A Genomic Analysis of Rat Proteases and Protease Inhibitors

Xose S. Puente and Carlos López-Otín[1]

Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, 33006-Oviedo, Spain

Proteases perform important roles in multiple biological and pathological processes. The availability of the rat genome sequence has facilitated the analysis of the complete protease repertoire or degradome of this model organism. The rat degradome consists of at least 626 proteases and homologs, which are distributed into 24 aspartic, 160 cysteine, 192 metallo, 221 serine, and 29 threonine proteases. This distribution is similar to that of the mouse degradome but is more complex than that of the human degradome composed of 561 proteases and homologs. This increased complexity of rat proteases mainly derives from the expansion of several families, including placental cathepsins, testases, kallikreins, and hematopoietic serine proteases, involved in reproductive or immunological functions. These protease families have also evolved differently in rat and mouse and may contribute to explain some functional differences between these closely related species. Likewise, genomic analysis of rat protease inhibitors has shown some differences with mouse protease inhibitors and the expansion of families of cysteine and serine protease inhibitors in rodents with respect to human. These comparative analyses may provide new views on the functional diversity of proteases and inhibitors and contribute to the development of innovative strategies for treating proteolysis diseases.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to EMBL under accession nos. BN000318–BN000390.]

Proteolytic enzymes comprise a group of structurally and functionally diverse proteins that have the common ability to catalyze the hydrolysis of peptide bonds (Barrett et al. 1998). Although these enzymes were originally studied as the central executioners of nonspecific protein catabolism, our view of proteases has considerably expanded after the recognition of their participation in the catalysis of specific reactions of proteolytic processing (Neurath 1999). The highly selective and limited cleavage of specific substrates mediated by proteases is essential in every cell and organism. In fact, a number of important processes that regulate the activity and fate of many proteins are strictly dependent on proteolytic processing events. These include the ectodomain shedding of cell surface proteins; the appropriate intra- or extracellular localization of multiple proteins; the activation and inactivation of cytokines, hormones and growth factors; the regulation of transcription factor activity; or the exposure of cryptic neoproteins with functional roles distinct from the parent molecule from which they derive after proteolytic cleavage reactions (López-Otín and Overall 2002). These protease-mediated processing events, which are distinct from nonspecific protein degradation reactions, are vital in the control of essential biological processes such as DNA replication, cell-cycle progression, cell proliferation, differentiation and migration, morphogenesis and tissue remodeling, immunological reactions, ovulation, fertilization, neuronal outgrowth, angiogenesis, hemostasis, and apoptosis. Consistent with the biological relevance of proteases in the control of multiple biological processes, deficiencies or alterations in the regulation of these enzymes underlie important human diseases such as arthritis, cancer, and neurodegenerative and cardiovascular diseases (Hooper 2002). Most human diseases of proteolysis are the result of alterations in the spatiotemporal patterns of expression of proteases. Nevertheless, we have also recently cataloged >50 hereditary disorders that are caused by loss-of-function mutations in protease genes (Puente et al. 2003). Furthermore, it is remarkable that many infectious microorganisms, viruses, and parasites use proteases as virulence factors, thereby being of great interest for the pharmaceutical industry as potential drug targets (Shao et al. 2002; Anand et al. 2003; Imamura 2003; Wu et al. 2003).

Because of the essential functional roles of proteases in the control of cell behavior, survival, and death, together with their increasing relevance as therapeutic targets, there is a growing interest in the identification and functional characterization of the complete protease repertoire of living organisms. The virtual completion of several large-scale genome sequencing programs has made possible this kind of global analyses aimed at characterizing degradomes: the complete set of proteases produced by a cell, tissue, or organism (López-Otín and Overall 2002). Recently, we have performed the first genomic analysis of the human and mouse degradomes (Puente et al. 2003). Similar analyses have been performed for the study of proteases of *Plasmodium falciparum* (Wu et al. 2003). Preliminary data are also available for proteases present in other model organisms such as *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* (http://merops.sanger.ac.uk). These studies have provided new opportunities to appreciate the complexity of proteolytic systems. We have annotated 561 proteases and protease-homologs encoded in the human genome, whereas somewhat surprisingly, the mouse degradome is much more complex, being composed of 641 components (Puente et al. 2003) (http://web.uniovi.es/degradome). It is also remarkable that *Drosophila*, despite having a gene content much lower than that of humans, shows a similar number of protease genes owing to the expansion of a family of trypsin-like serine proteases in the fly genome (Ross et al. 2003). To date, 403 proteases and homologs have been annotated in *C.*

[1]**Corresponding author.**
E-MAIL CLO@correo.uniovi.es; FAX 34-985-103564.

Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1946304.

14:609–622 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

elegans and 609 in *Arabidopsis*, although these analyses are far from being complete and the functionality of most predicted proteases in these organisms has not yet been validated at the biochemical level. A comparative evaluation of available data has confirmed that many families of human and mouse proteases are recognizable in the genomes of all model organisms, confirming the existence of conserved proteolytic routines in all cases. However, beyond these universal protease-mediated functions, there are also specific functions that are carried out by unique proteases in different species, making necessary to clarify the genetic and molecular basis underlying the evolutionary differences between the complete protease sets of these organisms.

The recent availability of the rat euchromatic genome sequence (Rat Genome Sequencing Project Consortium 2004) has prompted us to extend our comparative analysis of degradomes to this animal model whose study has been decisive in our current understanding of multiple physiological and pathological processes (Jacob and Kwitek 2002). In this work, we perform a genomic analysis of the rat proteases, with the finding of a total of 626 proteases and homologs in this organism. We also perform a comparative analysis of the rat degradome with those of mouse and human, as well as a preliminary evaluation of the protease inhibitor content of these species. Finally, we discuss the potential relevance of these studies to define distinctive aspects of the rat, mouse, and human biology from a protease perspective.

RESULTS

Genomic Analysis of the Rat Degradome

The rat genome sequence (assembly 3.1) was searched for the presence of proteases by using TBLASTN and BLAT at the Ensembl and University of California at Santa Cruz (UCSC) genome browsers, respectively, and by using as queries the protease sequences derived from our previous analysis of human and mouse degradomes (Puente et al. 2003). We also used InterPro annotations of the rat genome as well as experimental information generated in our laboratory to identify putative new members of known protease families. The combined utilization of these methods led us finally to annotate a total of 626 genes encoding proteases or protease-homologs in the rat genome (Table 1). A total of 529 of the annotated proteins are classified as bona fide proteases possessing all structural requirements for catalytic activity, whereas 97 of them are suggested to be inactive protease homologs owing to the occurrence of changes in specific residues important for the catalytic properties of the different classes of proteases. These inactive protease homologs have been proposed to play important roles as regulatory or inhibitory molecules by titrating inhibitors from the milieu, thereby increasing net proteolytic activity, or by acting as dominant negatives with ability to bind substrates through the inactive catalytic or exosite ancillary domains (López-Otín and Overall 2002). In addition to the 626 annotated genes encoding proteases and homologs, we have also identified >150 protease pseudogenes in the rat genome. These predicted rat pseudogenes, derived by retrotransposition or by duplication and subsequent accumulation of frameshifts and stop codons, have not been further analyzed in this work, with the exception of those representing differences with putative mouse and human orthologous genes. Likewise, the rat genome also contains multiple aspartic protease-related sequences embedded within endogenous retroviral elements, but they have not been included in the present catalog of rat proteases. The annotated rat degradome includes >200 putative proteases and homologs absent in the last release of MEROPS database (6.40; September 24, 2003), the best resource for analysis of proteases from multiple organisms (http://merops.sanger.ac.uk). These differences are likely because we are using the nonannotated rat genome assembly as a tool for the identification of new proteases, whereas MEROPS and other public databases mainly use sequences derived from primary databases for the annotation of the protease complement in the analyzed organisms.

We next analyzed the distribution of the 626 annotated rat proteases and homologs within the different catalytic classes of proteolytic enzymes. It is well established that according to the mechanism of catalysis, proteases may be distributed into five distinct classes: aspartic, metallo, cysteine, serine, and threonine proteases (Barrett et al. 1998). The results derived from analyzing the distribution of rat proteases in these catalytic classes are shown in Figure 1 and Table 1. Serine proteases are the most abundant proteolytic enzymes in rat, with 221 members, whereas metallo and cysteine proteases also contain multiple members (192 and 160, respectively). In contrast, there are only 24 aspartic and 29 threonine proteases in rat, likely reflecting the highly specialized roles played by these enzymes. The distribution of rat proteases in catalytic classes is very similar to that previously reported for mouse proteases but differs from that corresponding to human proteases in both total number and relative distribution among classes (Table 1; Puente et al. 2003).

To carry out a more detailed evaluation of possible differences between rat and human and mouse proteases, we next

Table 1. Distribution of Proteases in Rat, Mouse, and Human Genomes

	Catalytic class					
	Total	Asp	Cys	Metallo	Ser	Thr
Rat	626	24	160	192	221	29
Mouse	641	27	163	198	227	26
Human	561	21	148	186	178	28
Rat/human orthologs	524	21	135	177	165	26
Rat/mouse orthologs	583	24	151	191	191	26
Rat specific	42	0	9	1	29	3
Rodent specific	62	3	16	15	28	0
Human specific	31	0	11	9	9	2
Human gene/rat pseudogene	9	0	3	3	3	0
Rat gene/human pseudogene	21	1	1	7	12	0
Mouse gene/rat pseudogene	12	1	2	4	5	0
Rat gene/mouse pseudogene	1	0	1	0	0	0

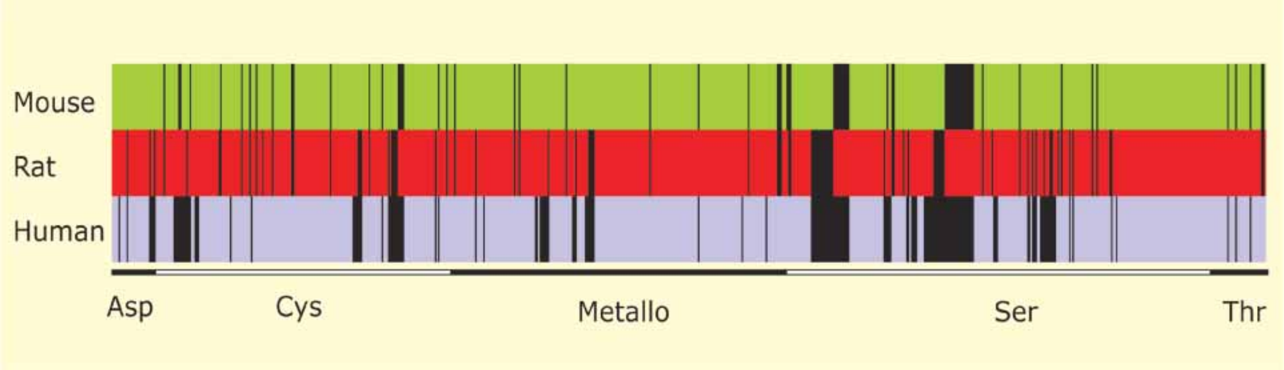

Figure 1 A global view of the rat degradome and comparison with those of mouse and human. The figure represents the complete set of proteases and protease–homologs from each species distributed into five catalytic classes. Proteases absent in one species are shown as black bars.

