Birth-and-Death of *KLK3* and *KLK2* in Primates: Evolution Driven by Reproductive Biology

Patrícia Isabel Marques[1,2,3], Rui Bernardino[4], Teresa Fernandes[4], NISC Comparative Sequencing Program[5,6], Eric D. Green[5], Belen Hurle[5], Victor Quesada[2,*], and Susana Seixas[1,*]

[1]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal

[2]Department of Biochemistry and Molecular Biology-IUOPA, University of Oviedo, Oviedo, Spain

[3]Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal

[4]Lisbon Zoo Veterinary Hospital, Lisbon, Portugal

[5]National Human Genome Research Institute, National Institutes of Health (NIH), Bethesda, Maryland

[6]NIH Intramural Sequencing Center (NISC), National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland

*Corresponding author: E-mail: sseixas@ipatimup.pt; quesadavictor@uniovi.es.

Accepted: November 25, 2012

Data deposition: GenBank accession numbers for all the BAC genomic entries and the assembly coordinates of the genomic segments from reference sequences (UCSC Genome Browser) used in gene annotation are provided in supplementary table S1, Supplementary Material online.

Abstract

The *kallikrein* (*KLK*) gene family comprises the largest uninterrupted locus of serine proteases in the human genome and represents a notable case for studying the evolutionary fate of duplicated genes. In primates, a recent duplication event gave rise to *KLK2* and *KLK3*, both encoding essential proteins for the cascade of seminal plasma liquefaction. We reconstructed the evolutionary history of *KLK2* and *KLK3* by comparative analysis of the orthologous sequences from 22 primate species, calculated d_N/d_S ratios, and addressed the hypothesis of coevolution with their substrates, the semenogelins (SEMG1 and SEMG2). Our findings support the placement of the *KLK2–KLK3* duplication in the Catarrhini ancestor and unveil the frequent loss of *KLK2* throughout primate evolution by different genomic mechanisms, including unequal crossing-over, deletions, and pseudogenization. We provide evidences for an adaptive evolution of KLK3 toward an expanded enzymatic spectrum, with an effect on the hydrolysis of semen coagulum. Furthermore, we found associations between mating system, the number of SEMG repeat units, and the number of functional *KLK2* and *KLK3*, suggesting complex evolutionary dynamics shaped by reproductive biology.

Key words: serine proteases, adaptive evolution, mating system, semen coagulation, semenogelins.

The birth-and-death of genes has a significant impact in genome evolution, particularly in gene families involved in physiological traits such as sensory systems, immunity, and reproduction (Zhang 2003; Demuth and Hahn 2009). According to this model, new genes are generated by duplication, and although some are maintained in the genome (acquiring novel or altered functions), others are disrupted or become nonfunctional through a variety of deleterious mechanisms (Nei and Rooney 2005; Kaessmann 2010). In this context of gene gain, diversification, and loss, the *kallikrein* (*KLK*) cluster, the largest locus in the human genome of phylogenetically related serine proteases (Yousef and Diamandis 2001), represents a remarkable case for the study of the evolutionary fate of duplicates. In humans, the *KLK* cluster spans over 265 kb on chromosome 19q13.4 and includes 15 genes ranging from 4.4 to 10.5 kb, most of them sharing a common gene structure with five coding exons (Yousef and Diamandis 2001; Lundwall and Brattsand 2008). KLKs act mainly as trypsin or chymotrypsin-like proteases in a number of biological processes such as skin desquamation, semen liquefaction, neuroplasticity, and regulation of blood pressure (Emami and Diamandis 2007). In primates, a recent gene duplication gave rise to two kallikreins, *KLK2* and *KLK3* (encoding prostate-specific antigen), which play a crucial role in the proteolytic cascade of seminal plasma liquefaction (Lundwall and Brattsand 2008). Briefly, upon ejaculation, the epididymal fluid is mixed with prostate and seminal vesicles secretions containing semenogelins

(SEMG1 and SEMG2) to form a coagulum that entraps spermatozoa. Later, these spermatozoa are released with the hydrolysis of SEMGs by KLK3 and KLK2. In addition, KLK2 is also thought to activate KLK3 (Lovgren et al. 1999; Lundwall and Brattsand 2008). Previous findings suggest that primate *KLK2* and *KLK3* (Clark and Swanson 2005), along with *SEMGs* (Jensen-Seaman and Li 2003; Dorus et al. 2004; Hurle et al. 2007), may be targets of natural selection and could provide an important example of birth-and-death evolution. Here, we reconstruct the evolutionary history of *KLK2* and *KLK3* in primates and test the hypothesis of their coevolution with *SEMGs* as a possible example of evolution driven by male reproductive biology.

