponding to this region, and PCR amplification followed by direct sequencing of the *MMP23* gene. We also investigated the presence of one or two copies of the neighbor gene *CDC2L*, which is also duplicated in the same region of the human genome resulting in two almost identical genes, *CDC2L1* and *CDC2L2* [23]. This analysis revealed that the chimpanzee genome appears to contain a single gene for both *MMP23* and *CDC2L*, as we detected a single hybridization band corresponding to both genes in chimpanzee, while

two bands were clearly present in human (Figs. 2A and 2B). By PCR amplification and direct sequencing of *MMP23* and *CDC2L* genes, we also confirmed that the chimpanzee contains a single copy for both genes, which likely correspond to *MMP23B* and *CDC2L1*. These data support the idea that this region has been specifically duplicated in the human lineage and might represent a region prone to genetic instability, as demonstrated by the observation of frequent translocations and deletions involving this region in different human tumors [23,25,26].

The last identified protease gene absent from the chimpanzee genome is *GGTLA1*, encoding a threonine protease of the  $\gamma$ -glutamyltranspeptidase family. This gene is located in human chromosome 22q11, a region that has undergone numerous intrachromosomal duplications leading to the generation of four closely related genes encoding  $\gamma$ -glutamyltranspeptidases and several pseudogenes [27]. The chimpanzee genome also contains several copies of this group of genes, as well as some pseudogenes, reflecting the dynamic nature of this region (Supplemental Table 1 and data not shown). The high sequence identity between the different human and chimpanzee genes and the variable number of pseudogenes in both species greatly hamper the analysis of this region. Moreover, this gene is located within the immunoglobulin  $\lambda$  chain C locus, a region of genomic instability that might be polymorphic between different individuals [28] and might contribute to explaining its absence in chimpanzee.

#### Analysis of differential pseudogenization in protease genes from human and chimpanzee

Because a common mechanism in the evolution of human proteases has been the pseudogenization of several reproductive and immune system genes that are active in other species

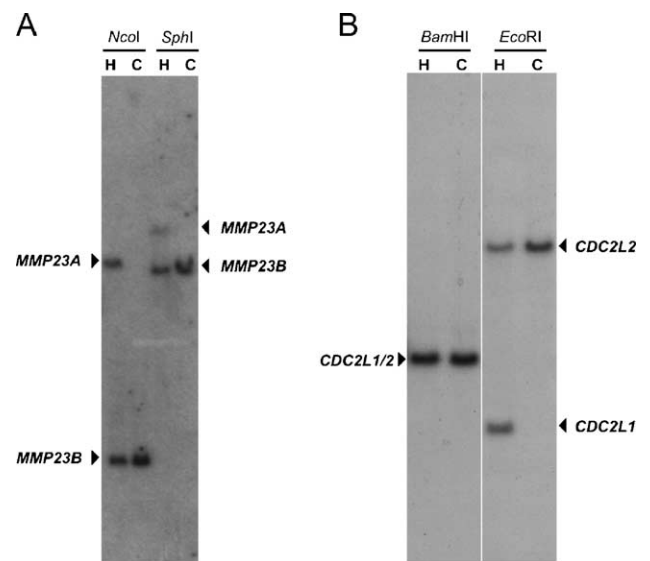


Fig. 2. *MMP23* and *CDC2L* genes are specifically duplicated in the human genome. (A) Southern blot analysis of human (H) and chimpanzee (C) genomic DNA digested with either *NcoI* or *SphI*, using a probe for human *MMP23*. (B) Southern blot hybridization of human and chimpanzee DNA with a probe for human *CDC2L*. Band identities are indicated by arrowheads.

[20,29], we next examined the status of these genes/pseudogenes in chimpanzee. Despite the above-mentioned limitations in the analysis, we can conclude that all examined human pseudogenes related to reproductive proteases are also inactive in chimpanzee, although there are some cases (*ADAM6*, *PRSS29P*) that have been pseudogenized by different mechanisms in human and chimpanzee (Supplemental Fig. 1). In the case of immune system genes, we have found that the gene for napsin B (*NAPSB*), which had been previously described as a human pseudogene [30], appears to be functional in chimpanzee. This pseudogene is located in a duplicated region in human chromosome 19q13 that contains the aspartic protease gene *NAPSA*, as well as the related pseudogene *NAPSB*, which lacks an in-frame stop codon and leads to the synthesis of a nonfunctional protein [30]. However, after PCR amplification and direct sequencing, we have verified that the corresponding region in the chimpanzee genome contains an in-frame stop codon at a position equivalent to that of human and chimpanzee *NAPSA* (Fig. 3A), strongly suggesting that chimpanzee *NAPSB* codes for a functional aspartic protease. Finally, the observation that the human *NAPSB* pseudogene is specifically expressed in lymphoid tissues [31] also suggests that its active chimpanzee counterpart might contribute to some of the functional differences between human and chimpanzee immune systems [32–35].