classified the rat proteases into families following the MEROPS criteria, and then performed a comparative analysis of members of each family with those annotated in the human and mouse genomes. According to this analysis, we have concluded that the rat proteases belong to 67 different families, the largest one being the S01 of serine proteinases, containing 160 members. Other families with many representatives are the C19 family of cysteine proteases and the M12 family of metalloproteases, possessing 54 and 56 components, respectively. In contrast, there are several families, such as C15, C26, C50, C56, C67, Cx1, M08, M18, M47, M49, M50, Mx1, S12, S14, S53, S59, S60, and Sx1, with a single member in the rat genome (see Supplemental Table 1 available online at www.genome.org). As expected, the comparative analysis between rat and mouse degradomes showed a very high percentage (93%) of rat genes with a strict ortholog in the mouse genome. However, we could not find a bona fide mouse ortholog for 43 rat genes. Likewise, there are 58 mouse genes lacking a rat counterpart, although there are some few cases in which the absence of rat or mouse specific genes might derive from gaps still occurring in the assemblies of both genomes. On the other hand, comparative analysis between rat and human degradomes has shown a total of 524 genes with a recognizable ortholog in the human genome. There are 102 protease genes specific for rat compared with human and 37 genes specific for human compared with rat. In most cases, differences between rat, mouse, and human degradomes derive from differential expansion of genes encoding related members of protease families present in the genome of the three species. Nevertheless, there are also interesting examples of creation of specific protease subfamilies in rat and mouse with no counterparts in human, as well as some cases of specific losses or inactivation of protease genes in one of the lineages and preferentially in human (Tables 1, 2). A detailed analysis of each protease family annotated in the rat genome and its comparison with those present in mouse and human is described below.

Aspartic Proteases

We have annotated a total of 24 rat aspartic proteases and protease-homologs divided into four families: A01, A02, A22, and Ax1 (Supplemental Table 1). The family A01 is composed of nine rat proteases, including chymosin, a digestive protease that is inactivated by mutations and frameshifts in the human genome (Kolmer et al. 1991). This fact represents an important difference between human and rodent proteases and could contribute to explain differences in the physiology of digestion between these species. In contrast, we have not found evidence of the presence in rat of the *Ren-2* gene encoding the submandibular renin. This gene is also absent in the human genome but is present in many strains of mice (Abel and Gross 1990), thereby representing a mouse-specific gene. On the other hand, rat pepsinogen F—a fetal aspartic protease—is closely related to mouse pepsinogen F (94% identities), but both are very distant to the diverse human pepsinogen A isozymes. These human gastric enzymes are encoded by highly related genes syntenic to rodent pepsinogens F but are very divergent from them in structure, regulation, and function (Kageyama 2002).

The A02 family of rat aspartic proteases contains five members—Ddi1, Ddi2, Ddi-rp, Nrip2, and Nrip3—absolutely conserved in mouse and human (Supplemental Table 1). These genes encode enzymes with some similarity to retroviral aspartic proteases, although they are not embedded within endogenous retroviral elements. Moreover, there are seven members of the A22 family of presenilins, which are also conserved in both mouse and human genomes. The repertoire of rat aspartic proteases is completed with a new family whose members have sequences similar to that of the PIP protein (prolactin inducible protein), which has been recently characterized as an aspartic protease (Caputo et al. 2000). The two rat PIP-related proteins lack residues proposed to be essential for PIP proteolytic activity and have been classified as nonprotease homologs. The mouse genome encodes two additional members of the PIP family called Sva and Sval3 proteins. In contrast, the human genome only contains a single copy *PIP* gene—although highly divergent of rat and mouse PIP—and lacks all the additional PIP-related genes found in rodents and expressed in male reproductive organs (Yoshida et al. 2001). Accordingly, the PIP family represents an example of gene family expansion in the rat and mouse degradomes.

Cysteine Proteases

We have annotated 160 rat cysteine proteases belonging to 18 different families (Supplemental Table 1). The catalog of rat cysteine proteases includes the three members of the hedgehog protein family, whose protease function is exclusively used for the autoproteolytic processing of their respective precursors (Perler 1998). We have also included in this list the rat ortholog of the human protein DJ-1, which is mutated in some forms of Parkinson's disease and has been suggested to have a functional role as cysteine protease. Nevertheless, the recent resolution of DJ-1 crystal structure has raised doubts about this function (Wilson et al. 2003), and consequently, we have classified this protein as a nonprotease homolog belonging to the class of cysteine proteases.

According to our genomic analysis and similar to the case of mouse, the C01 family of cysteine proteases is largely expanded in the rat genome due to the presence of two protease subfamilies—placental cathepsins and testins—that are absent in the hu-

Table 2. **Classification of Rat-Specific Genes Whose Human Ortholog Has Been Inactivated by Mutation or is Absent**

Process/protein	Gene	Rat locus	Human ortholog	Human locus
Digestion				
chymosin	Cymp	2q34	Inactive	1p13
distal intestinal serine protease	Disp	10q12	Inactive	16p13
trypsin 10	Try10	4q23	Inactive	7q34
trypsin 15	Try15	4q23	Inactive	7q34
pancreatic elastase	Ela1	7q35	Inactive	12q13
Reproduction				
fertilin-α	Adam-1a	12q16	Inactive	12q24
fertilin-α-b	Adam-1b	12q16	Absent	—
cyritestin	Adam-3b	16q12	Inactive	16q12
ADAM4	Adam-4	6q24	Inactive	14q24
ADAM4B	Adam-4b	6q24	Inactive	14q24
ADAM5	Adam-5	16q12	Inactive	8p11
ADAM6	Adam-6	6q32	Inactive	14q24
testases 1, 4, 5	Adam-24, -34, -35	16q12	Absent	—
testase 2	Adam-25	16q12	Inactive	8p22
testis serine protease 3	Tessp3	8q32	Inactive	3p21
testis serine protease 6	Tessp6	8q32	Inactive	3p21
testicular serine protease 1	Tesp1	9q13	Absent	—
testicular serine protease 2	Tesp2	9q13	Inactive	2q21
testicular serine protease 3	Tesp3	17p14	Inactive	9q22
placental cathepsins (10 genes)	Ctsj,m,q,q2,q21,r,1,2,6,7l	17p14	Absent	—
testins (3 genes)	Cmb22,23,24	17p14	Absent	—
collagenase-like B	Mcolb	8q11	Absent	—
implantation serine protease 2	Isp2	10q12	Absent	—
implantation serine protease 2L	Isp2l	10q12	Inactive	16p13
Ppnx	Ppnx	Xq14	Inactive	Yp11
glandular kallikreins (10 genes)	rGk1-3,rGk7-12,K-32	1q21	Absent	—
Host defense				
mast cell chymases (16 genes)	Mcpt1,1l1,1l2,1l3,1l4,2,3,4,4l1, 8,8l1,8l2,8l3,9,rMcpt4,Rmcp1	15p13	Absent	—
granzymes (9 genes)	Gzmbl1, bl2, cl1, cl2, cl3, n, o, Rnkp7, 7l	15p13	Absent	—
airway trypsin-like (3 genes)	Hatl1-3	14p21	Inactive	4q13

Protease genes are grouped according to their putative participation in three main biological processes.

man lineage (Fig. 2). The placental cathepsins provide an important example of local gene expansion taking place in the rat and mouse genomes, although there are some differences in the evolution of this family in both rodents. We have annotated 10 putative placental cathepsins in rat chromosome 17p14, but none of them is present in human. The mouse genome contains eight placental cathepsin genes located at chromosome 13B3 (Deussing et al. 2002; Sol-Church et al. 2002), seven of them being true orthologs of rat placental cathepsins. The rat specific genes should be *Ctsq2, Ctsq2l,* and *Cts7l,* whereas *Cts3* should be a mouse-specific gene. The testin subfamily of the C01 family of cysteine proteases provides another example of gene family showing differences between rat and human. We have identified three testins in the rat genome, located in close proximity to the placental cathepsins at 17p14. The mouse genome also contains three testin genes orthologous to those identified in rat, but none of them is present in the human lineage, indicating that testins are a rodent-specific subfamily of cysteine proteinases. It is remarkable that both placental cathepsins and testins are proteases associated with reproductive functions, providing additional evidence that most differences in protease genes underlie changes in reproductive strategies between the analyzed species. Analysis of the rat genome has also confirmed the presence of a single copy of cathepsin L–like genes in rodents, whereas there are two functional human cathepsin L–like genes (*CTSL* and *CTSL2*) at 9q21, the second one being associated with reproductive and immunological functions. It is remarkable that the rat cathepsin L–like is more closely related to human CTSL2 than to human CTSL, confirming and extending data previously reported for mouse cathepsin L–like (Santamaría et al. 1998).

Analysis of the C02 family of cysteine proteinases—the calpains—has also revealed differences between rat, mouse, and human (Supplemental Table 1). First, and similar to mouse, the rat genome lacks calpain-14, indicating that this gene is human-specific compared with rodents. A second interesting difference is the finding that calpain-13 has been specifically inactivated by mutation in the rat genome. The C12 family of deubiquitinating enzymes contains five members in rat as in mouse, including Uchl4, which is absent in human. The rat caspase repertoire (C14 family) is identical to that of mouse but distinct from human. We have confirmed the absence of caspase-5 and caspase-10 in the rat genome, both being functional enzymes in human. Rat and mouse also lack a nonprotease homolog of caspases present in the human genome and called ICEY. In contrast, there is a functional gene for rat caspase-12 at 8q11, orthologous to the mouse caspase-12 gene at 9A1, whereas the human caspase-12 gene has acquired several deleterious mutations that abrogate its protease function (Fischer et al. 2002). Interestingly, the C15 family of pyroglutamyl-peptidases shows a unique feature of the rat genome: The rat pyroglutamyl-peptidase II gene at 1q22 has accumulated mutations and frameshifts that have prompted us to classify it as a pseudogene, whereas it is a functional gene in both human and mouse.