KLK2 and *KLK3* Gains and Losses

To better understand the evolutionary dynamics of *KLK2* and *KLK3* genes in primates, we sequenced and/or annotated the orthologous genomic segments spanning these genes in a total of 22 primate species (supplementary table S1, Supplementary Material online). We confirmed the presence of *KLK2* and *KLK3* in all Catarrhini, except for *Colobus guereza*, *Gorilla gorilla,* and *Nomascus leucogenys*, and the presence of a single *KLK2* ortholog sequence in Platyrrhini and Strepsirrhini (fig. 1*A*). This result reinforces the hypothesis of a *KLK3* origin by *KLK2* duplication after the Catarrhini split approximately 42 million years ago (Olsson et al. 2004;

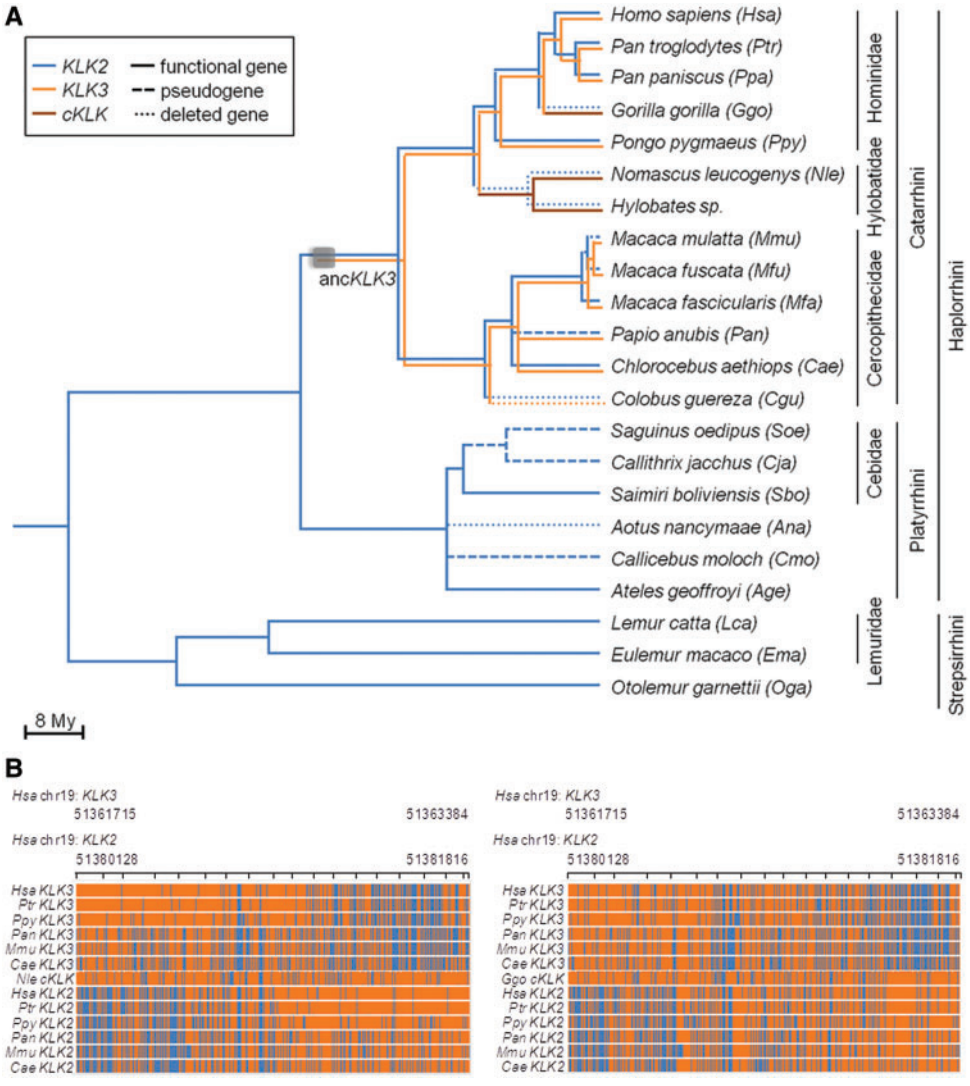

Fig. 1.—Phylogenetic analysis of *KLK2* and *KLK3* in primates. (*A*) Phylogenetic tree showing primate divergence times (Hedges et al. 2006) and functional status of *KLK2* and *KLK3*. The criteria to define a nonfunctional *KLK* gene were the identification of at least one disrupting mutation. Gray square indicates a duplication event. The ancestral *KLK3* branch is indicated (anc*KLK3*). (*B*) Alignment of exons IV–V for *KLK2* and *KLK3* in Catarrhini. The corresponding human genomic positions for these regions are represented at the top. Positions conserved with *Gorilla gorilla* (left panel) or *Nomascus leucogenys* (right panel) are in orange. Nonconserved positions are in blue. Sites conserved in all species were omitted.