Another interesting case is that of caspase-12 (*CASP12*), which has acquired two different mutations in humans—a frameshift mutation in exon 3 and a nonsense mutation in exon 4—while its mouse and rat orthologs are functional enzymes. Therefore, *CASP12* had been classified as a pseudogene in humans [36]. However, it has been recently described that as much as 20% of the South African and African American population contains a polymorphism in which the stop codon (TGA) in exon 4 is replaced by an Arg codon (CGA) [37]. This polymorphism leads to the production in some humans of a full-length caspase proenzyme by deletion of exon 3 through alternative splicing [37]. Direct sequencing of this region in

chimpanzee DNA samples from different geographical regions has allowed us to confirm that chimpanzee *CASP12* presents an Arg codon (CGA) instead of the TGA codon present in most humans, suggesting that in chimpanzees *CASP12* is a functional gene (Fig. 3B). Furthermore, it must be emphasized that human caspase-12 does not contain a functional SHG box, a crucial element for the enzymatic activity of caspases, thereby indicating that even the full-length isoform of caspase-12 found in some humans lacks protease activity [36,37]. Interestingly, chimpanzee caspase-12 contains a complete SHG box as well as all residues involved in catalysis (Fig. 3B), suggesting that it is an active protease, similar to its murine and rat counterparts.

Finally, we have confirmed in the chimpanzee genome the previous experimental observation that a human serine protease-like gene (haptoglobin-related protein, *HPR*) is a pseudogene in chimpanzee [38]. *HPR* is a highly polymorphic gene in human and exhibits multiple copies in the black population. By contrast, the haptoglobin primate gene (*HPP*) is present in chimpanzee and absent in human [39]. This *HPR* gene has been identified as the *Trypanosoma* lytic factor and could contribute to explaining the human resistance to the disease caused by *T. brucei* in other species including chimpanzee [40]. This finding prompted us to analyze other genes for putative lytic factors and allowed us to conclude that *APOL1*, another proposed lytic factor for *T. brucei* [41], is absent in chimpanzee (see Supplemental Fig. 2).

#### Analysis of insertions, deletions, and disease variants in chimpanzee proteases

The insertion or deletion of specific residues can influence the activity or stability of a protein and constitutes a mechanism underlying numerous human diseases [42]. The analysis of human and chimpanzee degradomes allowed us to identify 26 chimpanzee proteases that contain inserted or deleted residues compared to the human sequence (Supple-

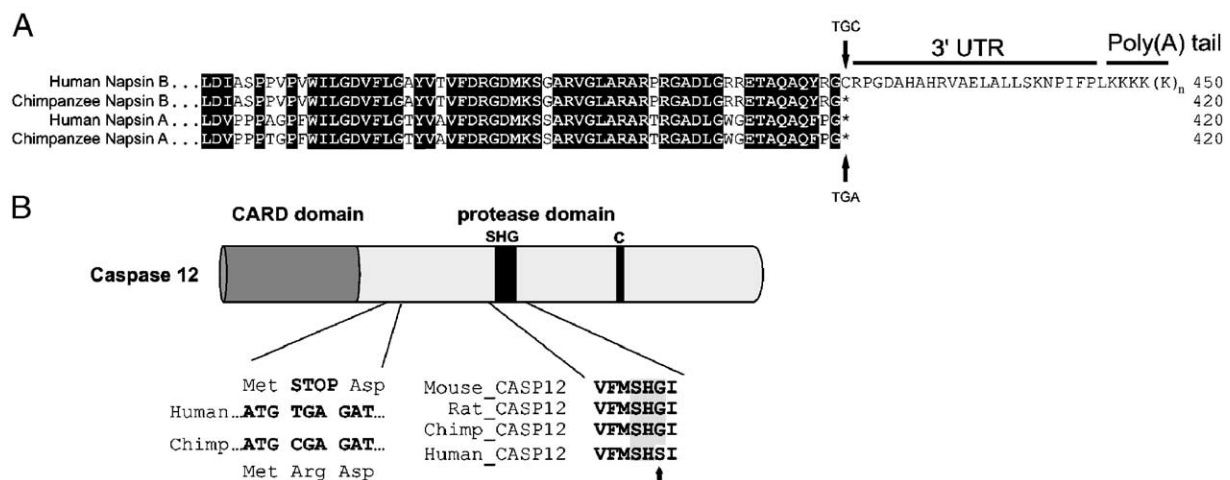


Fig. 3. Napsin B and caspase-12 are functional proteases in chimpanzee. (A) Multiple amino acid alignment of the C-terminal region of human and chimpanzee napsins. Human napsin B (*NAPSB*) lacks an in-frame stop codon. The conceptual translation of the 3'UTR region and poly(A) tail is shown. (B) Domain organization of caspase-12 showing the premature stop codon in exon 4 present in more than 90% of the human population. Alignment of the SHG box region caspase-12 from different mammals shows the absence of a critical residue for proteolysis in human caspase-12.