The C19 family of ubiquitin specific proteases (USPs) deserves a particular analysis due to its large size and extreme complexity. We have annotated a total of 54 rat USPs, many of them representing novel in silico predictions, although they are fully supported by EST-evidence and recent experimental data from our laboratory (V. Quesada and C. López-Otín, unpubl.). The catalog of rat USPs includes Usp4, Usp18, and Usp19 absent in the current genome as-

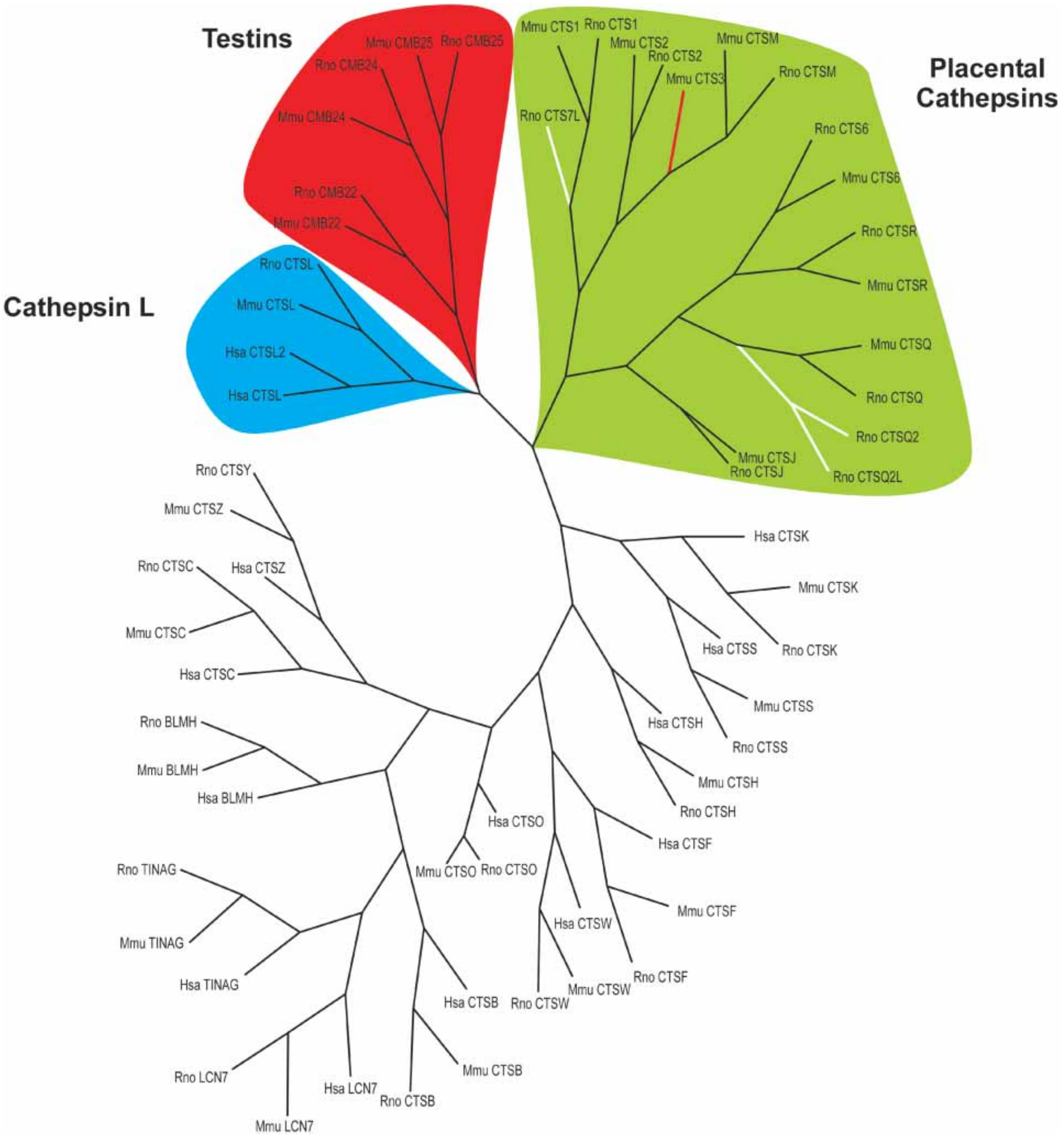

Figure 2 Unrooted phylogenetic tree of the C01 family of cysteine proteases. The testin subfamily of rodent-specific proteases is shown in red, and placental cathepsins, which are also rodent-specific, are shown in green. Branches corresponding to rat-specific members are shown in white; those mouse-specific, in red. Cathepsin L–like proteases, the only group containing one extra member in human, are shown in blue.

sembly but experimentally verified, as well as Usp9y—the only protease gene reported to date in chromosome Y—the sequence of which is not yet available in rat. Analysis of rat USPs shows some interesting differences with their human and mouse counterparts. First, we have confirmed the absence of USP6 (tre2) in the rat genome. This gene is also absent in mouse and has been recently characterized as a hominoid-specific gene (Paulding et al. 2003). Likewise, USP41 is also absent in rat and mouse but present in the human lineage. The rat genome, as previously described for mouse, also lacks true orthologs of human USP17 and USP-17-like sequences encoded within highly polymorphic and tandemly repeated intronless regions located at 4p15 and 8p23 (Okada et al. 2002). The closest relatives of *USP17* genes in the rat genome are those coding for proteins called DUBs (deubiquitinating enzymes), a group of cytokine-inducible cysteine proteases produced by lymphocytes (Baek et al. 2001). We have annotated three members of this subfamily of hematopoietic proteases in rat chromosome 1q32, whereas at least six members have been found in the mouse and none in human. Accordingly, the DUB subfamily of cysteine proteases provides an example of divergent evolution in rat, mouse, and human lineages. Lastly, it is remarkable that Usp26, located on the X chromosome, shows an extreme divergence with its mouse and human orthologs. In fact, the percentage of Usp26 identities in rat–mouse (77%) and rat–human (36%) pairs are among the lowest ones analyzed in this work. This observation may be explained by the fact that this gene is exclusively expressed in spermatogonia (Wang et al. 2001), likely playing a reproductive function and being subjected to strong selection pressures that lead to its rapid evolution (Swanson and Vacquier 2002).

Similar to the case of USPs, the C48 family of sentrin/SUMO-specific proteases (SENPs) also exhibits marked differences between these three species (Supplemental Table 1). We have annotated 13 SENP genes in the rat lineage, a number similar to that present in mouse (14) but considerably higher than the seven members described in human. Senp5-like, Senp16, Senp17, Senp18, and Senp19 are rat-specific proteases, whereas Senp9, Senp12, Senp13, Senp14, and Senp15 are mouse-specific proteases. Nevertheless, it is remarkable that these proteases are still poorly characterized at the biochemical level, and they should remain as bioinformatic predictions until their functional relevance is corroborated. The gene encoding the cylindromatosis protein (*CYLD1*) is conserved in rat and has been recently classified as the only member of the C67 family of cysteine proteases. Lastly, and in addition to these large-scale expansions in clusters of rat and mouse gene families, there are additional single copy genes encoding nonprotease homologs belonging to the cysteine protease class that are specifically inactivated in the rat and mouse genomes. These include the Gln-fructose-6-P transamidase 3 and several components of the recently described otubain family (Balakirev et al. 2003).

Metalloproteases

We have annotated 192 rat metalloproteases subdivided into 26 distinct families (Supplemental Table 1). The M01 family of aminopeptidases contains 12 members in rat absolutely conserved in the mouse genome. Both species lack a functional gene for aminopeptidase MAMS/L-RAP, a leukocyte-derived protease proposed to be associated with antigen processing. The M10 family of matrix metalloproteinases (MMPs)—one of the most relevant protease families in human pathology (Brinckerhoff and Matrisian 2002; Overall and López-Otín 2002)—shows some differences between rodents and humans. We have found a rat ortholog of mouse collagenase-like B, located at 8q11, that is absent in human. It seems that rat and mouse collagenase-like B are the result of a recent duplication event in rodents leading to the generation of a pair of genes (collagenase-like A and collagenase-like B) distantly related to human fibroblast collagenase and expressed in placenta (Balbín et al. 2001). Conversely, we have not found evidence of the presence of a gene coding for rat matrilysin-2 (*MMP-26*; Uría and López-Otín 2000). The gene is also absent in mouse and represents a human-specific gene compared with rodents. Likewise, we have only found a single copy of the *MMP-23* gene in the rat genome, a similar situation to that of mouse, reinforcing the proposal that the two *MMP-23*-like genes found in the human lineage are the result of a very recent duplication event (Gururajan et al. 1998; Velasco et al. 1999). Nevertheless, it must be pointed that this region is artifactually collapsed in the available assemblies of the human genome and is erroneously considered as containing a single gene. Therefore, the possibility that this region is duplicated in the rat and mouse genomes and has also been collapsed during the assembly can not be definitely ruled out.

The ADAM (a disintegrin and metalloproteinase) subfamily of M12 metalloproteinases shows important differences between rat, mouse, and human (Table 2; Supplemental Table 1). The genes for *ADAM-1, -3, -4, -5, -6,* and *-25* are pseudogenes in human and active in rat and mouse. *ADAM-1* is duplicated in both rat and mouse, *ADAM-6* only in mouse, and *ADAM-20* in human (*ADAM-20* and *ADAM-21*). In addition, there is a subgroup of ADAMs—called testases and located at rat chromosome 16q12—that are rodent specific. We have identified seven putative testase genes in the rat genome, although three of them (*Adam-26, Adam-36,* and *Adam-37*) have inactivating mutations and have been annotated as pseudogenes. The mouse genome contains nine members of this protease subfamily that are specifically expressed in testis, although they are intronless and their functional relevance remains to be demonstrated in most cases. It is remarkable that all these differences correspond to ADAMs expressed in reproductive tissues. In marked contrast with these species-specific differences in ADAMs, the genomic analysis of a group of ADAM-related metalloproteases known as ADAMTSs (ADAMs with thrombospondin domains) has revealed the absolute conservation in rat of the 19 distinct components previously identified in both human and mouse genomes (Llamazares et al. 2003).