Table 1
Identified *KLK2* Deleterious Mutations

Species	Deleterious Mutations[a]
Hsa, Ptr, Ppa, Ppy, Mfa, Cae, Sbo, Age, Lca, Ema, Oga	None
Mmu	D120A[b,c] and R109X[d]
Mfu	G51R,[c,d] L54P,[c,d] and L171fsX181[d]
Pan	V247fsX337
Cja	M1I, W47X, IVS3+1G>A,[e] IVS3-1G>A,[e] and L178_C184delinsVfsX190
Soe	M1I,[f] S213T,[c,f] R250X,[f] and IVS4+1G>A[f]
Cmo	M1L, I25T, C184fsX189, and IVS4-2A>T[e]

[a]Mutations are displayed according to the recommended nomenclature for the description of human sequence variations (den Dunnen and Antonarakis 2000).
[b]Catalytic triad mutation (Clark and Swanson 2005).
[c]Possible damaging as predicted by Polyphen2 (Adzhubei et al. 2010).
[d]Polymorphic site.
[e]Splice site mutation.
[f]Previously identified mutations (Olsson et al. 2004).

Valtonen-Andre et al. 2005; Pavlopoulou et al. 2010). Notably, we identified two *KLK3–KLK2* fusions in *G. gorilla* and *N. leucogenys* yielding single chimeric *KLK* genes (*cKLK*) (supplementary fig. S1*A* and S1*B*, Supplementary Material online). We located the breaking point in both species to a few bases in intron IV, in the vicinity of a LINE2 element common to *KLK2* and *KLK3* sequences (*G. gorilla* IVS4+622_781 and *N. leucogenys* IVS4+787_1026; fig. 1*B*). These genomic rearrangements were confirmed by direct sequencing of three additional *G. gorilla* individuals and five Hylobatidae samples, indicating a likely fixation of *cKLK* in these taxa (supplementary fig. S1*C*, Supplementary Material online). In both cases, the first four exons of *cKLK* are orthologous to *KLK3*, whereas the last exon is more similar to *KLK2* (fig. 1*B*). At the protein level, these genomic rearrangements account only for minor amino acid replacements relative to the expected KLK3 sequence (S231P, R239K, S241A, L242V, and V258A). Because these replacements are not predicted to alter protein structure or function, *cKLK* is likely a functional *KLK3*-like gene. Our findings confirm previous reports, which suggested the partial loss of *KLK2* in *G. gorilla* and *Hylobates* sp. (Clark and Swanson 2005). On the other hand, a detailed analysis of the alignments of *C. guereza* genomic sequences with the *Homo sapiens* reference genome showed the complete loss of *KLK2* and *KLK3* in this species, possibly by two deletion events (supplementary fig. S2, Supplementary Material online). In Cercopithecoidea, we identified several loss-of-function events in *KLK2* through a variety of deleterious mechanisms (table 1). These include a premature stop codon in *Macaca mulatta* (R109X) and a frameshift mutation (L171fsX181) and two nonsynonymous substitutions (G51R and L54P) in *M. fuscata*. The mutation of the KLK2 catalytic triad (D120A) previously described in *M. mulatta* (Clark and Swanson 2005) was not observed, and no evidence for the accumulation of deleterious mutations in *M. fascicularis* was found. In *Papio anubis,* we identified a frameshift mutation leading to a 75-codon longer open reading frame (V247fsX337), which is unlikely to be translated into a KLK2 (supplementary fig. S3, Supplementary Material online). Additional examples of *KLK2* loss were observed in Platyrrhini, either by gene deletion or disruption (fig. 1*A* and table 1). In this taxon, the single example of *KLK2* loss by deletion was found in *Aotus nancymaae,* whereas several deleterious mutations were detected in *Callicebus moloch, Callithrix jacchus*, and *Saguinus oedipus.* In *Cal. moloch,* these mutations affect the starting codon (ATG-TTG), alter the activation site (I25T), and produce a premature stop codon (C184fsX189). In *Callithrix jacchus,* we identified a disrupted start codon (ATG-ATA) and a premature stop codon (W47X). In *S. oedipus,* a sister species of *Callithrix jacchus,* we have confirmed that *KLK2* is a pseudogene due to the accumulation of several mutations predicted to impair the translation of a functional serine protease (Olsson et al. 2004). All these species have an alternative starting codon 18 bp upstream of the consensus site; however, this is not expected to lead to an active KLK2 due to the occurrence of additional damaging mutations (supplementary fig. S3, Supplementary Material online, and table 1). In Strepsirrhini, no deleterious mutations were detected, suggesting a functional *KLK2* (fig. 1*A* and supplementary fig. S3, Supplementary Material online).

KLK2 and *KLK3* Phylogenetic Analysis

To address the extent of the selective pressures exerted on *KLK2* and *KLK3*, we calculated d_N/d_S (ω; d_S—synonymous substitution rate and d_N—nonsynonymous substitution rate) ratios under alternative models of gene evolution. To this end, we performed a series of branch models to test whether *KLK2* and *KLK3* experienced different selective pressures during primate evolution. First, we estimated a single ω for the entire phylogeny (one-ratio model), in which we assumed no differentiation in *KLK2* and *KLK3* selective constrains. The observed ω value below 1 ($\omega_{KLK} = 0.54$) pointed out to an overall conservation of *KLK2* and *KLK3* (table 2). Then, to examine whether the two paralogs were subjected to