mental Table 1). In most cases, the insertion/deletion event affected just one or two residues, and in 42% of the cases it was located in a highly repetitive region that could be unstable and even polymorphic in humans. However, we found two genes, *ADAMTS20* and *MMP9*, which showed a significant number of deleted amino acids compared to the human sequence. In the case of *ADAMTS20*, a multidomain metalloprotease expressed in testis and brain [43], we identified a deletion of exons 22 and 23 in a region that does not show gaps in the assembly. PCR amplification and direct sequencing of this region allowed us to confirm this deletion, which results in the production of a mature protein lacking residues 1031 to 1125 (human numbering) (Fig. 4A). Deletion of exons 22 and 23 results in a truncated protein lacking two of the C-terminal thrombospondin-1 repeats that are conserved in the human and mouse sequences (Fig. 4A). In the case of *MMP9*, we observed the deletion of eight residues (483 to 490, human numbering) in the hinge region [13] (Fig. 4B). These results suggest that insertion/deletion events in protease sequences could represent an additional mechanism to generate diversity among proteolytic enzymes from human and chimpanzee.

Finally, detailed analysis of the 64 genes mutated in human hereditary diseases of proteolysis [6] has revealed that all of them have a conserved ortholog in chimpanzee, including caspase-10, mutated in a lymphoproliferative syndrome, and there are no orthologs in mouse and rat. Nevertheless, there are human disease variants in protease genes present in chimpanzee. This is the case with *PRSS1*, which encodes cationic trypsinogen. A mutation, N29T, enhances intrapancreatic trypsin activity and causes autosomal dominant hereditary pancreatitis [44]. We have verified by PCR amplification and sequence analysis the presence of Thr at position 29 in chimpanzee *PRSS1*, thereby providing an additional example of the usefulness of the chimpanzee genome sequence to identify differences in genes associated with human diseases.

### Discussion

In this work, we have performed a comparative analysis of the chimpanzee and human degradomes, the complete set of protease genes from these organisms. The chimpanzee degradome contains 559 protease-coding genes, a number similar to that of the human degradome (561 genes). Despite the high sequence conservation between human and chimpanzee proteases (99.1%), a number of protease genes implicated in immunological functions show a higher degree of divergence. Moreover, we have identified 7 differential protease genes between human and chimpanzee, and most of them are associated with the immune system. An interesting observation in this regard is that the two proposed *Trypanosoma* lytic factors of human serum, haptoglobin-related protein and apoL1 [40,41], are inactive or absent from the chimpanzee genome. This difference might explain the absence of *Trypanosoma* lytic factors in serum from chimpanzee and the susceptibility of chimpanzees to the disease caused by *T. brucei*. This finding highlights the utility of genome comparison to identify genes responsible for the observed phenotypic differences between closely related species. Additionally, we have identified other genes associated with the immune system that are either absent in one species, including *PRSS33*, *CASP12*, or *NAPSB*, or highly divergent and subject to positive selection, including proteases from neutrophilic granules such as *PRTN3* or *GZMH* or proteases implicated in the processing of proinflammatory cytokines such as *CASP5*, which might contribute to differences in host defense between both organisms. These findings support previous observations that indicate that host defense has been a major driving force during evolution [20,32–35], probably due to exposure to different pathogenic agents and/or parasites. The identification of these differential protease genes between human and chimpanzee represents an starting point to explore the physiological consequences associated with these