Analysis of the M14 family of carboxypeptidases has shown that rat carboxypeptidase O has been inactivated by mutation, hence being annotated as a pseudogene. Mouse carboxypeptidase O is also a pseudogene, although the mutations leading to inactivation of this gene have been distinct in rat and mouse. The human carboxypeptidase O is functional and provides another example of human-specific gene as compared to rodents. Additional differences derived from the genomic analysis of rat metalloproteases include the inactivation in rat and mouse of a putative procollagenase III N-endopeptidase apparently functional in human (Scott et al. 1996), and the absence or inactivation in rodents of three nonprotease homologs belonging to the M67 family of metalloisopeptidases (Cope et al. 2002). Conversely, the Afg3-like protein 1—an ATP-dependent metalloprotease—is specifically inactivated in human but is apparently functional in the rat and mouse lineages. Last, it is of interest that the rat genome of an additional methionyl aminopeptidase located at chromosome 18p12, which is absent in human and mouse and could represent a rat-specific gene, was found.

Serine Proteases

The present analysis of the rat genome has allowed us to annotate 221 serine proteases, belonging to 16 families (Supplemental Table 1). Most of them are part of the densely populated S01 family, providing an explanation to the observation that all differences between rat serine proteases and their mouse and human counterparts derive from changes in members of this large protein family. Relevant differences are found in the kallikrein subfamily, which has been largely, albeit distinctly, expanded in the rat and mouse genomes with respect to the human lineage (Fig. 3). We have annotated 23 kallikreins in rat, whereas there are 26 in mouse and 15 in human. These kallikreins can be divided into two groups according to their relative location within two closely linked clusters present at chromosomes 1q21 in rat, 7B2 in mouse, and 19q13 in human (MacDonald et al. 1996; Yousef and Diamandis 2001; Olsson and Lundwall 2002). The first cluster contains 13 kallikrein genes in rat, 12 in mouse, and 15 in human. Among the human-specific kallikreins, the presence of KLK3 or PSA (prostate-specific antigen), an important biochemical marker in prostate cancer, is remarkable. This gene is inactivated in rat by a frameshift and appears to be absent in mouse. Likewise, the orthologs of human *KLK2* are inactivated in both rat and mouse genomes, whereas *KLK1* is active in human and rat but no mouse ortholog has been identified. Beyond these

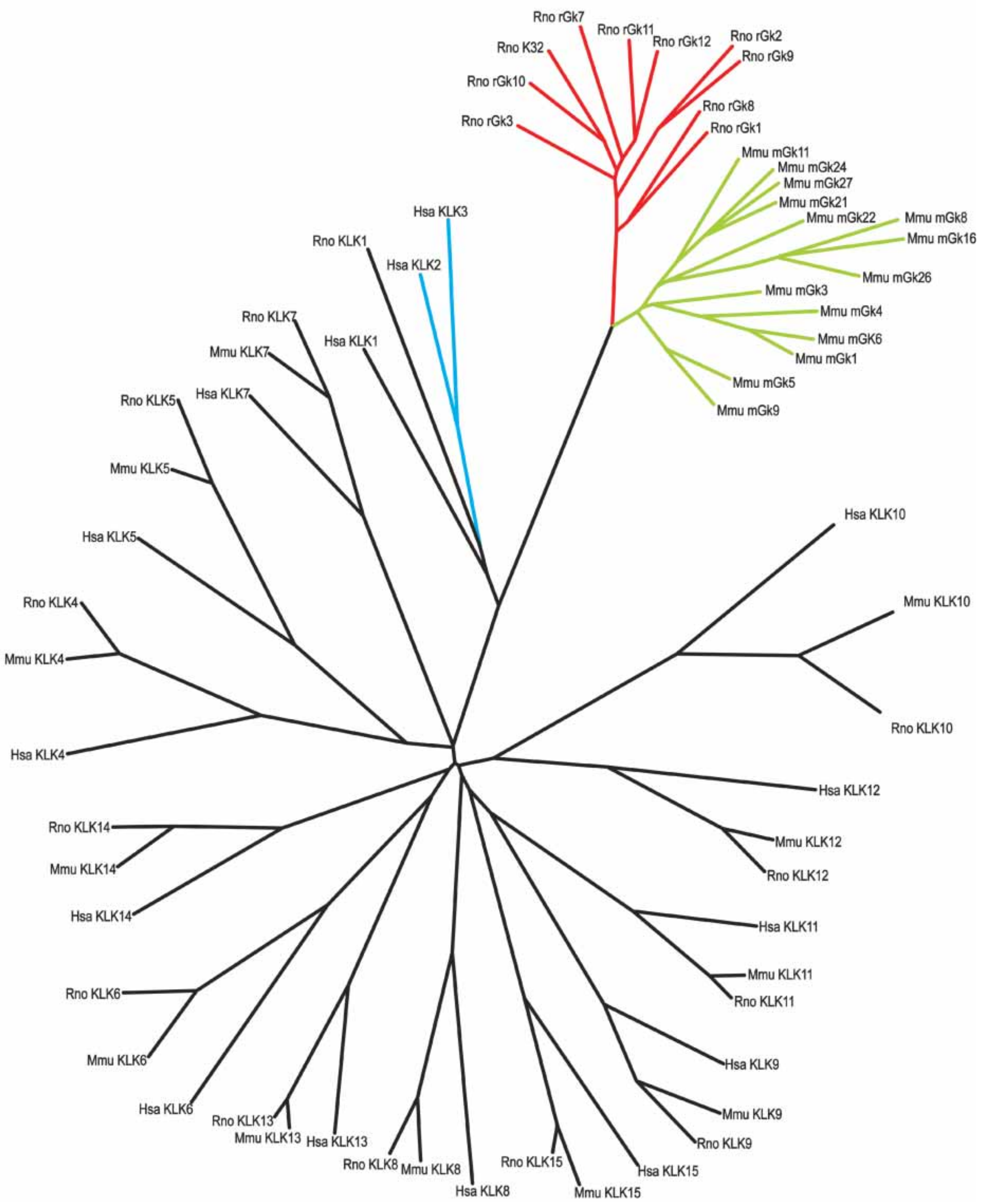

Figure 3 Phylogenetic tree of rat glandular kallikreins and comparison with those of mouse and human. The tree illustrates an independent expansion of rat and mouse glandular kallikreins. Rat-specific and mouse-specific glandular kallikreins are shown in red and in green, respectively. Human-specific kallikreins, including KLK2 and prostate specific antigen/KLK3, are shown in blue.

gene-specific changes, the differences in the second kallikrein cluster encoding glandular kallikreins are much more relevant. We have annotated 10 rat genes and 12 pseudogenes belonging to this subfamily, whereas none of them are present in the human genome. The mouse genome contains 14 genes and 12 pseudogenes at the corresponding 7B2 cluster. All these glandular kallikrein genes appear to have evolved independently in both rodents (Fig. 3). Therefore, this protease family represents an excellent example of differential evolutionary diversification in rat, mouse, and human lineages. The functional relevance of these serine proteases is largely unknown in most cases, although their expression is usually regulated by sex hormones, indicating their possible implication in reproductive processes (Yousef and Diamandis 2001).

There are also evident rat–human differences in hematopoietic serine proteases (Fig. 4). At least 24 mast cell chymase and granzyme-like genes located at 15p13 in the rat genome are absent in human. Some of them are conserved in mouse, but comparative analysis indicates poor conservation between both sets of rodent hematopoietic serine proteases. Because mast cell proteases may be involved in host defense—especially during bacterial infections—the expansion and differential evolution of this gene family in the rat and mouse genomes may have important consequences for the development of distinctive immune responses in rodents compared with humans (Lunderius and Hellman 2001). Several trypsins and HAT-like (<u>h</u>uman <u>a</u>irway <u>t</u>rypsin) proteases have also been specifically, albeit distinctly, expanded in rat and mouse with respect to human (Supplemental Table 1).

The genomic overview of rat serine proteases has also allowed us to find additional S01 family members that are functional in rat and mouse but inactive or absent in human (Table 2; Supplemental Table 1). Among them, we can mention the rat and mouse genes coding for implantation serine protease-2 (*ISP-2*); distal intestinal serine protease (*DISP-1*); and testis serine proteases *TESP-2, TESP-3, TESSP-3,* and *TESSP-6* classified as pseudogenes in human; as well as those coding for *DISP-2* and *TESP-1,* which are absent in the human genome. The absence or inactivation in human of all rodent ISP, DISP, and TESP serine proteases indicates that the functions of these enzymes—mainly associated with reproductive processes—are specific of rat and mouse compared with human. Conversely, an ovochymase-like protease, a form of pancreatic elastase, and tryptases-α and -δ1 are only present in human. The case of tryptase-α is interesting because although it is absent in rat and mouse, it has been also reported that human tryptase-α gene deficiency is common and affects ~29% of individuals surveyed (Soto et al. 2002). Three well-known nonprotease homologs—apolipoprotein (a), azurocidin, and haptoglobin-related protein—are absent in rat and mouse and present in human. In addition, and in contrast to the situation in the mouse genome, we have not found evidence of duplication of complement factors C1r and C1s in the rat genome. One set of these murine factors (*c1rA* and *c1sA*) are orthologs of the rat and human genes, whereas the others (*c1rB* and *c1sB*) are exclusively expressed in the male genital tract, indicating a role for these

