

Pervasive Adaptive Evolution in Primate Seminal Proteins

Nathaniel L. Clark*, Willie J. Swanson

Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

Seminal fluid proteins show striking effects on reproduction, involving manipulation of female behavior and physiology, mechanisms of sperm competition, and pathogen defense. Strong adaptive pressures are expected for such manifestations of sexual selection and host defense, but the extent of positive selection in seminal fluid proteins from divergent taxa is unknown. We identified adaptive evolution in primate seminal proteins using genomic resources in a tissue-specific study. We found extensive signatures of positive selection when comparing 161 human seminal fluid proteins and 2,858 prostate-expressed genes to those in chimpanzee. Seven of eight outstanding genes yielded statistically significant evidence of positive selection when analyzed in divergent primates. Functional clues were gained through divergent analysis, including several cases of species-specific loss of function in copulatory plug genes, and statistically significant spatial clustering of positively selected sites near the active site of kallikrein 2. This study reveals previously unidentified positive selection in seven primate seminal proteins, and when considered with findings in *Drosophila*, indicates that extensive positive selection is found in seminal fluid across divergent taxonomic groups.

Citation: Clark NL, Swanson WJ (2005) Pervasive adaptive evolution in primate seminal proteins. PLoS Genet 1(3): e35.

Introduction

Studies of adaptive evolution have revealed multiple classes of reproductive proteins under positive selection, including those involved in gamete recognition, seminal fluid factors, and proteins in the female reproductive tract [1–5]. The unknown pressures driving this adaptive evolution may be shared among taxonomic groups. For example, evidence of positive selection in gamete recognition proteins is found across divergent taxonomic groups, including mollusks, echinoderms, green algae, and mammals [1,4–7]. Positive selection in seminal proteins is observed in *Drosophila* and in primate semenogelin proteins [2,8–10]. However, the extent of selection in primates remains unknown, and it has not been determined whether seminal fluid proteins in divergent taxa experience such adaptive pressures.

Seminal fluid proteins in *Drosophila* initiate striking reproductive responses in females [11]. Inseminated proteins have been shown to affect sperm storage in the female reproductive tract, copulatory plug formation, ovulation, oogenesis, female receptivity to re-mating, and female lifespan. These can be important effects for sperm competition and sexual conflict, both of which may drive adaptive evolution. Additional seminal factors show antibacterial activity and may serve in pathogen defense, another adaptive driving force.

There is reason to believe that similar forces act on primate seminal fluid, leaving signatures of positive selection. As in drosophilids, several primate species form a post-mating copulatory plug, which could serve in sperm competition by excluding subsequent ejaculates from competing males. Plugs are present in diverse primates [12], including prosimians, New World and Old World monkeys, and in the chimpanzee, the closest living relative to humans. Consistent with adaptation, positive selection is seen in semenogelin proteins functioning in this pathway [9,10,13].

To identify positive selection in primate proteins, we used a measure of selective pressure, the d_N/d_S ratio [14,15].

Positive selection for amino acid diversification results in the rate of nonsynonymous substitutions exceeding that of synonymous substitutions. This effect is measured on coding sequences as the nonsynonymous substitution rate divided by the synonymous substitution rate—the d_N/d_S ratio. A value greater than one is indicative of positive selection, and a value less than one indicates purifying selection. In the absence of selection (neutral evolution), a value of one is expected. When measured over the entire length of a gene, the d_N/d_S ratio is a conservative measure of positive selection; in the presence of strong positive selection at some sites, conservation at others will lower the ratio. For this reason, when measured over the entire gene, we consider an elevated d_N/d_S to be suggestive of positive selection acting on a portion of the gene [3]. We then measure statistical significance using multiple species alignments through a likelihood method (CODEML) allowing different d_N/d_S values for codon sites [5,16,17]. Criticisms of this method include the argument that it gives false positive results under certain parameter combinations [18]. A more extensive study conducted by Wong et al. [19] found that this problem was limited to an early version of the program and to problems with convergence, and that the maximum likelihood method has good power and accuracy in detecting positive selection.

In this study we aimed to determine the extent of positive selection in primate seminal fluid proteins and to character-

Received April 13, 2005; Accepted August 4, 2005; Published September 16, 2005
DOI: 10.1371/journal.pgen.0010035

Copyright: © 2005 Clark and Swanson. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: TG, transglutaminase

Editor: Takashi Gojobori, National Institute of Genetics, Japan

*To whom correspondence should be addressed. E-mail: clarknl@u.washington.edu



Synopsis

Proteins found in seminal fluid accompanying sperm show dramatic effects on reproduction, such as manipulating female behavior. Even in primates they participate in competition between sperm of different males, and serve to protect sperm from infection by pathogens. These types of roles require the proteins to constantly adapt to stay ahead of the competition. Such adaptive pressures on proteins leave characteristic signatures in the DNA sequences that encode them. The authors used these signatures to identify adaptive evolution in primate seminal proteins and found extensive signs of adaptation when comparing thousands of seminal genes between human and chimpanzee. They further characterized outstanding genes in several primate species, including a diversity of apes and monkeys. Several of these proteins have no known function, yet by visualizing the adaptation on their three-dimensional surfaces, the authors uncovered clues to what is driving their evolution. In addition, they found several cases in which certain species lost their functional copies of these genes. Interestingly, species that showed loss of function do not participate in sperm competition. Past studies found widespread adaptation in fruit fly seminal fluid, and this study reveals extensive adaptation in primate seminal proteins. Could this be a phenomenon common among animals?

ize outstanding candidates in further detail. Eight candidates were chosen from a pairwise human-chimpanzee $d_{N/dS}$ screen of seminal proteins. More detailed analysis with several species sequences provided strong evidence that positive selection acts on several primate seminal proteins.

Results

A Selective Pressure Screen

A list of proteins present in human seminal fluid was compiled from mass spectrometry studies of seminal plasma and prostasomes [20,21]. A total of 161 proteins were identified, 129 in prostasomes and 43 in seminal plasma, with 11 found in both studies. Human coding regions for these genes were aligned with chimpanzee orthologous sequences in order to estimate selective pressure between these two lineages as indicated by the $d_{N/dS}$ ratio. Estimates of pairwise $d_{N/dS}$ ratios were calculated using both the Nei and Gojobori method and a maximum likelihood method [14,16]. Both gave similar estimates (Table S1).