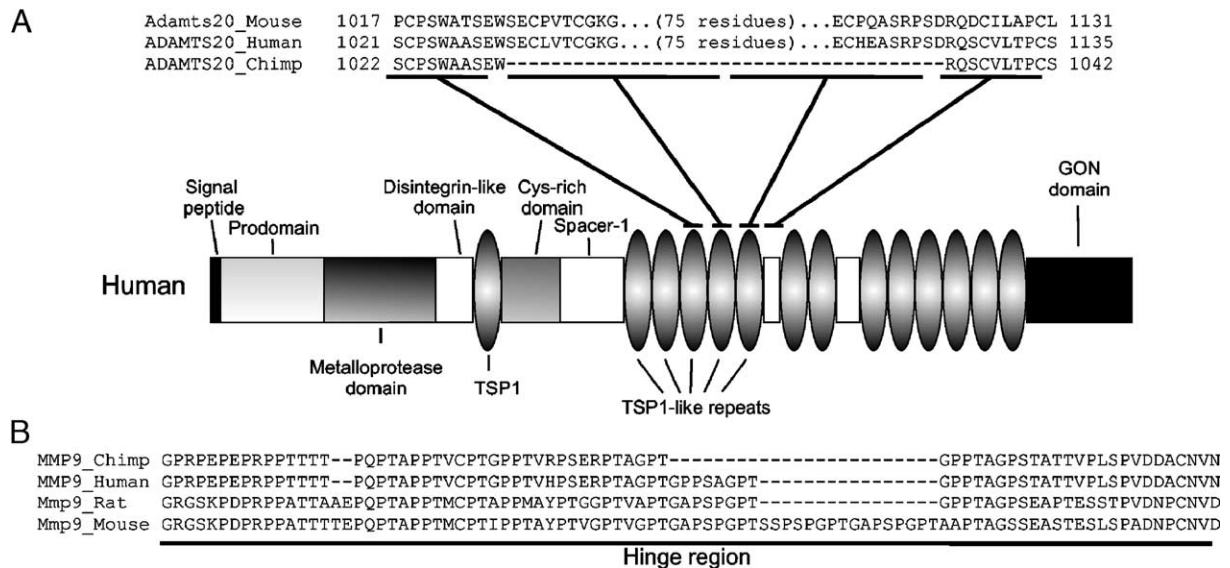


Fig. 4. *ADAMTS20* and *MMP9* show deletions in chimpanzee. (A) Amino acid alignment and domain structure organization of human and chimpanzee *ADAMTS20* showing the specific deletion of two TSP1-like repeats in chimpanzee due to a deletion through exons 22 and 23. (B) Amino acid alignment of the hinge region of *MMP9* from different mammalian organisms shows that chimpanzee *MMP9* has the shortest hinge region.

changes, although further studies will be required to understand fully the relevance of the findings described in this work. In this sense, recent studies have provided increasing evidence of the importance of the immune system in complex pathological processes such as cancer susceptibility and progression [45–47]. Therefore, it is tempting to speculate that differences in protease genes implicated in immunological functions might contribute to explain the reduced incidence of solid tumors in chimpanzee [48,49].

During this protease gene comparative analysis, we also confirmed and extended previous observations suggesting that pseudogenization of specific genes has contributed to the evolution of mammalian genomes [4]. Thus, ADAM6 and ISP2, implicated in reproduction in mouse and rat, have been inactivated in both human and chimpanzee genomes. However, it is interesting to note that they have been inactivated by different mechanisms, suggesting that the pseudogenization of these genes occurred after the human–chimpanzee split. We have also confirmed the presence in chimpanzee of genes encoding important reproductive proteases that are absent in mouse and rat. This is the case of prostate-specific antigen (KLK3), which is involved in the proteolytic cleavage of semenogelins (SEMG1 and SEMG2), the major structural proteins in human semen. Interestingly, analysis from chimpanzee assemblies has corroborated the recent finding of marked differences between human and chimpanzee semenogelins [50]. Thus, SEMG1 is extended at the C-terminal end, originating a considerably longer protein than its human counterpart. By contrast, SEMG2 is truncated due to the presence of a premature stop codon. Therefore, despite the absence of marked changes in reproductive proteases between human and chimpanzee, differences in their targeted substrates may be responsible for some of the changes in the reproductive biology of both species.

Likewise, we have also identified two genes, coding for napsin B and caspase-12, which appear to be functional in chimpanzee, but have been inactivated by mutations in the human genome. We still do not know the functional significance of the presence of napsin B in chimpanzee, but due to its putative expression in lymphoid tissues, it might participate in host defense functions. The case of *CASP12* is also noteworthy because although it has long been considered a human pseudogene due to the presence of two different mutations [36], a recent study has demonstrated that as much as 20% of the South African and African American population may produce a complete protein [37]. Furthermore, the fact that the intact isoform of human caspase-12 has changes in key residues suggests that humans have lost the original function of this gene, while in chimpanzees it should be a bona fide protease. This gene has been implicated in immunological functions, as it is expressed in the immune system, and individuals producing caspase-12 have an increased risk of severe sepsis [37]. However, the presence of an active caspase-12 protease in chimpanzee and rodents could lead to the activation of inflammatory mediators and result in differences in the inflammatory response between humans and other species. Additionally, it has been shown

that caspase-12 mediates  $\beta$ -amyloid-induced apoptosis in mice [51], raising the hypothesis that the inactivation of human *CASP12* might contribute to an increased risk to develop Alzheimer disease.

The comparison between human and chimpanzee proteases also allowed us to identify numerous insertion/deletion events that can influence the activity of these enzymes. An interesting finding is the specific deletion of two complete exons from *ADAMTS20* that results in the absence of two thrombospondin-1 repeats conserved from rodents to humans. The functional significance of this deletion is unclear, but it has been proposed that thrombospondin repeats could be implicated in substrate specificity or interaction with extracellular matrix proteins, which could affect the activity of chimpanzee *ADAMTS20*. Interestingly, this region has been shown to be important in the activity of this protease, as inactivating mutations in murine *Adamts20* and in *gon1*, the *Caenorhabditis elegans* ortholog of *ADAMTS20*, are localized to this C-terminal region [52,53]. The fact that *ADAMTS20* is expressed at high levels in testis [43] agrees with previous studies that show that genes implicated in reproductive functions have been subjected to numerous changes during evolution [20,50] and could contribute to explaining some of the differences in reproduction between human and chimpanzee. Additionally, we found that chimpanzee MMP9—a multifunctional protease with an important role in the regulation of inflammatory processes [47,54]—has eight residues deleted within the hinge region. Similar to the case of *ADAMTS20*, the hinge region in MMPs is believed to be important in substrate specificity by keeping the metalloprotease domain at an appropriate distance from the hemopexin domain [55]. Together, these results suggest additional mechanisms by which genetic changes might generate functional diversity in this group of enzymes by affecting substrate specificity, proteolytic activity, or sensitivity to specific inhibitors.