Table 2
Parameter Estimates and Likelihood Scores under Different Branch Models

Model	Parameters for Branches	Likelihood (*l*)
One ratio	$\omega_{KLK} = 0.54$	−4,390.93
Two ratios	$\omega_{KLK2} = 0.55$	−4,390.90
	$\omega_{KLK3} = 0.53$	
Three ratios	$\omega_{KLK2} = 0.48$	−4,386.40
	$\omega_{pKLK2} = 1.16$	
	$\omega_{KLK3} = 0.53$	
Four ratios	$\omega_{KLK2} = 0.47$	−4,384.35
	$\omega_{pKLK2} = 1.16$	
	$\omega_{KLK3} = 0.43$	
	$\omega_{ancKLK3} = 0.90$	

Models Compared	$-2\Delta l$	*P*
One vs. two ratios	0.06 (df = 1)	0.806
Two vs. three ratios	9.00** (df = 1)	0.003
Three vs. four ratios	4.10* (df = 1)	0.043

NOTE.—ω_{KLK}, ω for all *KLK2* and *KLK3* lineages; ω_{KLK2}, ω for all *KLK2* lineages; ω_{KLK3}, ω for all *KLK3* lineages; ω_{pKLK2}, ω for *KLK2* pseudogene lineages; $\omega_{ancKLK3}$, ω for the ancestral *KLK3* lineage; df - degrees of freedom.

*Significant $P < 0.05$.

**Significant $P < 0.01$.

different selective pressures, we applied a different model (two-ratio model) considering two branches within the phylogeny comprising either the *KLK2* or *KLK3* clades. Both ω values were below 1 ($\omega_{KLK2} = 0.55$; $\omega_{KLK3} = 0.53$) (table 2) and did not differ from the previous reported model ($-2\Delta l = 0.06$; $P > 0.05$). Given the evidences for *KLK2* pseudogenization in several primate species, we anticipated a contrast in the relaxation of selective constraints in pseudogenes and their functional orthologs. In our models, we considered this hypothesis by subdividing *KLK2* clade into functional and pseudogenized (*pKLK2*). This model (three-ratio model) had a significant higher likelihood and an improved fit to *KLK2* and *KLK3* evolution ($-2\Delta l = 9$, $P < 0.01$). Furthermore, ω estimates corroborated the neutral evolution of *KLK2* pseudogenes ($\omega_{pKLK2} = 1.16$), with a possible relaxation of selective constrains after the duplication event ($\omega_{KLK2} = 0.48$; $\omega_{KLK3} = 0.53$; table 2). Therefore, to test whether *KLK3* had been subjected to different selective pressures following the duplication, the ancestral *KLK3* branch (anc*KLK3*) was regarded as an independent clade. The last model (four-ratio model) provided the best fit to the evolutionary history of *KLK2* and *KLK3* ($-2\Delta l = 4.10$, $P < 0.05$). According to the ω values estimated, an episode of reduced selective constrains occurred immediately after the duplication event ($\omega_{ancKLK3} = 0.90$); stronger selective pressures are operating at *KLK3* and *KLK2* ($\omega_{KLK3} = 0.43$; $\omega_{KLK2} = 0.47$), and a complete release of selective constraints is observed among *KLK2* pseudogenes ($\omega_{pKLK2} = 1.16$; table 2). Importantly, if the ancestral *KLK3* experienced brief episodes of adaptive evolution, it is unlikely to produce an ω value greater than 1, because most residues were subjected to strong constrains and only a few were under positive selection. To test the adaptive hypothesis of the ancestral *KLK3*, we performed a branch-site model, in which branches on the phylogeny are divided a priori into foreground (ancestral *KLK3* branch) and background and selective pressures are allowed to vary over sites and branches. Even though the majority of sites are constrained or neutrally evolving, four codon positions (13, 41, 72, and 207) show a footprint of positive selection with posterior probability higher than 85% (table 3). A similar approach was applied to the functional *KLK2* and *KLK3* data sets using the site models test. In these cases, variable ω ratios among sites were calculated for each gene and neutral and selection models compared (M1 vs. M2 and M7 vs. M8). In both cases, selection models fit significantly better the *KLK2* and *KLK3* data than neutral models (table 3), and eight (18, 67, 69, 109, 177, 205, 210, and 250) and five (45, 189, 203, 238, and 248) codon positions were identified as being positively selected in *KLK2* and *KLK3*, respectively (table 3).

To uncover the adaptive impact of the amino acids replacements targeted by positive selection, we mapped the corresponding residues onto three-dimensional models of KLK2 and KLK3. From the eight sites identified for KLK2, amino acids 177 and 210 are located in the catalytic pocket and 109 in the kallikrein loop (fig. 2*A*). Among KLK3 selected sites, the D207S replacement that occurred shortly after the duplication is located at the base of the substrate-biding pocket (fig. 2*B*). Noteworthy, D207S altered enzyme activity to a chymotrypsin-like specificity and modified substrate affinity to medium size hydrophobic (tyrosine, leucine, valine, and phenylalanine) or basic residues (arginine, lysine, and histidine) (Debela et al. 2006). On the other hand, KLK2 conserved the aspartate residue at position 207, which is known to confer trypsin-like specificity to kallikrein-related peptidases and to display a strong preference for arginine in substrates (Janssen et al. 2004; Debela et al. 2006; Emami and Diamandis 2007). The findings of KLK3 adaptive evolution and of its expanded enzymatic spectrum on top of the restricted KLK2 spectrum provide strong arguments for a significant impact of *KLK3* emergence in the hydrolysis of semen coagulum and in the extensive SEMGs fragmentation as currently seen in humans.