Rates of nonsynonymous versus synonymous substitution in these genes revealed several coding sequences with elevated $d_{N/dS}$ ratios (Figure 1A). Of 161 seminal fluid proteins, 17 had a $d_{N/dS}$ greater than one, and 36 greater than 0.5. The median $d_{N/dS}$ value was 0.19. A study of *Drosophila* seminal proteins showed similar variation of selective pressure (Figure 1B) [2]. Primate genes with elevated $d_{N/dS}$ ratios were involved in immune response (*complement component 7, interleukin 1 receptor-like 2*), semen coagulum (*semenogelins I and II, prostate-specific transglutaminase 4, prostatic acid phosphatase*), cellular structure (*desmoglein 1, profilin 1*), and other roles, including several proteins of unknown function. Results for all 161 genes are shown in Table S1.

Secreted proteins may tend to have higher $d_{N/dS}$ values, as they encounter adaptive pressures from exterior forces, such as interactions in the female reproductive tract. The subset of 43 proteins with secretion signal sequences showed a higher mean $d_{N/dS}$ (0.30) than those without (0.15). This difference is

significant as determined by permutations ($p = 0.0091$). This 2-fold increase for secreted proteins was also observed in *Drosophila* seminal fluid [2].

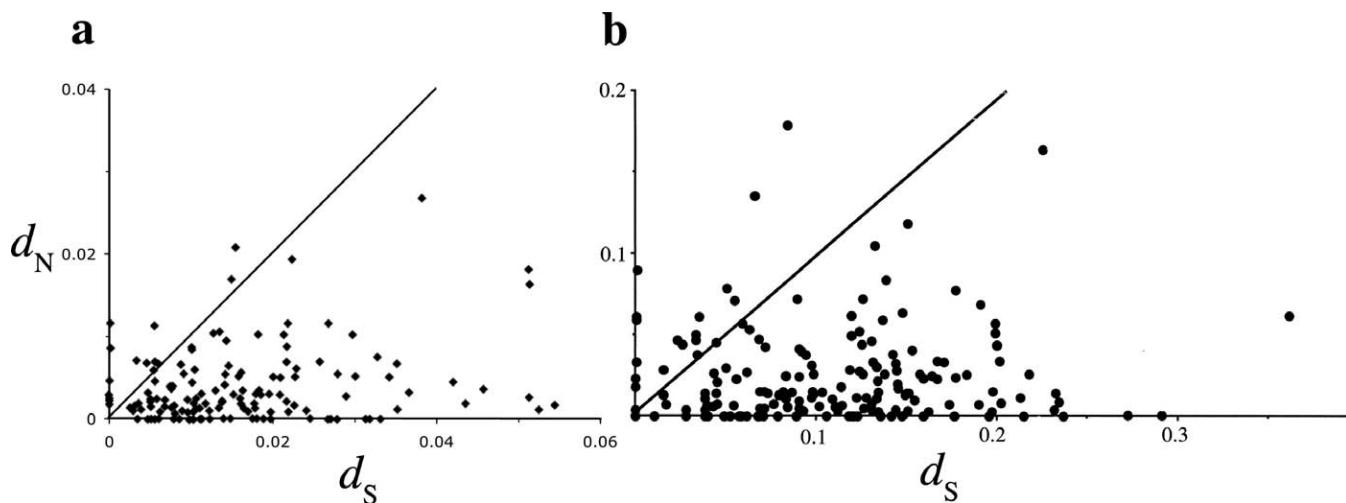
Since the mass spectrometry studies are not expected to provide an exhaustive catalog of seminal fluid proteins, a similar screen was performed on prostate-expressed genes identified from an expression study of noncancerous human prostate [22]. Of 2,858 prostate-expressed genes, 290 showed a $d_{N/dS}$ greater than one, while the median value was 0.15. Secreted proteins again showed a higher mean $d_{N/dS}$ (0.24 versus 0.17, $p = 0.00038$).

The pairwise estimates were aimed at predicting candidate genes under selection, and eight were selected for in-depth analysis. Criteria used to select these candidates were a high $d_{N/dS}$ value, a high d_N , and evidence for high or specific prostate expression. Seven of eight candidates were taken from the mass spectrometry set because of the direct evidence that they are present in ejaculate. One gene, *kallikrein 2 (KLK2)*, was taken from the prostate-expressed set, since it has a known role in seminal fluid dynamics. An additional gene, *prostate-specific antigen (PSA)*, was analyzed due to its importance in copulatory plug dissolution, despite its low pairwise $d_{N/dS}$ value (Table 1). Since it was not chosen from the screen results, *PSA* is not considered a candidate. Overall, we chose three candidate genes involved in semen coagulation and five candidates whose functions are unknown.

Positive Selection in Candidate Genes

To assess statistical significance of positive selection in candidate genes, we sequenced primate coding regions to provide eleven species sequences on average. We then assessed the selective pressure acting on these sequences using $d_{N/dS}$ ratios. Using a method that predicts a uniform $d_{N/dS}$ ratio across all codon sites, several pairwise comparisons of *prolactin-induced protein (PIP)* and *β-microseminoprotein (MSMB)* sequences have $d_{N/dS}$ ratios significantly greater than one, suggesting positive selection (unpublished data) [23]. This is a conservative approach, since it is unlikely that all codon sites are subject to the same selective pressure during evolution. More sensitive methods allow testing for variation in $d_{N/dS}$ at codon sites by comparing neutral models to selection models of codon evolution. Model parameters were estimated using a maximum likelihood method employed in the CODEML program of the PAML package [5,16,17]. For each gene, three different comparisons of neutral and selection models gave similar results (M1 versus M2, M7 versus M8, and M8A versus M8). From these comparisons, significant signs of positive selection were found in seven of eight candidate genes (Table 1). Since candidates were chosen based on high human-chimpanzee $d_{N/dS}$ values, there could be a statistical bias when sequences from the initial screen are included in the multiple alignments. When human and chimpanzee sequences were removed, six of the seven remained statistically significant, showing positive selection. The analysis that failed this conservative test, that of *prostate-specific transglutaminase 4 (TGM4)*, may have suffered a lack of power, because the total tree length (0.47) was below optimal (~1) for this maximum likelihood method, due to the removal of two taxonomic groups [24].