Finally, we have found two genes (*PRSS33* and *GGTLA1*) of putative immunological relevance that are absent from the chimpanzee genome. The gene encoding the EOS serine protease is expressed in macrophages and is present in all sequenced mammalian genomes to date [20,24]. In human, this gene is located at chromosome 16p13 in a region containing four serine protease genes in tandem, as well as two pseudogenes and multiple repetitive elements. This region appears to be very unstable and has undergone an extensive reorganization during evolution, as deduced from the different numbers of genes contained in the equivalent region in rodents and primates. It is likely that inversions have occurred more than once during evolution and resulted in the elimination of *PRSS33* from the chimpanzee genome. Our preliminary studies suggest that *PRSS33* is present in gorilla, orangutan, and mandrill (data not shown), which implies that somehow *PRSS33* was deleted after the split from the last human–chimpanzee common ancestor. Likewise, the gene encoding *GGTLA1* appears to be absent in chimpanzee. This gene encodes a member of the  $\gamma$ -glutamyltranspeptidase family, a group of enzymes that have been implicated in T-cell-mediated immune responses [56].

In summary, this work provides the first overview of the chimpanzee degradome and constitutes a first approach to understanding the genetic basis underlying differences in protease-mediated processes between human and chimpanzee. We have analyzed more than 550 genes and showed that most differences are found in protease genes implicated in immunological functions. Human–chimpanzee differences in proteolytic genes vary from deletion of whole genes to small insertion/deletion events or to single nucleotide changes that lead to gene inactivation in one species or the other. Nevertheless, the small number of differences between human and chimpanzee degradomes suggests that changes in regulatory elements or in protease inhibitors can also contribute to differences in proteolysis between both organisms. Studies now in progress aimed at analyzing these differential genes further will help to define their precise contribution to the physiological differences between human and chimpanzee.

## Materials and methods

### Bioinformatic analysis

The human degradome database has been previously described [6,20] and is detailed in Supplemental Table 1. Due to the high percentage of identities between human and chimpanzee genomes, human protease cDNA sequences obtained from the RefSeq and GenBank databases were first curated to remove sequencing errors that would result in artifactual differences between human and chimpanzee proteases. During this procedure, we realized that about 60% of the RefSeq deposited sequences contained nucleotide changes or frameshifts that affected the protein sequence and were not supported by the human genome sequence, EST sequences, or the HapMap polymorphism database, reflecting errors in the RefSeq entries. The two available chimpanzee genome assemblies, ARACHNE and PCAP [3], were screened for the presence of chimpanzee orthologs of human protease genes, by using the curated set of human protease cDNA sequences and a combination of BLAT and BLAST algorithms [57,58]. After *in silico* searches, the predicted chimpanzee cDNA and protein sequences were extracted and compared to the human orthologs. All sequences and alignments were manually inspected to identify putative frameshifts or incomplete sequences due to gaps in the assemblies. Significantly divergent proteases were defined as those showing a percentage of identity at the protein level below 2 standard deviations of the mean (96.88%). The  $K_a/K_s$  ratio was calculated for the full-length protein using the program K-Estimator [59].

Identification of chimpanzee-specific proteases was carried out by three approaches: (a) screening of the chimpanzee genome with protease amino acid sequences using the TBLASTN algorithm, (b) analysis of chimpanzee regions syntenic to human pseudogenes that are active in other species, and (c) specific comparison of clusters of protease genes for the presence or absence of proteolytic genes.

### DNA samples, PCR amplification, and DNA sequencing

Chimpanzee genomic DNA was obtained from WES cells and from chimpanzees from different geographical regions (a gift from Dr. J. Bertranpetit, CRG-UPF, Barcelona, Spain), human DNA from a healthy male individual, mouse (C57BL/6J) DNA from animals from our inbred colonies, and dog DNA from MDCK cells. Chimpanzee BAC clones CH251-545K10, CH251-350O2, CH251-252K11, CH251-541E6, CH251-696F3, RP43-59F8, and RP43-40A16 were obtained from Children's Hospital Oakland Research Institute (<http://bacpac.chori.org>). Conflictive regions, defined as those presenting a frameshift, premature stop codon, or absence of initiation codon in both ARACHNE and PCAP assemblies, were PCR-amplified from chimpanzee and human genomic DNA by using specific oligonucleotides (see Supplemental Table 2) and a GeneAmp 2400 PCR system. Nucleotide

sequencing was performed in an ABI Prism 310 sequencer (Perkin–Elmer Life Sciences).