Implications of *KLK2* and *KLK3* Evolution in Primate Reproductive Biology

Primate mating behavior drives the intensity of sperm competition and, with that, the evolution of genes involved in reproduction. Specifically, polyandrous species exhibit physiological traits better adapted for fertilization, like larger testis relative to body size (Harcourt et al. 1981; Clark and Swanson 2005). For *SEMGs*, a relationship between molecular evolution rates and female promiscuity has been already shown. The SEMGs are

Table 3
Model Comparisons of Variable ω Ratios among Sites

Models Compared	$-2\Delta l$	Parameter Estimates under Selection	Positively Selected Sites[a]
KLK2			
M1 vs. M2	6.56* (df = 2)	$p_1 = 0.69$, $\omega = 0.16$ $p_2 = 0$, $\omega = 1.00$ $p_3 = 0.31$, $\omega = 1.65$	109
M7 vs. M8	9.78** (df = 2)	$p_0 = 0.69$, $p = 19.76$ $q = 99.00$ ($p^1 = 0.31$), $\omega = 1.66$	18, 67, 69, 109, 177, 205, **210**, 250
KLK3			
M1 vs. M2	5.59 (df = 2)	$p_1 = 0.73$, $\omega = 0$ $p_2 = 0$, $\omega = 1.00$ $p_3 = 0.27$, $\omega = 1.83$	45, **189**
M7 vs. M8	6.37* (df = 2)	$p_0 = 0.73$, $p = 0.01$ $q = 2.82$, ($p^1 = 0.27$), $\omega = 1.83$	**45**, **189**, 203, 238, 248
Branch-site (MA)	13.06**	$p_0 = 0.51$, $\omega_{bg} = 0.13$, $\omega_{fg} = 0.13$ $p_1 = 0.41$, $\omega_{bg} = 1.00$, $\omega_{fg} = 1.00$ $p_{2a} = 0.04$, $\omega_{bg} = 0.13$, $\omega_{fg} = 17.06$ $p_{2b} = 0.03$, $\omega_{bg} = 1.00$, $\omega_{fg} = 17.06$	13, 41, 72, **207**

NOTE.—ω_{bg}, ω for background branches; ω_{fg}, ω for foreground branch (ancestral *KLK3* branch).
[a]Sites with posterior probabilities > 0.85 are indicated in regular type; *P* values > 0.90 are underlined and *P* values > 0.95 are in bold.
*Significant $P < 0.05$.
**Significant $P < 0.01$.

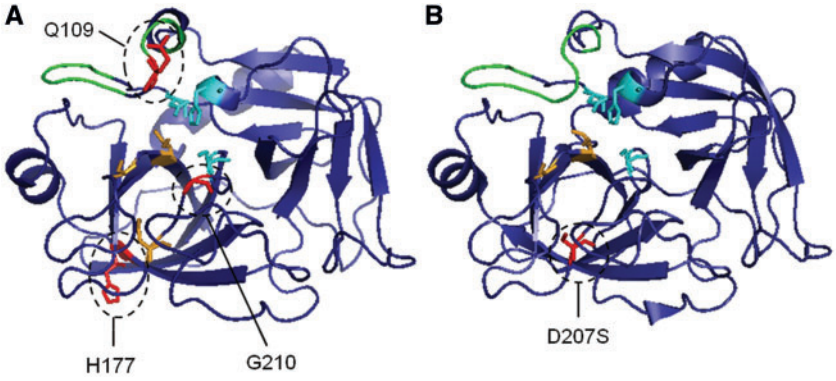

FIG. 2.—Positive selected sites in biologically relevant regions. (*A*) Human KLK2 three-dimensional model showing amino acid replacements predicted to be under positive selection (Q109, H177, and G210). (*B*) Human KLK3 three-dimensional model showing D207S substitution predicted to be under positive selection in the ancestral branch. The catalytic triad is represented in light blue (H65, D120, and S213) and the binding sites in orange (S228, G230, and D207 in KLK2 or S207 in KLK3).

highly polymorphic modular proteins, with a number of repeat units varying within and between species. This in turn dictates the degree of crosslinking between SEMGs, which influences the semen coagulum thickness. The correlation is such that the higher the promiscuity of a given species, the higher the likelihood of longer SEMGs, more crosslinking events, and a rigid copulatory plug, possibly influencing the fertilization of a recently inseminated female by rival males (Jensen-Seaman and Li 2003; Dorus et al. 2004; Hurle et al. 2007). Indeed, the number of SEMG repeats shows a significant rank correlation with residual testis size, which is a good proxy for primate mating system (Anderson et al. 2004; Dixson and Anderson 2004; Wlasiuk and Nachman 2010) (fig. 3*A* and supplementary table S2, Supplementary Material online). Considering the role of KLK2 and KLK3 in the hydrolysis of SEMGs (Rawlings et al. 2012), we tested whether the presence or absence of functional genes correlates with the number of SEMG1 and SEMG2 repeat units. In most cases, active KLK2 and KLK3 are associated with higher repeat numbers and polyandry, whereas the lack of one or both of them is linked to lower repeat numbers and monoandry (fig. 3*B* and *C* and supplementary table S2, Supplementary Material online).