The codon classes predicted to be under positive selection had $d_{N/dS}$ values ranging from 2 to 14 and were estimated to contain large proportions of codons for some genes (*MSMB, PIP*) and smaller proportions for others (*TGM4*) (Table 1). The

**Figure 1.** Plots of d_N Versus d_S for Primate and *Drosophila* Seminal Fluid Genes

(A) Genes encoding seminal fluid proteins identified by mass spectrometry in human versus chimpanzee.

(B) *Drosophila simulans* male-specific accessory gland genes versus *D. melanogaster* [2].The diagonal represents neutral evolution, a d_N/d_S ratio of one. Most genes are subject to purifying selection and fall below the diagonal, while several genes fall above or near the line suggesting positive selection. Comparison of the two plots shows elevated d_N/d_S ratios in seminal fluid genes of both taxonomic groups.

DOI: 10.1371/journal.pgen.0010035.g001

rapid evolution of *MSMB* was noted in past studies of primate, rodent, and bird sequences [25,26], and was attributed to either low selective constraint or positive selection. We found highly significant signs of positive

selection within primates ($p < 0.001$), with an estimated 42% of codons showing a d_N/d_S ratio of 2.90. Three diversified paralogs of the *MSMB* gene exist in New World monkeys [27], and their functions are unknown. When only Old World

Table 1. Seven Candidate Genes from the Screen Show Signs of Positive Selection

Gene	Functional Information	Human-Chimpanzee d_N/d_S^a	Species in Divergent Analysis ^b	Ratio of d_N/d_S (% Codons) ^c	<i>p</i> -Value: All Species ^d	<i>p</i> -Value: Minus Human and Chimp	Codon Sites under Selection ^e
<i>TGM4</i> (PS)	Formation of copulatory plug, Activation of PSA	2.14 (10/2)	1–3, 5, 6, 8, 9, 13, 16–18	14 (0.3%)	0.0018	NS	582T
<i>KLK2</i> (PS)		1.42 (3/1)	1–3, 5, 9, 13, 15–17, 19	2.11 (24%)	0.0046	0.016	50L, 52K, 91H, 92Q, 108R, 118V, 193G
<i>ACPP</i> (PS)	Dissolution of semen coagulum	0.51 (3/3)	1–6, 9, 13, 16	6.47 (4.6%)	0.0072	0.03	103I, 165L, 245I
<i>PSA</i> (PS)	Dissolution of semen coagulum	0.38 (3/3)	1–5, 8, 9, 11, 13, 15, 16	2.94 (6.7%)	NS	NS	NS
<i>DBI</i>	Unknown	Infinity (2/0)	1–9, 11, 13, 15–18	2.91 (30%)	0.015	0.015	32H, 38S
<i>Synaptic vesicle VAT-1 homolog</i>	Unknown	2.05 (7/1)	1–9, 17, 18	N/A	NS	NS	NS
<i>PIP</i> (HP)	Host defense	1.13 (5/2)	1–9	7.56 (25%)	0.000054	0.00025	10K, 19R, 27V, 29A, 31Q, 48I, 60A, 65D, 90R, 106N, 112L, 116K
<i>TMPRSS2</i> (PS)	Protease, may be cleaved and secreted	0.87 (9/4)	1–11, 16	4.22 (5.1%)	0.0014	0.0052	136G, 255E, 319A, 443S
<i>MSMB</i> (HP)	Sperm-coating, immunoglobulin-binding	0.84 (2/1)	1–6, 8, 9, 14, 16, 17	2.90 (42%)	0.000032	0.000038	9G, 22K, 31E, 39T, 41T, 62D, 88S, 90S, 91E, 94I

Since the initial screen selected for genes with high human-chimpanzee d_N/d_S values, an analysis was performed on divergent sequences excluding those from the initial screen, thus correcting for this potential bias. For genes with significant positive selection, codon sites under selection were predicted using a Bayes Empirical Bayes approach. Since we report only those positively selected sites with high statistical support, there are fewer sites listed than were estimated in the codon model.

^aHuman-chimpanzee pairwise d_N/d_S was determined during the selective pressure screen; values in parentheses are the numbers of substitutions (nonsynonymous/synonymous).^bDivergent analysis used coding sequences from several species. Species key: 1, human; 2, chimpanzee; 3, bonobo; 4, gorilla; 5, orang-utan; 6, *Hylobates syndactylus*; 7, *H. lar*; 8, *H. gabriellae*; 9, *Macaca mulatta*; 10, *M. nemestrina*; 11, *M. fascicularis*; 12, *M. fuscata*; 13, *Papio anubis*; 14, *P. hamadryas*; 15, *Cercopithecus cebus*; 16, *Erythrocebus patas*; 17, *Saguinus labiatus*; 18, *Ateles geoffroyi*; and 19, *Lemur catta*.^cThe d_N/d_S ratio of the class of codons under positive selection is given with the percentage of codon sites predicted to be in that class.^dThe *p*-value rejecting the model of neutral evolution (M8A) over that of selection (M8) is given.^eSites with posterior probabilities greater than 0.9 are indicated in regular type; $p > 0.95$ indicated in bold type; and $p > 0.99$ indicated in bold, underlined type.

N/A, not applicable; NS, not significant; HP, high expression in prostate; PS, prostate-specific expression.

DOI: 10.1371/journal.pgen.0010035.t001



monkey and ape sequences are analyzed, significant positive selection is still observed ($p = 0.029$), and selection is predicted at similar codon sites.

We looked for lineage-specific variation in selective pressure by estimating d_N/d_S along phylogenetic lineages. For *TGM4*, a model estimating independent d_N/d_S ratios for each lineage fit the data better than a model with a uniform ratio ($p = 0.0031$). This indicates that variable selective pressure acted on *TGM4* during its evolution, with branch-specific d_N/d_S values ranging from 0.1 to 1.95 (Figure 2). *Prostatic acid phosphatase (ACPP)* also shows significant variation in d_N/d_S , with elevated values in the chimpanzee and rhesus macaque lineages—1.16 and 0.64, respectively ($p = 0.016$). Finally, *PSA* does not have a high pairwise human-chimpanzee d_N/d_S , but it shows significant variation in selective pressure during its evolution. A branch model shows *PSA* lineages with d_N/d_S ratios exceeding one and was a significantly better fit than a model with uniform ratios for all lineages ($p = 0.004$). The extreme values in all three of these genes could be due to either positive selection or a reduction in functional constraint.