### Southern blot hybridization

Genomic DNA (10  $\mu$ g) from the indicated species or BAC DNA (3  $\mu$ g) was digested for 16 h, and DNA fragments were separated in agarose gels and transferred to nylon membranes. Hybridization was performed with probes amplified from genomic DNA with specific oligonucleotides (see Supplemental Table 2), washed at 65°C with 0.1 $\times$  SSC, 0.1% SDS for 30 min, and autoradiographed.

## Acknowledgments

We thank J. Bertranpetit for DNA samples, T. Graves for help with BAC clones, and F. Rodríguez and S. Alvarez for excellent technical assistance. This work was supported by grants from the CICYT–Spain, Principado de Asturias, Fundación la Caixa, and European Union (Cancer Degradome-FP6 and -FP5). The Instituto Universitario de Oncología is supported by Obra Social Cajastur-Asturias and Red de Centros de Cancer-Instituto Carlos III, Spain. X.S.P. belongs to the Ramón y Cajal Program of the MCyT (Spain).

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi: 10.1016/j.ygeno.2005.07.009.

## References

- [1] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al., Initial sequencing and analysis of the human genome, *Nature* 409 (2001) 860–921.
- [2] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al., The sequence of the human genome, *Science* 291 (2001) 1304–1351.
- [3] The Chimpanzee Sequencing and Analysis Consortium, Initial sequence of the chimpanzee genome and comparison with the human genome, *Nature* 437 (2005) 69–87.
- [4] M.V. Olson, A. Varki, Sequencing the chimpanzee genome: insights into human evolution and disease, *Nat. Rev. Genet.* 4 (2003) 20–28.
- [5] S. Caenepeel, G. Charyczak, S. Sudarsanam, T. Hunter, G. Manning, The mouse kinome: discovery and comparative genomics of all mouse protein kinases, *Proc. Natl. Acad. Sci. USA* 101 (2004) 11707–11712.
- [6] X.S. Puente, L.M. Sánchez, C.M. Overall, C. López-Otín, Human and mouse proteases: a comparative genomic approach, *Nat. Rev. Genet.* 4 (2003) 544–558.
- [7] T. Angata, E.H. Margulies, E.D. Green, A. Varki, Large-scale sequencing of the CD33-related Siglec gene cluster in five mammalian species reveals rapid evolution by multiple mechanisms, *Proc. Natl. Acad. Sci. USA* (2004).
- [8] Z. Zhang, P.E. Burch, A.J. Cooney, R.B. Lanz, F.A. Pereira, J. Wu, R.A. Gibbs, G. Weinstock, D.A. Wheeler, Genomic analysis of the nuclear receptor family: new insights into structure, regulation, and evolution from the rat genome, *Genome Res.* 14 (2004) 580–590.
- [9] R.D. Emes, S.A. Beatson, C.P. Ponting, L. Goodstadt, Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents, *Genome Res.* 14 (2004) 591–602.
- [10] A. Fischer, Y. Gilad, O. Man, S. Paabo, Evolution of bitter taste receptors in humans and apes, *Mol. Biol. Evol.* 22 (2005) 432–436.
- [11] G. Suske, E. Bruford, S. Philipsen, Mammalian SP/KLF transcription factors: bring in the family, *Genomics* 85 (2005) 551–556.