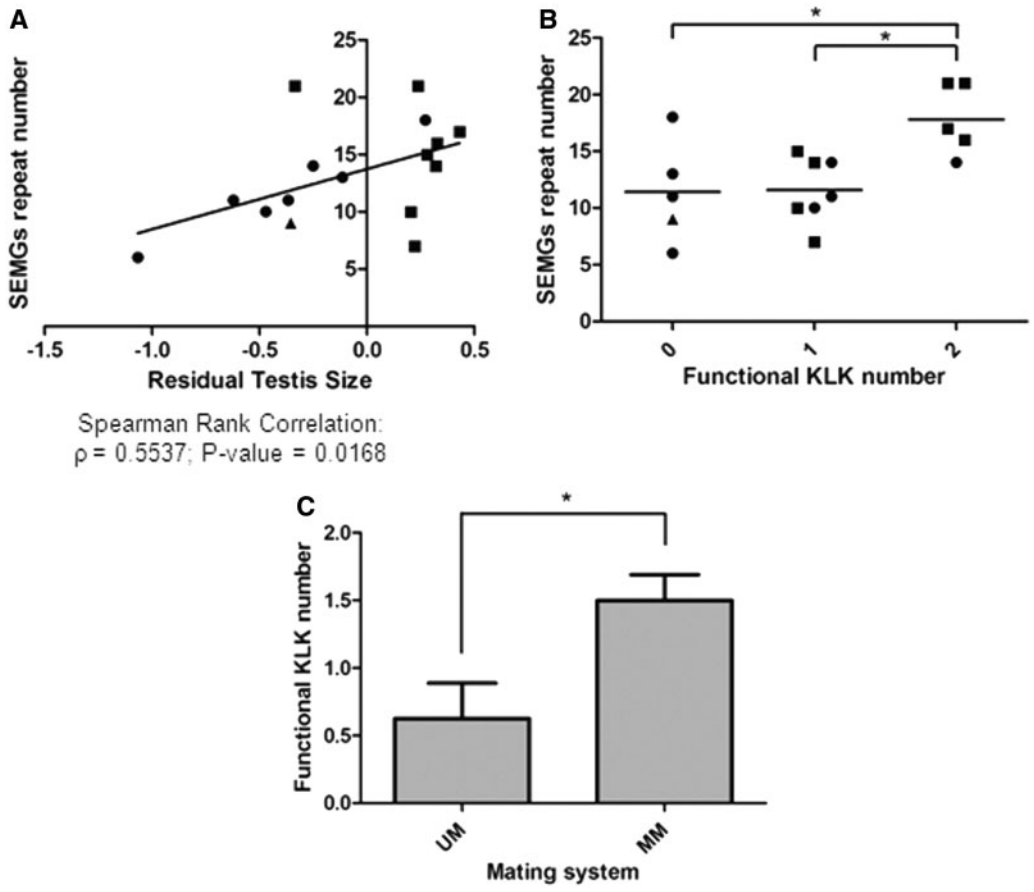

Fig. 3.—Evolution of primate KLK2 and KLK3 related to mating factors. (*A*) Correlation of residual testis size (Anderson et al. 2004; Dixson and Anderson 2004; Wlasiuk and Nachman 2010) with the combined SEMG repeat units (Jensen-Seaman and Li 2003; Hurle et al. 2007). (*B*) Correlation between the number of SEMG1 and SEMG2 repeat units (Jensen-Seaman and Li 2003; Hurle et al. 2007) and the presence of functional KLK2 and KLK3. $*P < 0.05$. (*C*) Correlation between the mating system (Wlasiuk and Nachman 2010) and the presence of functional KLK2 and KLK3. UM, unimale; MM, multimale. $*P < 0.05$. (●), monoandrous; (■), polyandrous; and (▲), ambiguous.

Interestingly, we also observed a trend for higher KLK numbers with more prominent semen coagulation and increase residual testis size (supplementary fig. S4, Supplementary Material online).

We propose a model in which *KLK2* and *KLK3* coevolved with *SEMGs* in a sperm competition-driven process. In a polyandrous species with many SEMG repeats and prominent semen coagulation, the robust and orchestrated activity of KLK2 and KLK3 may be important for sperm release. Conversely, in a monoandrous species with few SEMG repeats, the loss of *KLK2* might represent a biological response to maintain a gelatinous coagulum. Here, the loss of *KLK2* may not be arbitrary because KLK3 has a larger spectrum of cleavage sites in SEMGs than KLK2 (Rawlings et al. 2012), therefore being more effective in semen coagulum liquefaction.