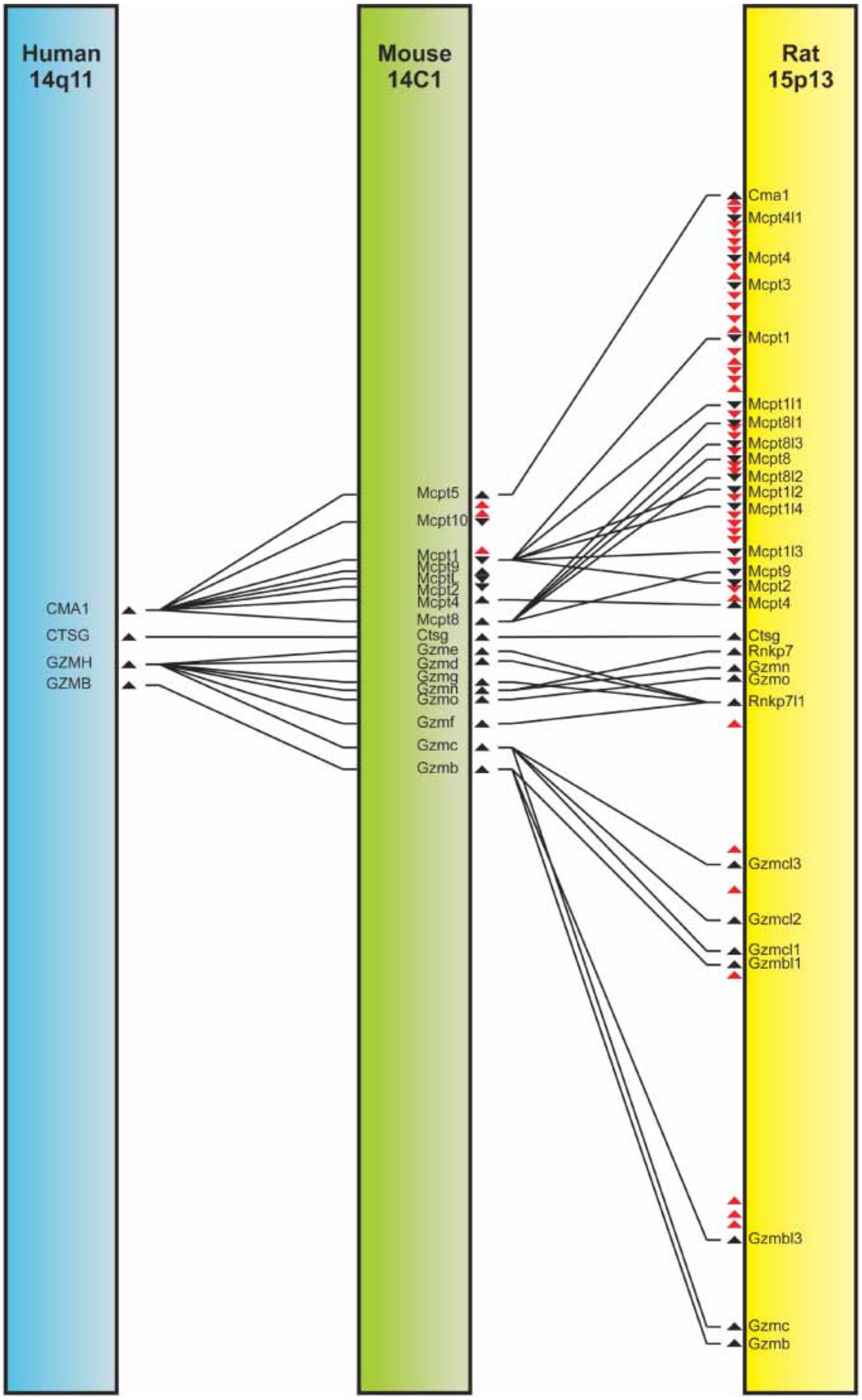

Figure 4 Comparison of rat hematopoietic serine proteases clustered at 15p13 with those of mouse at 14C1 and human at 14q11. Gene position and orientation are indicated by *arrowheads*; black ones represent genes, and red ones denote pseudogenes. Connecting lines indicate orthology or gene expansion.

proteases in reproduction independent of complement activation (Garnier et al. 2003). We have also found the presence in the rat genome of a nonprotease homolog called Ppnx, located in the X-inactivation center region (Xic) and strongly expressed in testis and undifferentiated ES cells (Chureau et al. 2002). There is no human ortholog in the corresponding Xic region, and a similar sequence located in human chromosome Yp11 has been classified as a pseudogene due to the accumulation of premature stop codons. A final and interesting example of differential serine proteases between rodents and human is illustrated by the *Ela1* gene encoding pancreatic elastase. This gene is active in rat and mouse but has been transcriptionally silenced in the human genome due to a mutation that inactivates crucial enhancer and promoter elements (Rose and MacDonald 1997).

Threonine Proteases

We have annotated 29 genes encoding threonine proteases in the rat genome (Supplemental Table 1), including 22 members of the T01 family of proteasome components, three components of the T02 family of glycosylasparaginases, and four members of the T03 family of γ-glutamyltransferases (GGTs), which catalyze the degradation of glutathione to glutamic acid and cysteinylglycine. We have identified three genes belonging to the T01 family that are rat specific, although they are intronless and may represent inactive pseudogenes. In contrast, the three rat genes of the T02 family are conserved in both human and mouse genomes. Lastly, there are some differences in the number of GGT genes present in rodent and humans. Two of them, *Ggtl3* and *Ggt6,* located at rat chromosomes 3q41 and 10q24 are orthologous of genes found in the mouse and human lineages. Likewise, we have annotated two rat GGT genes absent in the current genome assembly, which are also conserved in mouse. Nevertheless, the equivalent region of the human genome (22q11) has been very dynamic in its evolution, undergoing successive duplications that have given rise to four functional GGT genes and several pseudogenes in this genome region.

Analysis of Ancillary Domains Present in Rat Proteases

In addition to gene duplication-based mechanisms, proteases from all living organisms have also evolved through incorporation into their structures of a large variety of ancillary domains that facilitate their interaction with substrates, inhibitors, or receptors, or play some kind of regulatory role. Analysis of ancillary domains present in the annotated rat proteases has shown that ~50% of these enzymes are associated with at least one recognizable domain among those present in the Pfam database. Detailed analysis of these domains has revealed the presence of 61 distinct modules linked to the catalytic domains of rat proteases. Most of them are specifically associated with one or several protease families of a single catalytic class, but there are cases of ancillary domains shared by protease families belonging to two distinct catalytic classes. Thus, the UBQ domain is present in aspartic and cysteine proteases; the PA domain, in aspartic proteases and metalloproteases; the EF_HAND domain, in cysteine and serine proteases; and the CUB, MAM, EGF, EGF_CA, AAA, CCP, and FN2 domains, in metalloproteases and serine proteases.

According to the diversity and dynamism in ancillary domain accretion, it was tempting to speculate that rat proteases could have selectively accreted or lost specific domains when compared with human or mouse proteases, facilitating the evolution of distinct enzymes with ability to perform new functions. However, extending previous data from comparative analysis between human and mouse degradomes (Puente et al. 2003), we have not found evidence of changes in domain organization between pairs of orthologous rat–mouse or rat–human proteases. This observation is further supported after analysis of recently identified proteases with complex mosaic architectures such as polyserase-I, which contains a type II transmembrane motif, a LDLR module, and three tandem serine protease domains in a single polypeptide chain; or Adamts-20, with a metalloproteinase-, a cysteine-rich-, a disintegrin-, a GON-, and 15 TSP1-domains embedded in its amino acid sequence. These complex structural designs are absolutely conserved in the rat, mouse, and human orthologs of both proteases (Cal et al. 2003; Llamazares et al. 2003), indicating that the incorporation of these ancillary domains in primitive proteases predates the divergence between rat, mouse, and human lineages.

Distribution of Protease Genes in the Rat Genome

The protease genes are unevenly distributed in the rat genome. Some chromosomes such as 1, 8, and 15 are densely populated with protease genes. This fact may be explained, at least in part, by the occurrence of several protease clusters in these chromosomes. The largest cluster is located at 15p13 (mast cell chymase locus) and contains 28 functional genes and several pseudogenes. Another densely populated cluster maps at 1q21, in which a primordial serine protease gene duplicated repeatedly during evolution to give rise to 23 kallikrein-like genes and two related pseudogenes. Likewise, the matrix metalloproteinase (MMP) cluster at 8q11 contains 10 genes, including the rodent-specific genes encoding collagenase-like A and collagenase-like B. Despite these examples of gene families formed and expanded by local duplications, most protease gene families have been very dynamic in their evolution and, after duplication, the different paralogous genes have translocated to multiple chromosomes.

The genomic analysis of rat proteases might be also useful to further evaluate the completeness and accuracy of the current assembly of the rat genome. We have found that 13 previously known rat protease genes (*Ctsy, CtsO, Ctsj, Ctsq2, Autl4, Tnfaip3, Rnpepl1, Mep1b, Amsh2, Jamml2, Prss21, Ggt,* and *Ggtla1*)—the existence of which has been corroborated experimentally—are lost in the rat genome assembly used in our genomic analysis. Likewise, seven protease genes (*Usp4, Usp19, Clpp, Dpp9, C6–1a, Psmb1,* and *Aga*) are present in the rat genome sequence but not placed in the assembly, whereas there are other genes with structure that is not complete in the assembly. Some of these conflicts may derive from gaps still occurring in the rat genome assembly or from artifactual collapse of duplicated regions. Beyond these few differences, the annotation of the rat degradome provides additional evidence on the relative completeness and high quality of the current version of the rat genome sequence.

Genomic Analysis of Rat Protease Inhibitors

The above findings indicating that the rat degradome is considerably more complex than that of human prompted us to evaluate the possibility that the rat protease inhibitor complement could be also more complex as an attempt to control the increased protease repertoire in this organism (Kaiserman et al. 2002; Puente et al. 2003). To this purpose and because we have observed that most differences between rat and human degradomes are the result of the expansion of rat gene families clustered in certain genome regions, we mainly focused our analysis on identified clusters of rat protease inhibitor genes (Supplemental Table 2). We first examined the different clusters of serpins, a large family of serine protease inhibitors (Silverman et al. 2001; Barbour et al. 2002; Forsyth et al. 2003). The evolutionary history of this family has been of great interest because of marked changes in the number of serpin genes among and within diverse species, as well as the high rate of mutation in their reactive centers likely driven by positive Darwinian selection. There are three clusters of serpin genes in the rat genome, located at 6q32, 17p12, and 13p13. The first one encoding serpin A inhibitors is largely expanded in rat and contains 18 functional genes as opposed to the 10 SERPIN A genes located at 14q32 in the human genome. The expansion of this cluster in mouse is even larger, with a total of 29 serpin A genes at 12F1. These expansions mostly derive from the fact that genes corresponding to the single-copy human serpin α1-proteinase inhibitor (*SERPINA1*) or α1-chymotrypsin (*SERPINA3*) have undergone successive duplications in the rat and mouse genomes. Likewise, the cluster of serpin B genes located at rat chromosome 17p12 is also expanded in rat, with eight functional genes in this region and only three SERPIN B genes in human 6p25 (Fig. 5). The expansion of this cluster has also been more dynamic in the mouse genome, resulting in at least 15 functional members located at 13A4 (Fig. 5). Analysis of the third serpin cluster, located in rat chromosome 13p13 and containing 10 functional serpin B-like genes, did not reveal any differences with the number of orthologous genes present in human 18q21, whereas there are three additional serpin B3-like genes in mouse 1E1. In addition, a detailed analysis of one gene from this cluster encoding serpin B10 or bomapin, has shown an interesting example of differences between rat, human, and mouse genes. This gene is functional in rat and human but is inactivated in mouse, as assessed by the presence of a TAG stop codon at position 21 of the serpin B10 sequence available from both public and private versions of the mouse genome sequence. Nevertheless, we have observed that ESTs derived from tissues of CZECH II mice do not have a stop codon at this position, strongly indicating that the serpin B10 gene may be functional in this mouse strain. It is also of interest that SERPIN B10 has a restricted pattern of expression in human bone marrow cells, with a predominant nuclear localization in these cells, whereas the rat ortholog (also called trespin) has a wide tissue distribution and a cytosolic localization (Chipuk et al. 2002). Accordingly, serpin B10 represents an example of orthologous genes that appear to have acquired different functions in different species or even in different strains of the same species.