- [12] C. López-Otín, C.M. Overall, Protease degradomics: a new challenge for proteomics, *Nat. Rev. Mol. Cell. Biol.* 3 (2002) 509–519.
- [13] C.M. Overall, C. López-Otín, Strategies for MMP inhibition in cancer: innovations for the post-trial era, *Nat. Rev. Cancer* 2 (2002) 657–672.
- [14] R. Vassar, Beta-secretase (BACE) as a drug target for Alzheimer's disease, *Adv. Drug Delivery Rev.* 54 (2002) 1589–1602.
- [15] W.P. Esler, M.S. Wolfe, A portrait of Alzheimer secretases—New features and familiar faces, *Science* 293 (2001) 1449–1454.
- [16] Y. Liu, L. Fallon, H.A. Lashuel, Z. Liu, P.T. Lansbury Jr., The UCH-L1 gene encodes two opposing enzymatic activities that affect alpha-synuclein degradation and Parkinson's disease susceptibility, *Cell* 111 (2002) 209–218.
- [17] E. Petek, C. Windpassinger, J.B. Vincent, J. Cheung, A.P. Boright, S.W. Scherer, P.M. Kroschel, K. Wagner, Disruption of a novel gene (IMMP2L) by a breakpoint in 7q31 associated with Tourette syndrome, *Am. J. Hum. Genet.* 68 (2001) 848–858.
- [18] W. Enard, P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, et al., Intra- and interspecific variation in primate gene expression patterns, *Science* 296 (2002) 340–343.
- [19] P. Khaitovich, B. Muetzel, X. She, M. Lachmann, I. Hellmann, J. Dietzsch, S. Steigele, H.H. Do, G. Weiss, W. Enard, et al., Regional patterns of gene expression in human and chimpanzee brains, *Genome Res.* 14 (2004) 1462–1473.
- [20] X.S. Puente, C. López-Otín, A genomic analysis of rat proteases and protease inhibitors, *Genome Res.* 14 (2004) 609–622.
- [21] A. Fortna, Y. Kim, E. MacLaren, K. Marshall, G. Hahn, L. Meltesen, M. Brenton, R. Hink, S. Burgers, T. Hernandez-Boussard, et al., Lineage-specific gene duplication and loss in human and great ape evolution, *PLoS Biol.* 2 (2004) E207.
- [22] G. Velasco, A.M. Pendás, A. Fueyo, V. Knäuper, G. Murphy, C. López-Otín, Cloning and characterization of human MMP-23, a new matrix metalloproteinase predominantly expressed in reproductive tissues and lacking conserved domains in other family members, *J. Biol. Chem.* 274 (1999) 4570–4576.
- [23] R. Gururajan, J.M. Lahti, J. Grenet, J. Easton, I. Gruber, P.F. Ambros, V.J. Kidd, Duplication of a genomic region containing the Cdc2L1-2 and MMP21-22 genes on human chromosome 1p36.3 and their linkage to D1Z2, *Genome Res.* 8 (1998) 929–939.
- [24] C. Chen, A.L. Darrow, J.S. Qi, M.R. D'Andrea, P. Andrade-Gordon, A novel serine protease predominately expressed in macrophages, *Biochem. J.* 374 (2003) 97–107.
- [25] N.C. Cheng, N. Van Roy, A. Chan, M. Beitsma, A. Westerveld, F. Speleman, R. Versteeg, Deletion mapping in neuroblastoma cell lines suggests two distinct tumor suppressor genes in the 1p35–36 region, only one of which is associated with N-myc amplification, *Oncogene* 10 (1995) 291–297.
- [26] B.J. Dave, D.L. Pickering, M.M. Hess, D.D. Weisenburger, J.O. Armitage, W.G. Sanger, Deletion of cell division cycle 2-like 1 gene locus on 1p36 in non-Hodgkin lymphoma, *Cancer Genet. Cytogenet.* 108 (1999) 120–126.
- [27] I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, R. Bruskiewich, D.M. Beare, M. Clamp, L.J. Smink, et al., The DNA sequence of human chromosome 22, *Nature* 402 (1999) 489–495.
- [28] P.G. Buckley, K.K. Mantripragada, M. Benetkiewicz, I. Tapia-Paez, T. Diaz De Stahl, M. Rosenquist, H. Ali, C. Jarbo, C. De Bustos, C. Hirvela, et al., A full-coverage, high-resolution human chromosome 22 genomic microarray for clinical and research applications, *Hum. Mol. Genet.* 11 (2002) 3221–3229.
- [29] R.A. Gibbs, G.M. Weinstock, M.L. Metzker, D.M. Muzny, E.J. Sodergren, S. Scherer, G. Scott, D. Steffen, K.C. Worley, P.E. Burch, et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution, *Nature* 428 (2004) 493–521.
- [30] P.J. Tatnell, D.J. Powell, J. Hill, T.S. Smith, D.G. Tew, J. Kay, Napsins: new human aspartic proteinases: distinction between two closely related genes, *FEBS Lett.* 441 (1998) 43–48.
- [31] M. Cook, F. Buhling, S. Ansoorge, P.J. Tatnell, J. Kay, Pronapsin A and B gene expression in normal and malignant human lung and mononuclear blood cells, *Biochim. Biophys. Acta* 1577 (2002) 10–16.
- [32] B.P. Shum, L.R. Flodin, D.G. Muir, R. Rajalingam, S.I. Khakoo, S. Cleland, L.A. Guethlein, M. Uhrberg, P. Parham, Conservation and variation in human and common chimpanzee CD94 and NKG2 genes, *J. Immunol.* 168 (2002) 240–252.
- [33] D. Meyer-Olson, K.W. Brady, J.T. Blackard, T.M. Allen, S. Islam, N.H. Shoukry, K. Hartman, C.M. Walker, S.A. Kalams, Analysis of the TCR beta variable gene repertoire in chimpanzees: identification of functional homologs to human pseudogenes, *J. Immunol.* 170 (2003) 4161–4169.
- [34] E.J. Adams, S. Cooper, P. Parham, A novel, nonclassical MHC class I molecule specific to the common chimpanzee, *J. Immunol.* 167 (2001) 3858–3869.
- [35] T. Anzai, T. Shiina, N. Kimura, K. Yanagiya, S. Kohara, A. Shigenari, T. Yamagata, J.K. Kulski, T.K. Naruse, Y. Fujimori, et al., Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence, *Proc. Natl. Acad. Sci. USA* 100 (2003) 7708–7713.
- [36] H. Fischer, U. Koenig, L. Eckhart, E. Tschachler, Human caspase 12 has acquired deleterious mutations, *Biochem. Biophys. Res. Commun.* 293 (2002) 722–726.
- [37] M. Saleh, J.P. Vaillancourt, R.K. Graham, M. Huyck, S.M. Srinivasula, E.S. Alnemri, M.H. Steinberg, V. Nolan, C.T. Baldwin, R.S. Hotchkiss, et al., Differential modulation of endotoxin responsiveness by human caspase-12 polymorphisms, *Nature* 429 (2004) 75–79.
- [38] S.M. McEvoy, N. Maeda, Complex events in the evolution of the haptoglobin gene cluster in primates, *J. Biol. Chem.* 263 (1988) 15740–15747.
- [39] L.M. Erickson, H.S. Kim, N. Maeda, Junctions between genes in the haptoglobin gene cluster of primates, *Genomics* 14 (1992) 948–958.
- [40] A.B. Smith, J.D. Esko, S.L. Hajduk, Killing of trypanosomes by the human haptoglobin-related protein, *Science* 268 (1995) 284–286.
- [41] L. Vanhamme, F. Paturiaux-Hanocq, P. Poelvoorde, D.P. Nolan, L. Lins, J. Van Den Abbeele, A. Pays, P. Tebabi, H. Van Xong, A. Jacquet, et al., Apolipoprotein L-I is the trypanosome lytic factor of human serum, *Nature* 422 (2003) 83–87.
- [42] H.L. Paulson, N.M. Bonini, K.A. Roth, Polyglutamine disease and neuronal cell death, *Proc. Natl. Acad. Sci. USA* 97 (2000) 12957–12958.
- [43] M. Llamazares, S. Cal, V. Quesada, C. López-Otín, Identification and characterization of ADAMTS-20 defines a novel subfamily of metalloproteinases—disintegrins with multiple thrombospondin-1 repeats and a unique GON domain, *J. Biol. Chem.* 278 (2003) 13382–13389.
- [44] R. Pfitzer, E. Myers, S. Applebaum-Shapiro, R. Finch, I. Ellis, J. Neoptolemos, J.A. Kant, D.C. Whitcomb, Novel cationic trypsinogen (PRSS1) N29T and R122C mutations cause autosomal dominant hereditary pancreatitis, *Gut* 50 (2002) 271–272.
- [45] M. Balbín, A. Fueyo, A.M. Tester, A.M. Pendás, A.S. Pitiot, A. Astudillo, C.M. Overall, S.D. Shapiro, C. López-Otín, Loss of collagenase-2 confers increased skin tumor susceptibility to male mice, *Nat. Genet.* 35 (2003) 252–257.
- [46] L.M. Coussens, C.L. Tinkle, D. Hanahan, Z. Werb, MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis, *Cell* 103 (2000) 481–490.
- [47] L.M. Coussens, Z. Werb, Inflammation and cancer, *Nature* 420 (2002) 860–867.
- [48] D.S. Beniashvili, An overview of the world literature on spontaneous tumors in nonhuman primates, *J. Med. Primatol.* 18 (1989) 423–437.
- [49] A. Varki, A chimpanzee genome project is a biomedical imperative, *Genome Res.* 10 (2000) 1065–1070.
- [50] M.I. Jensen-Seaman, W.H. Li, Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen, *J. Mol. Evol.* 57 (2003) 261–270.
- [51] T. Nakagawa, H. Zhu, N. Morishima, E. Li, J. Xu, B.A. Yankner, J. Yuan, Caspase-12 mediates endoplasmic-reticulum-specific apoptosis and cytotoxicity by amyloid-beta, *Nature* 403 (2000) 98–103.
- [52] C. Rao, D. Foerzler, S.K. Loftus, S. Liu, J.D. McPherson, K.A. Jungers, S.S. Apte, W.J. Pavan, D.R. Beier, A defect in a novel ADAMTS family