Overall, our data provide support for the occurrence of an event of gene birth by duplication linked to the origin of *KLK3* in a common ancestor of Catarrhini and to an adaptive process associated to the expanded spectrum of KLK3 proteolysis. It further points to multiple events of *KLK2* death through different genomic mechanisms: Unequal crossing-over between *KLK3* and KLK*2* led to the loss of *KLK2* and to the rise of a *cKLK*, whereas large deletions caused the excision of *KLK2* and relaxation of selective constraints led to *KLK2* pseudogenization. In spite of the proposed specialized role of KLK2 and KLK3 in the cascade of seminal plasma liquefaction, their substrate affinity to arginine and common patterns of expression suggest some level of redundancy for *KLK2*; however, such an argument would not explain the loss of *KLK2* observed in more ancient primate species or the skewed activity of *KLK2* in Catarrhini.

Materials and Methods

The genomic sequences from *H. sapiens, Pan troglodytes, P. paniscus, G. gorilla, Pongo pygmaeus, M. mulatta, M. fascicularis,* and *Callithrix jacchus* were retrieved from

public databases. The genomic sequences from *N. leucogenys, M. fuscata, Pap. anubis, Chlorocebus aethiops, C. guereza, Saimiri boliviensis, A. nancymaae, Cal. moloch, Ateles geoffroyi, Eulemur macaco, Lemur catta*, and *Otolemur garnettii* were obtained by Sanger-based shotgun sequencing (supplementary table S1, Supplementary Material online). BAC clones spanning the KLK2–KLK3 genomic fragment were isolated from the following libraries (see http://bacpac.chori.org, last accessed December 10, 2012), as described (Thomas et al. 2002, 2003): *P. troglodytes* (CHORI-251), *Sai. boliviensis* (CHORI-254), *Ate. geoffroyi* (UC-1), *N. leucogenys* (CHORI-271), *Cal. moloch* (LBNL-5), *C. guereza* (CHORI-272), *Pon. pygmaeus* (CHORI-253), *Pap. anubis* (RPCI-41), and *Chl. aethiops* (CHORI-252). Specifically, each library was screened using pooled sets of oligonucleotide-based probes designed from the established sequence of KLK locus. After isolation and mapping, BACs were shotgun sequenced on an ABI 3130 automated sequencer and subjected to sequence finishing, as described (Blakesley et al. 2004). KLK genes were annotated based on alignments to human RefSeq cDNA and protein sequences with the BATI algorithm (Blast, Annotate, Tune, Iterate) using four Perl scripts—Tbex, BlastSniffer, GeneTuner, and bgmix—available at http://degradome.uniovi.es/downloads.html (last accessed December 10, 2012). Briefly, BATI allows the annotation in the target genome of all orthologs and paralogs from the input set of cDNA and protein sequences. Tbex compares all the input sequences with the target genomic sequence by tblastn. BlastSniffer rebuilds each putative gene from the tblastn hits considering all the possible hit combinations and sets a raw score for each of them. GeneTuner shows the result from the previous step in the context of the template genome allowing the user to define exon/intron boundaries. Finally, bgmix creates a composite file with all the tblastn comparisons and highlights those hits overlapping defined exons. This helps the identification and annotation of novel putative genes that have not been annotated in an iterative process.

The genomic sequences spanning the *SEMG1–SEMG2* cluster were retrieved from Hurle et al. (2007) with the exception of *N. leucogenys* and *Cal. moloch* (AC198263 and AC207864, respectively), which were sequenced according to the methods described earlier for this study.

Maximum-likelihood estimates of d_N/d_S (ω) were carried out using the codeml program from the software package Phylogenetic Analysis by Maximum Likelihood—PAML version 4.2 (Yang 2007). To run PAML, we first reconstructed a phylogenetic tree using all the sequences except for *Hylobates* sp. and *L. catta* whose sequences were incomplete. To carry out a comprehensive analysis of pseudogenes, their sequences were only included after the removal of positions affected by premature stop codons and frameshift mutations. The phylogenetic tree was built using the maximum-likelihood method, implemented in DNAml, from the software package Phylogeny Inference Package (PHYLIP; http://evolution.genetics.washington.edu/phylip.html). The tree was consistent with the known primate phylogeny. To test for variable selective pressures among branches, we performed the branch model using either the null model (one ratio) or nested models (two-ratio, three-ratio, and four-ratio models) (Yang 1998; Bielawski and Yang 2003). The values of $\omega > 1$ were considered as evidences of positive selection, the values of $\omega < 1$ were regarded as an indication of purifying selection, and the values of $\omega \sim 1$ were inferred as neutral. The significance of each nested model was obtained from twice the variation of likelihoods ($-2\Delta l$) using a χ^2 statistic. To evaluate lineage-specific changes at amino acid sites, we performed the branch-site model for the anc*KLK3*. This model assumes that the branches on the phylogeny are divided a priori into foreground (anc*KLK3)* and background (remaining branches in the phylogeny) and allows ω to vary both among sites in the protein and across branches. For the branch-site model (Yang and Nielsen 2002), comparisons with critical χ^2 were carried out as described (Zhang et al. 2005). To test for variation in ω between sites of *KLK3* and functional *KLK2,* we used different codon models for each gene alone and compared neutral and selection models: M1–M2 and M7–M8 (Nielsen and Yang 1998; Yang et al. 2000). The Bayes empirical Bayes was used to calculate posterior probabilities of site classes, to identify sites under positive selection for the significant likelihood ratio tests (Yang et al. 2005). KLK3 three-dimensional model (2ZCH.pdb) was retrieved from RCSB PDB Protein Data Bank (http://www.rcsb.org/pdb/home/home.do). KLK2 three-dimensional model was generated by SwissModel (http://swissmodel.expasy.org/workspace) using KLK2 and KLK3 human sequences and KLK3 three-dimensional model (2ZCH.pdb) (Schwede et al. 2003).