Spatial Distributions of Selected Sites on Three-Dimensional Structures

Positively-selected codon sites were predicted by a Bayes Empirical Bayes method for all genes showing significant positive selection [28]. Observed levels of divergence and number of sequences were appropriate for accurate prediction of sites according to a power analysis of Bayes prediction [29]. The spatial relationship of these selected sites was evaluated by mapping them onto three-dimensional protein structures. This analysis was done to find connections

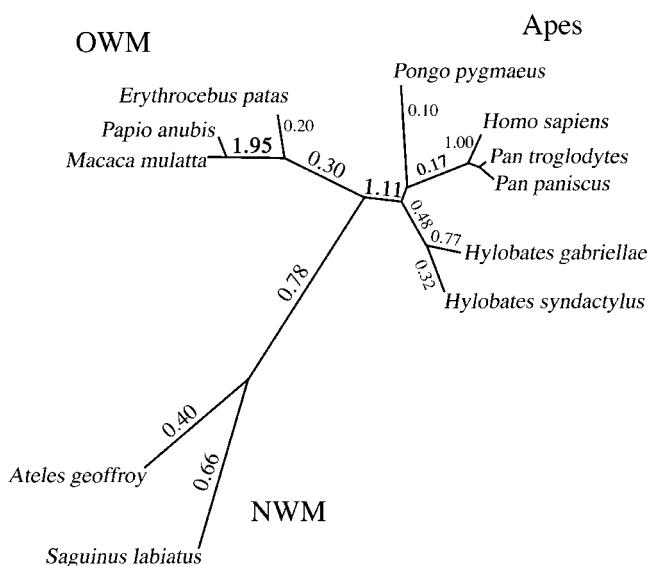


Figure 2. Variable Selective Pressure is Seen Between Lineages for Semen Coagulum Protein *TGM4*

This primate phylogeny shows selective pressure on *TGM4* with estimated d_N/d_S ratios indicated on branches. Ratios greater than one are suggestive of either relaxed constraint or positive selection. Ratios are only shown for long branches, those with at least eight substitutions. A null model with a uniform d_N/d_S ratio across all lineages is rejected in favor of these estimates ($p = 0.003$). Branch lengths are estimated from *TGM4* coding sequences. NWM, New World monkeys; OWM, Old World monkeys.

DOI: 10.1371/journal.pgen.0010035.g002

between positive selection and functional sites, because previous studies of MHC, lysin, and ZP3 proteins showed that predicted sites of positive selection fall into regions or binding clefts where diversification is biologically relevant [4,5]. We mapped selected sites onto five primate seminal proteins employing either solved crystal structures or threaded structural models, and used only predicted sites with a high level of support ($p > 0.9$). The spatial patterns and locations yielded intriguing patterns of positive selection.

Positively selected sites of the *KLK2* protein fall near the active site residues and are found in known functional regions (Figure 3A). One selected site is in a known substrate binding cleft and two are in the kallikrein loop [30]. These locations and the pattern of clustering suggest that there was selective pressure for *KLK2* to change substrate binding affinity. To assess statistical significance of surface clustering, we compared the mean pairwise distance between positively selected sites and a null distribution generated from randomly drawn surface (solvent exposed) sites (Figure 3B). Comparing the observed mean to the null distribution (10,000 permutations) lends statistical support to the hypothesis that these positively selected sites are clustered on the surface of *KLK2* ($p = 0.0043$). The spatial distribution of selected sites on *KLK2* provides an example of how positive selection can lead to inferences about evolution of protein function. In this case, a change in substrate is suggested.

MSMB is one of the most abundant proteins in human seminal plasma, yet its function remains unknown. It is also evolving rapidly, with an estimated 42% of codons under positive selection (Table 1). Positively selected sites are found all over the exterior of a threaded structure of *MSMB* (Figure 4), in contrast to the clustering seen on *KLK2*. When a clustering test is performed on *MSMB* selected sites, the observed mean distance falls just short of being significantly dispersed ($p = 0.066$). This dispersed pattern suggests that selection has acted uniformly on the surface of *MSMB* and no distinct functional regions can be inferred.

Because few sites could be mapped onto structural models of transmembrane serine protease 2 (TMPRSS2), ACPP, and acyl-coA-binding protein (DBI), spatial distributions were less distinct, and clustering was not seen; however, some functional hypotheses may be made. Selected sites were predicted in two domains of TMPRSS2, the serine protease and the low-density lipoprotein receptor. When selected sites in the protease domain are mapped onto a threaded three-dimensional structure, they all occupy exterior positions on the same face, opposite of the protease active site. No interactions are confirmed in this region, but TMPRSS2 is thought to be activated through cleavage at a site located on this face [31]. The selective pressure on ACPP may be related to its substrate; a selected surface site (V77) neighbors two active site residues (R79 and H257) in the solved crystal structure [32]. Although this selected site was only moderately supported in Bayes Empirical Bayes analysis ($p = 0.824$), it is intriguing because of its proximity to the active site in an otherwise conserved region.