- member is the cause of the belted white-spotting mutation, *Development* 130 (2003) 4665–4672.
- [53] R. Brelloch, S.S. Anna-Arriola, D. Gao, Y. Li, J. Hodgkin, J. Kimble, The gon-1 gene is required for gonadal morphogenesis in *Caenorhabditis elegans*, *Dev. Biol.* 216 (1999) 382–393.
- [54] W.C. Parks, C.L. Wilson, Y.S. Lopez-Boado, Matrix metalloproteinases as modulators of inflammation and innate immunity, *Nat. Rev. Immunol.* 4 (2004) 617–629.
- [55] H. Tsukada, T. Pourmotabbed, Unexpected crucial role of residue 272 in substrate specificity of fibroblast collagenase, *J. Biol. Chem.* 277 (2002) 27378–27384.
- [56] B.P. Lawrence, Y. Will, D.J. Reed, N.I. Kerkvliet, Gamma-glutamyltranspeptidase knockout mice as a model for understanding the consequences of diminished glutathione on T cell-dependent immune responses, *Eur. J. Immunol.* 30 (2000) 1902–1910.
- [57] W.J. Kent, BLAT—The BLAST-like alignment tool, *Genome Res.* 12 (2002) 656–664.
- [58] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (1997) 3389–3402.
- [59] J.M. Comeron, K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals, *Bioinformatics* 15 (1999) 763–764.