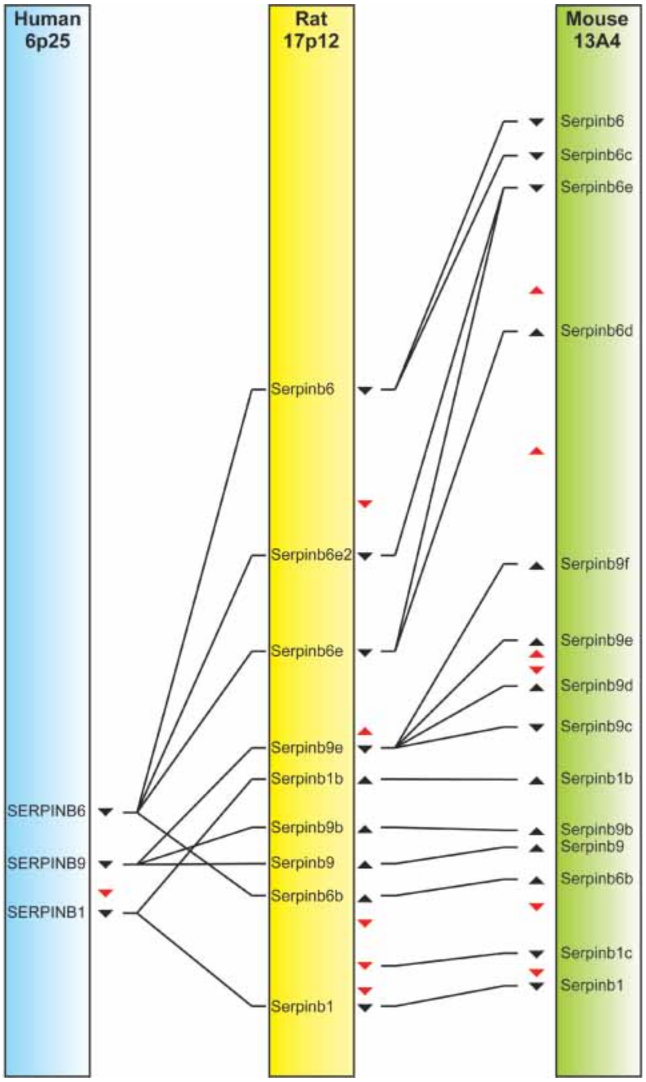

Figure 5 Comparative analysis of a cluster of rat serpins at 17p12 with the corresponding region of human and mouse genomes. The order and orientation of genes are indicated by black *arrowheads*, and pseudogenes are represented by red *arrowheads*. Orthology is indicated by connecting lines.

In addition to the observed differences in serpins, analysis of other families of protease inhibitors also revealed large expansions in rat and mouse versus human (Supplemental Table 2). For example, in the case of cysteine protease inhibitors clustered at rat chromosome 11q11, there are nine cystatin A-like genes in rat and mouse, whereas there is a single copy in the human genome. Likewise, there are four murinoglobulins—protease inhibitors of the α2-macroglobulin family—in rat and mouse but none in human. The family 2 cystatins predominantly expressed in the reproductive tract and clustered at rat 3q41 also exhibit differences with human cystatins located at 20p11 (Cornwall and Hsia 2003). Thus, rat cystatins CRES3, -T, -10, and -SC lack recognizable or functional orthologs in human, whereas there is one testatin-like inhibitor and four salivary gland cystatins present in human and absent in rat and mouse. Furthermore, a group of rat cystatin S–like inhibitors located also at 3q41 and absent in human has been largely expanded in the rat genome with 10 members in rat and only four in mouse. Rat calpastatin—the only endogenous inhibitor specific for calpains—is conserved in human and mouse, although it exhibits a species-specific pattern of alternative splicing (Takano et al. 1999). In contrast to these variations in serine and cysteine protease inhibitors, analysis of all endogenous inhibitors of rat metalloproteinases reported to date did not reveal any difference with those of mouse and human. Thus, the four tissue inhibitors of metalloproteinases (TIMPs), the RECK inhibitor of MMPs, and the tissue carboxypeptidase inhibitor (TCI) are conserved in all three species.

In summary, we have annotated 183 rat, 199 mouse, and 156 human protease inhibitors. This genomic analysis of protease inhibitors indicates that similar to the case of the rat and mouse degradomes, there is a marked expansion of several families of protease inhibitors in the genome of both rodent species with respect to that of human. Furthermore, we have also uncovered significant differences in the manner in which the analyzed protease inhibitor families have evolutionarily expanded in the genome of rat and mouse.

DISCUSSION

In this work, we have performed a genomic analysis of the rat degradome, the complete set of proteases present in the genome of this model organism that holds a special place in biomedical research. This genome-wide analysis has led us to conclude that the rat degradome is composed of a minimum of 626 proteases and homologs, distributed in 24 aspartic, 160 cysteine, 192 metallo, 221 serine, and 29 threonine proteases. These numbers are not definitive and could slightly change if new enzymes with unusual structures and catalytic mechanisms, which would have escaped to the predictive methods applied herein, are described in the future. We must also emphasize that many rat proteases annotated in this work are the result of bioinformatic predictions, and further experimental work will be required to confirm their functional relevance as proteolytic enzymes. The complexity of the rat degradome may be further increased through alternative splicing or differential polyadenylation events taking place in some of the identified protease genes, or by the occurrence in these genes of polymorphic variants that may contribute to modify protease functions or alter their regulatory mechanisms. The functional relevance of all these processes has been studied with some detail in several human protease genes, such as the angiotensin-converting enzyme (*ACE*) gene, in which alternative splicing events or use of alternative promoters generate isoforms with marked structural and functional differences (Riordan 2003). Illustrative examples of naturally occurring polymorphisms in human protease genes include those found in calpain-10 and *ADAM-33,* which confer increased susceptibility to complex diseases such as type 2 diabetes and asthma, respectively (Horikawa et al. 2000; Van Eerdewegh et al. 2002). Unfortunately, these additional sources of degradome variability remain largely unexplored in organisms distinct from human. The availability of the catalog of rat protease genes generated in this work may be helpful for the identification of polymorphic variants with ability to influence physiological or pathological functions mediated by these enzymes.

The comparative analysis of the rat degradome with those of mouse and human has also provided interesting insights into some protease-mediated processes that could contribute to explain differential aspects in the biology of these organisms. As expected, the rat protease repertoire is very similar to that of mouse, but we have also uncovered a number of differences. Most of these changes derive from the different expansion of several gene families, such as placental cathepsins, glandular kallikreins, and hematopoietic serine proteases, in both rodent species. Beyond these changes in differential evolution of clusters of gene families during the 12 to 24 million years since these organisms shared a common ancestor, we have also found some examples of specific loss or inactivation of genes in rat or mouse lineages. These include the one encoding a methionyl aminopeptidase–like protein present in rat and absent in mouse and those for renin-2, pyroglutamyl-peptidase II, and pancreatic endopeptidase E present in mouse and absent in rat. The fact that most of these differential genes between mouse and rat are associated with reproductive and host defense functions confirms and extends previous studies indicating that these processes have been fundamental motors of evolutionary innovation (Emes et al. 2003). These proposals are further supported after the comparative analysis of rat and human degradomes. A total of 102 proteases present in rat are absent in human, whereas 37 human proteases do not have a clear counterpart in rat. A detailed analysis of these differences has allowed us to conclude that at least in quantitative terms, they mainly derive from the creation or expansion of specific protease subfamilies in rat compared with human. Thus, the rat genome contains a number of gene families encoding proteases that are absent in human. These include placental cathepsins and testins located at 17p14, hematopoietic serine proteases at 15p13, and glandular kallikreins at 1q21. Likewise, additional expansions are found in testases, sentrin-specific proteases, and nonprotease-homologs of the PIP family. Overall, these large-scale changes with respect to human proteases are shared by rat and mouse, although some of the above-mentioned small-scale variations between rat and mouse protease genes might have contributed to shape functional differences between both rodent species.

The genomic analysis of the rat proteases has also led us to identify some genes that are specific of rodent versus human or vice versa. Among them, we should mention the cases of chymosin, caspase-12, cyritestins (Adam-3b and Adam-5), implantation serine proteases (ISPs), distal intestinal serine proteases (DISPs), and testis serine proteases (TESPs), present in rat and mouse but inactive or absent in human (Table 2). Conversely, caspases-5 and -10, USP6, matrilysin-2, carboxypeptidase O, aminopeptidase MAMS/L-RAP, and KLK3/PSA are absent or inactive in rat and mouse but are functional in human. The large number of protease genes inactivated in human compared with rat and mouse should be in agreement with preliminary comparative analysis of human and chimpanzee genomes, which has revealed that genetic losses in the human lineage may have caused some of the differences between these species (Olson and Varki 2003). Once more, it is remarkable that most differences between rat and human degradomes correspond to proteases involved in reproductive and immunological functions. In relation to the first type of processes, it is well established that proteolytic enzymes play multiple and diverse roles in menstruation, fertilization, ovulation, implantation, placentation, pregnancy, and involution of reproductive organs (Hulboy et al. 1997). Differences in the type, number, or expression levels of reproductive proteases in rat, mouse, and human could have contributed to facilitate speciation events and may also help to explain some functional differences between these species. On the other hand, the observed variations in proteases associated with host-defense reactions may reflect evolutionary diversification processes aimed at expanding the repertoire of immunological mechanisms in response to new physiological conditions or to new sources of pathogens or environmental stress. Furthermore, the differences in aspartic proteases involved in digestive functions could also contribute to explain rodent and human differences in gastric physiology (Kageyama 2002). It is also interesting that the evolutionary expansion in the repertoire of rat and mouse proteases has been accompanied by a marked expansion in the number of protease inhibitors belonging to different families, such as serpins, cystatins, and murinoglobulins. This fact may be part of an evolutionary strategy aimed at establishing additional regulatory mechanisms for the expanded rat and mouse degradomes compared with that of human.

Beyond these changes in protease and protease-inhibitor gene numbers, variations in regulatory motifs present in the annotated rat, mouse, and human orthologous genes may result in diverging spatiotemporal expression patterns of some of these genes and contribute to the development of evolutionary innovations in protease-mediated functions in each species. To date, these species-specific regulatory differences in protease and protease-inhibitor genes are largely unknown, but the availability of the rat, mouse, and human genome sequences, together with the introduction of protease-specific chips for the global analysis of expression and activity of these enzymes and their inhibitors (López-Otín and Overall 2002), may be very helpful for a better understanding of putative changes in their regulatory mechanisms.