Structural analysis of selected sites and biochemical characterization are complementary approaches for elucidating the biological roles of these proteins. For example, our evidence of selection in *KLK2* implies a testable change in substrate binding during primate evolution. As more coding sequences are determined, prediction of selected sites will

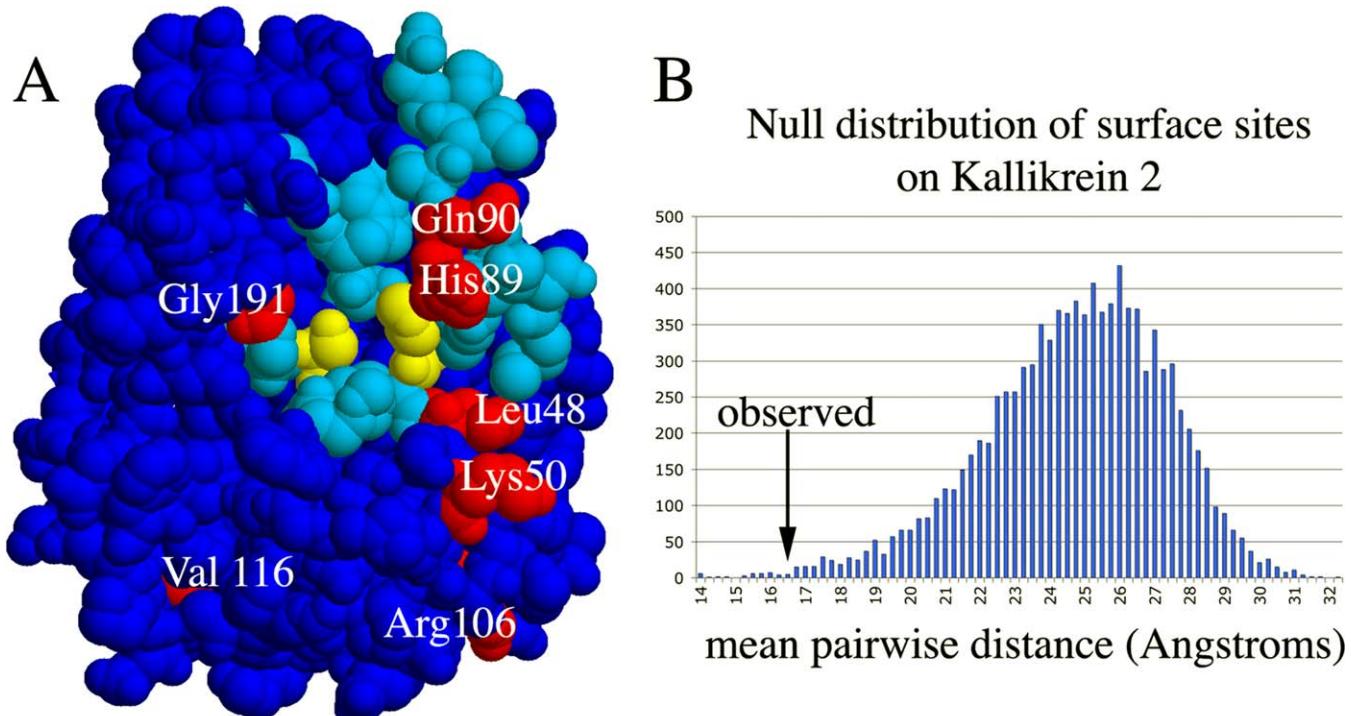


Figure 3. Positive Selection at Sites Involved in Substrate Binding in KLK2

(A) Several amino acid sites predicted to be under positive selection (red) are near the protease active site (yellow). Three selected sites are found in known structural components of kallikrein proteins (light blue residues): Gly191 is part of the S1 substrate binding pocket, and His89 and Gln90 are part of the kallikrein loop [30]. Selected sites are labeled with the human residue on this threaded model.

(B) Positively selected sites are significantly clustered on the surface of KLK2. The observed mean pairwise distance between predicted positively selected sites is significantly lower than random sets of surface sites ($p = 0.0043$). This spatial clustering suggests that positive selection acted during KLK2 evolution to alter substrate binding.

DOI: 10.1371/journal.pgen.0010035.g003

improve, allowing site-specific selective pressures to be evaluated in functional contexts.

Loss of Function

Evidence for loss of function in several species was seen for two candidate genes, *TGM4* and *KLK2*. Interestingly, both of these genes are involved primate semen coagulation. Prostate-specific TGM4 forms semen coagulum and copulatory plugs through cross-linking by its transglutaminase (TG) domain. Sequence from gorilla showed a homozygous, 11-basepair deletion in exon 7 at the start of the TG domain, a frameshift that would lead to early termination at amino acid 293 of 684. This deletion is likely fixed in gorilla populations, since four additional gorillas showed the same homozygous deletion. Abrogation of transglutaminase activity is likely since this exon contains ~20% of the TG domain and the remaining 80% falls downstream. Similarly, the sequenced *Hylobates lar* individual was homozygous for an early stop at codon position 411 downstream of the TG domain and before the first transglutaminase C-terminal domain.

In *KLK2*, the *Macaca mulatta* individual showed a homozygous change altering the active site residue D120 to alanine, which would eliminate proteolytic activity. This change was not seen in the four other Old World monkeys examined, including *Macaca nigra*. This suggests that abrogation of KLK2 activity occurred in those macaques closely related to *M. mulatta* or in that species alone.

Other evidence suggests loss of the *KLK2* gene from gorilla and lesser apes. Although *KLK2* sequence was obtained in

several divergent New and Old World monkeys, only the first and last exons (1 and 5) were obtained from gorilla, despite several amplification conditions and primer combinations. When conditions were relaxed, PCR products from exons 2 through 4 of the paralog *PSA* were obtained instead, suggesting that *KLK2* was lost in the gorilla lineage. Similar difficulties were encountered in all three analyzed species of

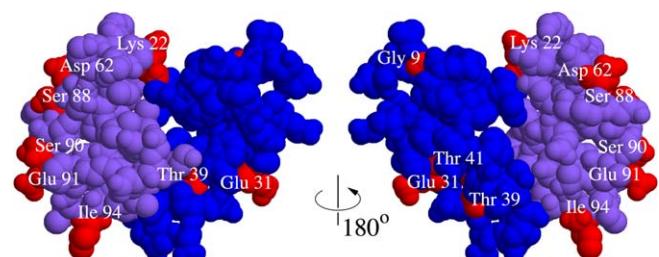


Figure 4. Positively Selected Sites on MSMB are Spread across the Protein Surface.

According to sites models of codon evolution, 42% of MSMB residues experienced adaptive pressure to alter their amino acids. Those predicted with high support are shown in red on this threaded structural model of human MSMB. Blue and purple residues demarcate two structural domains of the protein [62]. The amino acid sites show no clustering and are almost significantly dispersed throughout the protein ($p = 0.066$). This pattern is quite different from that shown by KLK2 (Figure 3). Although MSMB is one of the most abundant human seminal proteins, its function remains unknown.