Statistical analysis was performed by means of *t*-test, analysis of variance, and Spearman rank correlation.

Supplementary Material

Supplementary tables S1 and S2 and figures S1–S4 are available at *Genome Biology and Evolution* online (http://gbe.oxfordjournals.org/).

Acknowledgments

The authors thank the Lisbon Zoo for their collaboration in providing the primate samples. This work was supported by a fellowship SFRH/BD/68940/2010 from the Portuguese Foundation for Science and Technology (FCT) to P.I.M., by POPH-QREN—Promotion of Scientific Employment, by the European Social Fund, and by national funds of the Ministry of Education and Science to P.I.M. and S.S., and in part by the Intramural Research Program of the National Human Genome Research Institute. IPATIMUP (an Associate Laboratory of the Portuguese Ministry of Education and Science) is partially supported by FCT.

Literature Cited

Adzhubei IA, et al. 2010. A method and server for predicting damaging missense mutations. Nat Methods. 7:248–249.

Anderson MJ, Hessel JK, Dixson AF. 2004. Primate mating systems and the evolution of immune response. J Reprod Immunol. 61:31–38.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics. 3: 201–212.

Blakesley RW, et al. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res. 14: 2235–2244.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. PLoS Genet. 1:e35.

Debela M, et al. 2006. Specificity profiling of seven human tissue kallikreins reveals individual subsite preferences. J Biol Chem. 281:25678–25688.

Demuth JP, Hahn MW. 2009. The life and death of gene families. Bioessays 31:29–39.

den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat. 15:7–12.

Dixson AF, Anderson MJ. 2004. Sexual behavior, reproductive physiology, and sperm competition in male mammals. Physiol Behav. 83:361–371.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nat Genet. 36:1326–1329.

Emami N, Diamandis EP. 2007. New insights into the functional mechanisms and clinical applications of the kallikrein-related peptidase family. Mol Oncol. 1:269–287.

Harcourt AH, Harvey PH, Larson SG, Short RV. 1981. Testis weight, body weight, and breeding system in primates. Nature 293:55–57.

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Hurle B, Swanson W, Program NCS, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res. 17:276–286.

Janssen S, et al. 2004. Screening a combinatorial peptide library to develop a human glandular kallikrein 2-activated prodrug as targeted therapy for prostate cancer. Mol Cancer Ther. 3:1439–1450.

Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. J Mol Evol. 57: 261–270.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res. 20:1313–1326.

Lovgren J, Airas K, Lilja H. 1999. Enzymatic action of human glandular kallikrein 2 (hK2). Substrate specificity and regulation by Zn2+ and extracellular protease inhibitors. Eur J Biochem. 262:781–789.

Lundwall A, Brattsand M. 2008. Kallikrein-related peptidases. Cell Mol Life Sci. 65:2019–2038.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121–152.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.

Olsson AY, Valtonen-Andre C, Lilja H, Lundwall A. 2004. The evolution of the glandular kallikrein locus: identification of orthologs and pseudogenes in the cotton-top tamarin. Gene 343:347–355.

Pavlopoulou A, Pampalakis G, Michalopoulos I, Sotiropoulou G. 2010. Evolutionary history of tissue kallikreins. PLoS One 5:e13781.

Rawlings ND, Barrett AJ, Bateman A. 2012. MEROPS: the database of proteolytic enzymes, their substrates, and inhibitors. Nucleic Acids Res. 40:D343–D350.

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 31: 3381–3385.

Thomas JW, et al. 2002. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. Genome Res. 12: 1277–1285.

Thomas JW, et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793.

Valtonen-Andre C, Olsson AY, Nayudu PL, Lundwall A. 2005. Ejaculates from the common marmoset (*Callithrix jacchus*) contain semenogelin and beta-microseminoprotein but not prostate-specific antigen. Mol Reprod Dev. 71:247–255.

Wlasiuk G, Nachman MW. 2010. Promiscuity and the rate of molecular evolution at primate immunity genes. Evolution 64:2204–2220.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.

Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22:1107–1118.

Yousef GM, Diamandis EP. 2001. The new human tissue kallikrein gene family: structure, function, and association to disease. Endocrine Rev. 22:184–204.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

Associate editor: George Zhang