In addition to the evolutionary relevance of these comparative analyses of the rat, mouse, and human degradomes, the annotation of the rat protease complement and the identification of the rat orthologs of human protease genes associated with pathological conditions could facilitate the development of new strategies for better understanding and treating these diseases. The human diseases of proteolysis have been classically linked to alterations in the control of the spatiotemporal patterns of expression of a number of proteases, including MMPs, cathepsins, or plasminogen activators (Puente et al. 2003). Thus, overexpression of many of these proteases is commonly found in cancer, arthritis, cardiovascular diseases, and inflammatory disorders. The identification of rat orthologs of these human proteases, together with the already available information on their mouse counterparts, could help to identify the regulatory elements mediating their abnormal expression in pathological conditions. These studies could also facilitate the creation of rat transgenic models useful to examine in vivo the consequences of protease overexpression in certain tissues. In this regard, it is noteworthy that the first available transgenic rats—created in 1990—have been very useful to examine the role of proteases in cardiovascular diseases (Mullins et al. 1990). In fact, the TGRmRen2 transgenic rats that overexpress mouse submandibular renin develop fulminant hypertension and are a widely used model for studying multiple aspects of cardiovascular physiology and pathology. Nevertheless, and in addition to these regulatory diseases of proteolysis, we have cataloged 58 human hereditary disorders directly caused by mutations in protease genes (Puente et al. 2003; http://web.uniovi.es/degradome). Interestingly, with the single exception of the caspase-10 gene—mutated in patients with an autoimmune lymphoproliferative syndrome but lacking a recognizable ortholog in rat—all the remaining genes linked to human hereditary diseases of proteolysis are conserved in rat. These observations may provide a useful framework for discussing the possibilities of creating rat models for these human diseases that could complement the information derived from studies performed in knockout or knockin mouse models. This could be of special interest for those human genetic disorders of proteolysis in which the available mouse models do not recapitulate the corresponding human disease (Puente et al. 2003). The recent development of methods for manipulating the rat germline (Zan et al. 2003) will offer new opportunities to investigate the mutational mechanisms and pathological alterations underlying these genetic disorders of proteolysis, as well as to evaluate the potential usefulness of new therapeutic strategies for this growing and relevant group of human diseases.

METHODS

Bioinformatic Screening of the Rat Genome

A nonredundant set of human and mouse protease sequences was built by combining information from literature, the MEROPS database, the proteome analysis database (http://www.ebi.ac.uk/proteome/), and annotations from our laboratory (Puente et al. 2003). Orthology assignation for each of the 641 mouse protease genes as well as the 31 human-specific proteases present in our custom degradome database was performed by using the Ensembl and the UCSC genome browsers against the rat Baylor assemblies 2 and 3.1, respectively; the nonredundant nucleotide database; the rat EST databases at GenBank; and the BLAT and TBLASTN algorithms. Similar analyses using a custom database of protease inhibitors were performed to search for protease inhibitors in the rat genome and to further compare them with those present in mouse and human. Orthology assignment was based on four different criteria: synteny, sequence identity (>85% rat versus mouse, >65% rat versus human, and reciprocal best-match), function conservation, and relevant supporting literature. If one of these criteria was not met, a detailed analysis, including conservation of neighbor genes in both species and examination of genome gaps, was performed before orthology or paralogy assignment.

Use of TBLASTN and InterPro Annotations to Identify Proteases

Rat genomic sequences were analyzed for the presence of unidentified protease members belonging to the 67 families currently recognized in mouse and human, using TBLASTN at the Ensembl genome browser. Each single mouse protease sequence and those human specific were used to query the rat genome, and all hits with P value $<10^{-2}$ were analyzed by using the BLASTP program against a custom degradome database. Hits not present in the custom database were further analyzed by using TBLASTN against the nonredundant nucleotide database at National Center for Biotechnology Information (NCBI) and TBLASTN of a 500,000-bp fragment containing the hit against similar proteases to build a predicted sequence. These strategies allowed us to extend the putative protease fragments, and manual inspection against the murine and human paralogs as well as available rat EST sequences was used to complete the full open reading frame. Pseudogenes were confirmed by presence of premature stop codons or frameshifts in sequences derived from the rat genome assembly v3.1, high-throughput genomic sequences at NCBI, and EST sequences if available. Profile recognition programs, including SMART and InterPro, were used to determine the presence of protease motifs, and multiple sequence alignments were used to finally classify a protein as protease or nonprotease homolog. For each single protease locus, a 500-kb genomic sequence flanking the target gene was analyzed for the presence of additional members of that family, as ~31% of rat, 34% of mouse, and 23% of human protease genes are organized in clusters. Additionally, InterPro annotations of the rat genome were used to identify putative new members of known families. Ensembl predictions containing InterPro protease motifs were manually inspected to distinguish between true proteases, pseudogenes, and false positives. A combination of detailed genomic sequence analysis and relevant literature was used to determine the number of protease genes present in clusters of protease genes with high sequence similarity, as they can be accidentally collapsed during assembly of the shotgun sequences. Several duplication events were found and discarded as assembly artifacts due to the presence of regions of 100% nucleotide sequence identity between both the exonic and intronic regions of protease genes. To further extend the bioinformatic search of protease genes, we built a hidden Markov model (HMM) for some of the 67 different protease families present in the rat degradome. For families containing less than three members, alignments were built by using sequences from other organisms if available. Additional HMMs from protease families not described in mammals were obtained from the Pfam protein families, version 7.8, database (http://www.sanger.ac.uk/Software/Pfam/). The selectivity of these models was tested against the SWISS-PROT, release 40, database, identifying known proteases when a P value cutoff of 0.1 was used. HMMs were used to screen the rat protein sequence predictions from Ensembl (Release 15.2.1; http://www.ensembl.org), and the rat-specific RefSeq data set at NCBI (version June 15, 2003; http://www.ncbi.nlm.nih.gov) using the HMMER 2.1 package (http://hmmer.wustl.edu). Finally, analysis of the presence of ancillary domains in rat proteases was performed by using the SMART and Pfam domain databases. Sequences derived from this work have been submitted to the EMBL database with accession nos. BN000318–BN000390.

Creation of Phylogenetic Trees

To establish orthology or paralogy in densely populated clusters of protease genes in which different members were specifically expanded, a phylogenetic tree was generated. Protein sequences corresponding to the full-length protease from different species were aligned by using the ClustalX program, together with a more distantly related protease to be used as root. Phylogenetic trees were constructed for each family by using the Protpars program from the Phylip package (http://evolution.genetics.washington.edu/phylip.html).

ACKNOWLEDGMENTS

We thank all members of our lab for their comments on the manuscript. Our work is supported by grants from Comisión *Interministerial de Ciencia y Tecnología-Spain, Gobierno del Principado de Asturias, Fundación "La Caixa," and European Union. The Insti-*

tuto Universitario de Oncología is supported by Obra Social Cajastur-Asturias, Spain.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abel, K.J. and Gross, K.W. 1990. Physical characterization of genetic rearrangements at the mouse renin loci. *Genetics* **124:** 937–947.

Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J.R., and Hilgenfeld, R. 2003. Coronavirus main proteinase (3CLpro) structure: Basis for design of anti-SARS drugs. *Science* **300:** 1763–1767.

Baek, K.H., Mondoux, M.A., Jaster, R., Fire-Levin, E., and D'Andrea, A.D. 2001. DUB-2A: A new member of the DUB subfamily of hematopoietic deubiquitinating enzymes. *Blood* **98:** 636–642.

Balakirev, M.Y., Tcherniuk, S.O., Jaquinod, M., and Chroboczek, J. 2003. Otubains: A new family of cysteine proteases in the ubiquitin pathway. *EMBO Rep.* **4:** 517–522.

Balbín, M., Fueyo, A., Knauper, V., López, J.M., Alvarez, J., Sánchez, L.M., Quesada, V., Bordallo, J., Murphy, G., and López-Otín, C. 2001. Identification and enzymatic characterization of two diverging murine counterparts of human interstitial collagenase (MMP-1) expressed at sites of embryo implantation. *J. Biol. Chem.* **276:** 10253–10262.

Barbour, K.W., Wei, F., Brannan, C., Flotte, T.R., Baumann, H., and Berger, F.G. 2002. The murine α_1-proteinase inhibitor gene family: Polymorphism, chromosomal location, and structure. *Genomics* **80:** 515–522.

Barrett, A.J., Rawlings, N.D., and Woessner, J.F. 1998. *Handbook of proteolytic enzymes*. Academic Press, San Diego, CA.

Brinckerhoff, C.E. and Matrisian, L.M. 2002. Matrix metalloproteinases: A tail of a frog that became a prince. *Nat. Rev. Mol. Cell Biol.* **3:** 207–214.

Cal, S., Quesada, V., Garabaya, C., and López-Otín, C. 2003. Polyserase-I: A human polyprotease with the ability to generate independent serine protease domains from a single translation product. *Proc. Natl. Acad. Sci.* **100:** 9185–9190.

Caputo, E., Manco, G., Mandrich, L., and Guardiola, J. 2000. A novel aspartyl proteinase from apocrine epithelia and breast tumors. *J. Biol. Chem.* **275:** 7935–7941.

Chipuk, J.E., Stewart, L.V., Ranieri, A., Song, K., and Danielpour, D. 2002. Identification and characterization of a novel rat ov-serpin family member, trespin. *J. Biol. Chem.* **277:** 26412–26421.

Chureau, C., Prissette, M., Bourdet, A., Barbe, V., Cattolico, L., Jones, L., Eggen, A., Avner, P., and Duret, L. 2002. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res.* **12:** 894–908.

Cope, G.A., Suh, G.S., Aravind, L., Schwarz, S.E., Zipursky, S.L., Koonin, E.V., and Deshaies, R.J. 2002. Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science* **298:** 608–611.

Cornwall, G.A. and Hsia, N. 2003. A new subgroup of the family 2 cystatins. *Mol. Cell Endocrinol.* **200:** 1–8.

Deussing, J., Kouadio, M., Rehman, S., Werber, I., Schwinde, A., and Peters, C. 2002. Identification and characterization of a dense cluster of placenta-specific cysteine peptidase genes and related genes on mouse chromosome 13. *Genomics* **79:** 225–240.

Emes, R.D., Goodstadt, L., Winter, E.E., and Ponting, C.P. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12:** 701–709.