DOI: 10.1371/journal.pgen.0010035.g004

genus *Hylobates*, suggesting a similar loss in lesser apes. *PSA* and *KLK2* are paralogous genomic neighbors and likely arose through tandem duplication, so that unequal crossing-over could lead to deletion of one of the paralogs.

Discussion

Several seminal fluid proteins show dynamic evolutionary histories, significant positive selection, and variable selective pressure between lineages. Multiple instances of loss of function also hint at changing selective pressure. Seminal protein adaptation could result from several potential pressures, including sexual selection, pathogen response, and coevolution with changing binding partners and substrates. It is hypothesized that sexual selection, namely sperm competition and sexual conflict, is a major driving force behind the adaptive evolution of *Drosophila* seminal fluid proteins [33], and could be responsible for primate divergence as well.

Copulatory Plug Candidate Genes

In some primate species, the degree of semen coagulation is high enough to form a firm copulatory plug, a mechanism of sperm competition. Four prostate-specific candidate genes (*TGM4*, *KLK2*, *PSA*, and *ACPP*) participate in formation or dissolution of human seminal coagulum [34,35]. Significant positive selection is seen in *TGM4*, *KLK2*, and *ACPP*, along with significant variation in selective pressure between lineages for *TGM4*, *ACPP*, and *PSA*. Additionally, both of the genes showing loss of function participate in the formation (*TGM4*) or dissolution (*KLK2*) of semen coagulum. Loss of function of gorilla *TGM4* is consistent with the lack of semen coagulation in gorilla [12] and with past evidence of early stop codons in alleles of gorilla *semenogelins I* and *II* [9,10]. Degeneration of semen coagulation may also be occurring in the lar gibbon, as its *TGM4* coding sequence shows an early stop codon. Loss of semen coagulation is consistent with the mating systems of gorillas and gibbons, since both species are considered monoandrous, so males are not competitive postmating.

After a copulatory plug has been set, breaking it down is a strategy for competing males to win fertilizations. Positive selection seen in *ACPP* and *KLK2* could be due to optimization of this function. *KLK2* proteolytically activates *PSA*, a protease that breaks down semen coagulum. The likely loss of function of *KLK2* observed in the rhesus macaque, gorilla, and lesser apes could result in reduced ability to dissolve semen coagulum. This change could reflect either lack of constraint for this function or adaptive value.

Conflict over Sperm Levels

Human seminal fluid factors, such as prostaglandin E, can locally suppress female immune response [36]. This function may be related to conflict between males and females over sperm levels. As sperm competition leads to higher sperm levels, chances of polyspermy increase, causing females to limit sperm numbers and strengthen barriers to fertilization. Candidate genes *TGM4* and *MSMB* could serve to protect sperm from immune attack in the reproductive tract; evidence suggests that they both bind to sperm surfaces. *TGM4* may deter attack by altering the sperm surface [37,38], and *MSMB* was found to be the main immunoglobulin binding factor in human seminal plasma [39,40]. Hence, these proteins may play roles in suppressing immune response against sperm, resulting in positive selection. Although highly

expressed in human prostate, *MSMB* is also found in other mucous tissues [41], so a role in general pathogen defense must also be considered.

Pathogen Response

Like other secretions, seminal fluid contains protective antipathogenic factors. One candidate, *PIP*, has a likely role in host defense. *PIP* shows strong signs of positive selection, with 25% of codons estimated at a high d_N/d_S of 7.56. Notably, when just apes and one Old World monkey are analyzed, *PIP* shows highly significant positive selection ($p = 0.00024$). This secreted aspartyl proteinase is expressed at high levels in prostate and other exocrine glands. It is thought to play a role in host defense by binding bacteria, and it may suppress T-cell apoptosis [42,43]. Protection of sperm and the male reproductive tract from pathogens may also drive divergence of other seminal fluid proteins.

Antagonistic Pleiotropy

The major source of seminal fluid proteins is the prostate, a common site of male cancer. Disease research may benefit from studies of selection, since positive selection is often associated with human disease genes. In an analysis of 7,645 genes, those showing signs of positive selection were overrepresented in genes associated with disease in the Online Mendelian Inheritance in Man catalog [44]. In addition, Nielsen et al. found several genes involved in tumor suppression and apoptosis among those showing the strongest signs of positive selection between human and chimpanzee [45]. Also, the cancer susceptibility genes *BRCA1* and *angiogenin* show signs of positive selection [46–49]. Such selection could result in antagonistic pleiotropy, a phenomenon in which adaptation in one respect brings deleterious effects in another. It is important to explore the possibility that adaptive evolution of seminal fluid factors contributes to disease through pleiotropic effects. Adaptation in prostate-expressed genes may benefit primates during their reproductive lifespan, but could lead to damaging side effects in later life.

Screen Utility

Overall, this human-chimpanzee selective pressure screen was successful in identifying seminal fluid genes with significant signs of positive selection. This is notable because the screen compared two closely related species with relatively few nucleotide differences per gene. Since seven of eight candidate genes showed statistically significant positive selection, we expect a fraction of other genes with elevated pairwise d_N/d_S ratios to be under positive selection.

Conclusion

The lower limit for primate seminal fluid proteins under positive selection is nine, from seven proteins in this study plus *semenogelins I* and *II*. Given the rate of support in this study, we speculate that there are others from the set of screened genes, as well as from genes not included in this screen. In conclusion, primate seminal fluid contains several proteins that exhibit dynamic evolutionary histories involving positive selection and loss of function. Extensive adaptive evolution in seminal proteins may be common to internally fertilizing taxa, since evidence of positive selection is seen in both *Drosophila* and primates.

Materials and Methods

Selective pressure screen. A list of 161 proteins identified in human seminal fluid was compiled from mass spectrometry studies of

seminal plasma and prostasomes [20,21]. Prostate-expressed genes were identified from an expression study of whole normal prostate (NCI CGAP Pr22) from the Prostate Expression Database (<http://www.pedb.org>) [22]. Of 4,277 unique ESTs from the study, 2,858 were traced to unique accession numbers in the reference sequence (RefSeq) database. Human exons encoding these seminal fluid proteins and prostate-expressed genes were retrieved from the UCSC Table Browser (<http://genome.ucsc.edu/cgi-bin/hgTables>). Each human exon was aligned to the best BLASTN hit over a threshold of 1×10^{-10} from chimpanzee whole genome shotgun contigs [50], and coding sequence alignments were created for evolutionary analysis.

Pairwise values of d_N/d_S for each human-chimpanzee coding sequence alignment were estimated by CODEML of the PAML package [17]. It was noted that some perceived substitutions resulted from poor-quality chimpanzee sequence, so substitution base calls in all gene candidates were manually verified in raw sequence chromatograms found in the sequence reads database [<http://www.ncbi.nlm.nih.gov/Traces/trace.cgi/>]. Presence of a signal sequence was predicted using SignalP [51]. Statistical significance of the difference between secreted versus nonsecreted d_N/d_S values was evaluated by a permutation test comparing the differences between average d_N/d_S values for random subsets through 100,000 permutations.

Sequencing of candidate genes. Coding portions of nine genes were sequenced in divergent primates. Total DNA from the following primates were obtained from the Coriell Institute for Medical Research (Camden, New Jersey, United States): *Pan troglodytes*, *P. paniscus*, *Gorilla gorilla*, *Pongo pygmaeus abelii*, *Macaca mulatta*, *M. nemestrina*, *Erythrocebus patas*, *Saguinus labiatus*, and *Atelos geoffroyi*. DNA samples from the following species were obtained through the Integrated Primate Biomaterials and Information Resource (IPBIR; <http://www.ipbir.org/>): *Hylobates gabriellae*, *H. lar*, *H. syndactylus*, *Cercopithecus cephus*, and *Papio anubis*. Gorilla DNA samples were a generous gift from Evan Eichler at the University of Washington, Seattle, Washington, United States. Sequences of the following genes were obtained from GenBank: *M. fascicularis PSA*, *Papio hamadryas MSMB*, *Saguinus oedipus MSMB*, and *M. fuscata PIP*. Human coding sequences were taken from RefSeq entries in the UCSC genome browser (<http://genome.ucsc.edu>). Lemur *KLK2* exons were retrieved from a BAC clone (GenBank accession number AC153325) sequenced by the NIH Intramural Sequencing Center (<http://www.nisc.nih.gov>).

PCR was used to amplify exon-containing fragments from total DNA of various primates. PCR primers were designed from human introns, and clade-specific primers were designed when possible. PCR conditions and primer sequences are available from the authors upon request. Single-band PCR products were sequenced using Big Dye v.3.1 (Applied Biosystems, Foster City, California, United States). Sequence analysis was done using Phred, Phrap, and Consed [52,53]. High-quality sequence was used to generate coding sequences for each species based on human splice sites. Splice acceptor and donor sites were systematically checked for preservation of GT and AG nucleotides. Multiple alignments were made for each gene using ClustalW [54]. The close relationship between primates allowed for confident multiple alignments with few gaps. For estimation of d_N/d_S at sites or lineages, we removed secretion signal sequences and those species sequences showing loss of function.

Evolutionary analysis. Phylogenetic relationships between the studied primates were taken from published studies [55–57]. Pairwise differences in d_N and d_S were calculated by MEGA version 3.0 [23], using a modified Nei-Gojobori (Jukes-Cantor) codon model with standard error computed analytically. Maximum likelihood evolutionary analysis was done with CODEML of the PAML 3.14 package [17], which estimates parameters for codon models of evolution. In order to ensure correct estimation of model parameters, we checked for convergence by running the optimization multiple times with different starting values of the omega parameter. The omega parameter estimates the d_N/d_S ratio and is used to determine selective pressure on codon sites, where a value greater than one is indicative of positive selection. Statistical significance is determined by a likelihood ratio test comparing a neutral model, where omega is limited to the interval (0, 1), to a selection model with an additional class of codons whose omega value is allowed to be greater than one. Different codon models were used for testing variation in d_N/d_S

between sites; the models were compared as follows: (neutral to selection) M1 to M2, M7 to M8, and M8A to M8. All three comparisons gave similar results; we report those for M8A to M8 in Table 1. Model M1 (neutral) allows two classes of codons, one with omega over the interval (0,1) and the other with an omega value of one. Model M2 (selection) is similar to M1 except that it allows an additional class of codons with a freely estimated omega value. Model M7 (neutral) estimates omega with a beta-distribution over the interval (0, 1), while model M8 (selection) adds parameters to M7 for an additional class of codons with a freely estimated omega value. M8A (neutral) is a special case of M8 that fixes the additional codon class at an omega value of one [58]. Significance of positive selection found among codon sites was estimated both with and without the sequences from the initial screen (human and chimp). Such exclusion avoids a bias of selecting lineages from the screen with high numbers of nonsynonymous substitutions. When significant signs of positive selection were found, specific codon sites subjected to positive selection were predicted using a Bayes Empirical Bayes approach employed in CODEML [28]. Such an approach gives more reliable probability calculations than past methods, since it takes into account sampling errors in estimates of model parameters. To evaluate variation in selective pressure over a phylogeny, the branch model of CODEML estimated d_N/d_S values for each branch. The branch model is compared to the null hypothesis, model M0, in which all lineages have the same d_N/d_S value.

Structural analysis. Threaded protein structures for KLK2, DBI, TMPRSS2, and MSMB were created with SwissModel using human primary sequence [59]. Structural analysis of ACPP sites was done on a solved crystal structure [32]. Protein structure images were produced using RasMol 2.7.2.1 [60].

Statistical significance of spatial clustering of amino acids was assessed by comparing the mean pairwise physical distance between positively selected sites to the mean distance between an equal number of random surface sites. A p -value was obtained by making this comparison 10,000 times. Surface sites were defined as those amino acids that are at least 20% solvent-exposed over their surface area. Solvent exposure was calculated using GETAREA 1.1 [61].

Supporting Information

Table S1. Selective Pressure Screen Comparing 161 Human and Chimpanzee Seminal Protein Genes

This table shows pairwise d_N/d_S estimates from two different methods. Found at DOI: 10.1371/journal.pgen.0010035.st001 (79 KB XLS).