Fischer, H., Koenig, U., Eckhart, L., and Tschachler, E. 2002. Human caspase 12 has acquired deleterious mutations. *Biochem. Biophys. Res. Commun.* **293:** 722–726.

Forsyth, S., Horvath, A., and Coughlin, P. 2003. A review and comparison of the murine α_1-antitrypsin and α_1-antichymotrypsin multigene clusters with the human clade A serpins. *Genomics* **81:** 336–345.

Garnier, G., Circolo, A., Xu, Y., and Volanakis, J.E. 2003. Complement C1r and C1s genes are duplicated in the mouse: Differential expression generates alternative isomorphs in the liver and in the male reproductive system. *Biochem. J.* **371:** 631–640.

Gururajan, R., Lahti, J.M., Grenet, J., Easton, J., Gruber, I., Ambros, P.F., and Kidd, V.J. 1998. 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:** 929–939.

Hooper, N.M. 2002. *Proteases in biology and medicine*. Portland Press, London.

Horikawa, Y., Oda, N., Cox, N.J., Li, X., Orho-Melander, M., Hara, M., Hinokio, Y., Lindner, T.H., Mashima, H., Schwarz, P.E., et al. 2000. Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nat. Genet.* **26:** 163–175.

Hulboy, D.L., Rudolph, L.A., and Matrisian, L.M. 1997. Matrix metalloproteinases as mediators of reproductive function. *Mol. Hum. Reprod.* **3:** 27–45.

Imamura, T. 2003. The role of gingipains in the pathogenesis of periodontal disease. *J. Periodontol.* **74:** 111–118.

Jacob, H.J. and Kwitek, A.E. 2002. Rat genetics: Attaching physiology and pharmacology to the genome. *Nat. Rev. Genet.* **3:** 33–42.

Kageyama, T. 2002. Pepsinogens, progastricsins, and prochymosins: Structure, function, evolution, and development. *Cell. Mol. Life Sci.* **59:** 288–306.

Kaiserman, D., Knaggs, S., Scarff, K.L., Gillard, A., Mirza, G., Cadman, M., McKeone, R., Denny, P., Cooley, J., Benarafa, C., et al. 2002. Comparison of human chromosome 6p25 with mouse chromosome 13 reveals a greatly expanded ov-serpin gene repertoire in the mouse. *Genomics* **79:** 349–362.

Kolmer, M., Ord, T., Alhonen, L., Hyttinen, J.M., Saarma, M., Villems, R., and Janne, J. 1991. Assignment of human prochymosin pseudogene to chromosome 1. *Genomics* **10:** 496–498.

Llamazares, M., Cal, S., Quesada, V., and López-Otín, C. 2003. 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:** 13382–13389.

López-Otín, C. and Overall, C.M. 2002. Protease degradomics: A new challenge for proteomics. *Nat. Rev. Mol. Cell. Biol.* **3:** 509–519.

Lunderius, C. and Hellman, L. 2001. Characterization of the gene encoding mouse mast cell protease 8 (mMCP-8), and a comparative analysis of hematopoietic serine protease genes. *Immunogenetics* **53:** 225–232.

MacDonald, R.J., Southard-Smith, E.M., and Kroon, E. 1996. Disparate tissue-specific expression of members of the tissue kallikrein multigene family of the rat. *J. Biol. Chem.* **271:** 13684–13690.

Mullins, J.J., Peters, J., and Ganten, D. 1990. Fulminant hypertension in transgenic rats harbouring the mouse Ren-2 gene. *Nature* **344:** 541–544.

Neurath, H. 1999. Proteolytic enzymes, past and future. *Proc. Natl. Acad. Sci.* **96:** 10962–10963.

Okada, T., Gondo, Y., Goto, J., Kanazawa, I., Hadano, S., and Ikeda, J.E. 2002. Unstable transmission of the RS447 human megasatellite tandem repetitive sequence that contains the USP17 deubiquitinating enzyme gene. *Hum. Genet.* **110:** 302–313.

Olson, M.V. and Varki, A. 2003. Sequencing the chimpanzee genome: Insights into human evolution and disease. *Nat. Rev. Genet.* **4:** 20–28.

Olsson, A.Y. and Lundwall, A. 2002. Organization and evolution of the glandular kallikrein locus in *Mus musculus*. *Biochem. Biophys. Res. Commun.* **299:** 305–311.

Overall, C.M. and López-Otín, C. 2002. Strategies for MMP inhibition in cancer: Innovations for the post-trial era. *Nat. Rev. Cancer.* **2:** 657–672.

Paulding, C.A., Ruvolo, M., and Haber, D.A. 2003. The Tre2 (USP6) oncogene is a hominoid-specific gene. *Proc. Natl. Acad. Sci.* **100:** 2507–2511.

Perler, F.B. 1998. Protein splicing of inteins and hedgehog autoproteolysis: Structure, function, and evolution. *Cell* **92:** 1–4.

Puente, X.S., Sánchez, L.M., Overall, C.M., and López-Otín, C. 2003. Human and mouse proteases: A comparative genomic approach. *Nat. Rev. Genet.* **4:** 544–558.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway Rat yields insights into mammalian evolution. *Nature* (in press).

Riordan, J.F. 2003. Angiotensin I–converting enzyme and its relatives. *Genome Biol.* **4:** 225–235.

Rose, S.D. and MacDonald, R.J. 1997. Evolutionary silencing of the human elastase I gene (ELA1). *Hum. Mol. Genet.* **6:** 897–903.

Ross, J., Jiang, H., Kanost, M.R., and Wang, Y. 2003. Serine proteases and their homologs in the *Drosophila melanogaster* genome: An initial analysis of sequence conservation and phylogenetic relationships. *Gene* **304:** 117–131.

Santamaría, I., Velasco, G., Cazorla, M., Fueyo, A., Campo, E., and López-Otín, C. 1998. Cathepsin L2, a novel human cysteine proteinase produced by breast and colorectal carcinomas. *Cancer Res.* **58:** 1624–1630.

Scott, I.C., Halila, R., Jenkins, J.M., Mehan, S., Apostolou, S., Winqvist, R., Callen, D.F., Prockop, D.J., Peltonen, L., and Kadler, K.E. 1996. Molecular cloning, expression and chromosomal localization of a human gene encoding a 33 kDa putative metallopeptidase (PRSM1). *Gene* **174:** 135–143.

Shao, F., Merritt, P.M., Bao, Z., Innes, R.W., and Dixon, J.E. 2002. A Yersinia effector and a Pseudomonas avirulence protein define a family of cysteine proteases functioning in bacterial pathogenesis. *Cell* **109:** 575–588.

Silverman, G.A., Bird, P.I., Carrell, R.W., Church, F.C., Coughlin, P.B., Gettins, P.G., Irving, J.A., Lomas, D.A., Luke, C.J., Moyer, R.W., et al. 2001. The serpins are an expanding superfamily of structurally similar but functionally diverse proteins: Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. *J. Biol. Chem.* **276:** 33293–33296.

Sol-Church, K., Picerno, G.N., Stabley, D.L., Frenck, J., Xing, S., Bertenshaw, G.P., and Mason, R.W. 2002. Evolution of placentally expressed cathepsins. *Biochem. Biophys. Res. Commun.* **293:** 23–29.

Soto, D., Malmsten, C., Blount, J.L., Muilenburg, D.J., and Caughey, G.H. 2002. Genetic deficiency of human mast cell α-tryptase. *Clin. Exp. Allergy* **32:** 1000–1006.

Swanson, W.J. and Vacquier, V.D. 2002. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3:** 137–144.

Takano, J., Kawamura, T., Murase, M., Hitomi, K., and Maki, M. 1999. Structure of mouse calpastatin isoforms: Implications of species-common and species-specific alternative splicing. *Biochem. Biophys. Res. Commun.* **260:** 339–345.

Uría, J.A. and López-Otín, C. 2000. Matrilysin-2, a new matrix metalloproteinase expressed in human tumors and showing the minimal domain organization required for secretion, latency, and activity. *Cancer Res.* **60:** 4745–4751.

Van Eerdewegh, P., Little, R.D., Dupuis, J., Del Mastro, R.G., Falls, K., Simon, J., Torrey, D., Pandit, S., McKenny, J., Braunschweiger, K., et al. 2002. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* **418:** 426–430.

Velasco, G., Pendás, A.M., Fueyo, A., Knäuper, V., Murphy, G., and López-Otín, C. 1999. 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:** 4570–4576.

Wang, P.J., McCarrey, J.R., Yang, F., and Page, D.C. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* **27:** 422–426.

Wilson, M.A., Collins, J.L., Hod, Y., Ringe, D., and Petsko, G.A. 2003. The 1.1-Å resolution crystal structure of DJ-1, the protein mutated in autosomal recessive early onset Parkinson's disease. *Proc. Natl. Acad. Sci.* **100:** 9256–9261.

Wu, Y., Wang, X., Liu, X., and Wang, Y. 2003. Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite. *Genome Res.* **13:** 601–616.

Yoshida, M., Kaneko, M., Kurachi, H., and Osawa, M. 2001. Identification of two rodent genes encoding homologues to seminal vesicle autoantigen: A gene family including the gene for prolactin-inducible protein. *Biochem. Biophys. Res. Commun.* **281:** 94–100.

Yousef, G.M. and Diamandis, E.P. 2001. The new human tissue kallikrein gene family: Structure, function, and association to disease. *Endocr. Rev.* **22:** 184–204.

Zan, Y., Haag, J.D., Chen, K.S., Shepel, L.A., Wigington, D., Wang, Y.R., Hu, R., Lopez-Guajardo, C.C., Brose, H.L., Porter, K.I., et al. 2003. Production of knockout rats using ENU mutagenesis and a yeast-based screening assay. *Nat. Biotechnol.* **21:** 645–651.

WEB SITE REFERENCES

http://merops.sanger.ac.uk; MEROPS database of proteases and protease inhibitors.

http://web.uniovi.es/degradome; Database of proteases and diseases of proteolysis.

http://www.ncbi.nlm.nih.gov; National Center for Biotechnology Information.

http://www.ensembl.org; Ensembl rat, human, and mouse genome home page.

http://www.sanger.ac.uk/Software/Pfam/; Protein domain analysis.

http://www.ebi.ac.uk/proteome; Comprehensive database for proteomes of fully sequenced organisms.

http://hmmer.wustl.edu; Hidden Markov Model software for protein analysis.

http://evolution.genetics.washington.edu/phylip.html; Package of programs for phylogenetic analysis.

Received September 8, 2003; accepted in revised form November 24, 2003.