Accession Numbers

The GenBank (<http://www.ncbi.nlm.nih.gov>) accession numbers of the genes discussed in this paper are *M. fascicularis PSA* (AY647976), *Papio hamadryas MSMB* (U49786), *Saguinus oedipus MSMB* (AJ010154, AJ010155, AJ010158), and *M. fuscata PIP* (AB098481).

The sequences generated in this study have been submitted to GenBank under accession numbers DQ150438 through DQ150526.

Acknowledgments

We would like to thank Dr. Evan Eichler's lab for providing quality gorilla and siamang DNA. We also thank Dr. Joshua Akey, Chris Saunders, Dr. Dick Hwang, Dr. Peter Nelson, and the members of the Swanson lab for valuable advice. We appreciate the helpful comments made by two anonymous reviewers. NLC is supported by the University of Washington National Institutes of Health Genetics Training Grant. WJS is supported by National Institutes of Health grant HD42563 and NSF grant #DEB 0410112.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. NLC and WJS conceived and designed the experiments. NLC performed the experiments. NLC analyzed the data. NLC and WJS contributed reagents/materials/analysis tools. NLC wrote the paper. ■

References

1. Galindo BE, Vacquier VD, Swanson WJ (2003) Positive selection in the egg receptor for abalone sperm lysin. Proc Natl Acad Sci U S A 100: 4639–4643.
2. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci U S A 98: 7375–7379.
3. Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. Genetics 168: 1457–1465.
4. Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A 98: 2509–2514.

5. Yang Z, Swanson WJ, Vacquier VD (2000) Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol* 17: 1446–1455.
6. Ferris PJ, Pavlovic C, Fabry S, Goodenough UW (1997) Rapid evolution of sex-related genes in *Chlamydomonas*. *Proc Natl Acad Sci U S A* 94: 8634–8639.
7. Metz EC, Palumbi SR (1996) Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol Biol Evol* 13: 397–406.
8. Tsaur SC, Wu CI (1997) Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. *Mol Biol Evol* 14: 544–549.
9. Jensen-Seaman MI, Li WH (2003) Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *J Mol Evol* 57: 261–270.
10. Kingan SB, Tatar M, Rand DM (2003) Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. *J Mol Evol* 57: 159–169.
11. Wolfner MF (2002) The gifts that keep on giving: Physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. *Heredity* 88: 85–93.
12. Dixson AL, Anderson MJ (2002) Sexual selection, seminal coagulation and copulatory plug formation in primates. *Folia Primatol (Basel)* 73: 63–69.
13. 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.
14. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3: 418–426.
15. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335: 167–170.
16. 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.
17. Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555–556.
18. Suzuki Y, Nei M (2004) False-positive selection identified by ML-based methods: Examples from the *Sig1* gene of the diatom *Thalassiosira weissflogii* and the *tax* gene of a human T-cell lymphotropic virus. *Mol Biol Evol* 21: 914–921.
19. Wong WS, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168: 1041–1051.
20. Fung KY, Glode LM, Green S, Duncan MW (2004) A comprehensive characterization of the peptide and protein constituents of human seminal fluid. *Prostate* 61: 171–181.
21. Utleg AG, Yi EC, Xie T, Shannon P, White JT, et al. (2003) Proteomic analysis of human prostasomes. *Prostate* 56: 150–161.
22. Hawkins V, Doll D, Bumgarner R, Smith T, Abajian C, et al. (1999) PEDB: The Prostate Expression Database. *Nucleic Acids Res* 27: 204–208.
23. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5: 150–163.
24. Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18: 1585–1592.
25. Nolet S, St-Louis D, Mbikay M, Chretien M (1991) Rapid evolution of prostatic protein PSP94 suggested by sequence divergence between rhesus monkey and human cDNAs. *Genomics* 9: 775–777.
26. Lazare C, Villemure M, Gauthier D, Naudé RJ, Mbikay M (2001) Characterization of ostrich (*Struthio camelus*) beta-microseminoprotein (MSP): Identification of homologous sequences in EST databases and analysis of their evolution during speciation. *Protein Sci* 10: 2207–2218.
27. Makinen M, Valtonen-Andre C, Lundwall A (1999) New World, but not Old World, monkeys carry several genes encoding beta-microseminoprotein. *Eur J Biochem* 264: 407–414.
28. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22: 1107–1118.
29. Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* 19: 950–958.
30. LaxmiKanthan G, Blaber SI, Bennett MJ, Scarisbrick IA, Juliano MA, et al. (2005) 1.70 Å X-ray structure of human apo kallikrein 1: Structural changes upon peptide inhibitor/substrate binding. *Proteins* 58: 802–814.
31. Afar DE, Vivanco I, Hubert RS, Kuo J, Chen E, et al. (2001) Catalytic cleavage of the androgen-regulated TMPRSS2 protease results in its secretion by prostate and prostate cancer epithelia. *Cancer Res* 61: 1686–1692.
32. Ortland E, LaCount MW, Lebioda L (2003) Crystal structures of human prostatic acid phosphatase in complex with a phosphate ion and alpha-benzylaminobenzylphosphonic acid update the mechanistic picture and offer new insights into inhibitor design. *Biochemistry* 42: 383–389.
33. Chapman T (2001) Seminal fluid-mediated fitness traits in *Drosophila*. *Heredity* 87: 511–521.
34. Balk SP, Ko YJ, Bubley GJ (2003) Biology of prostate-specific antigen. *J Clin Oncol* 21: 383–391.
35. Brillard-Bourdet M, Rehault S, Juliano L, Ferrer M, Moreau T, et al. (2002) Amidolytic activity of prostatic acid phosphatase on human semenogelins and semenogelin-derived synthetic substrates. *Eur J Biochem* 269: 390–395.
36. Kelly RW, Critchley HO (1997) Immunomodulation by human seminal plasma: A benefit for spermatozoon and pathogen? *Hum Reprod* 12: 2200–2207.
37. Mukherjee DC, Agrawal AK, Manjunath R, Mukherjee AB (1983) Suppression of epididymal sperm antigenicity in the rabbit by uteroglobin and transglutaminase in vitro. *Science* 219: 989–991.
38. Paonessa G, Metafora S, Tajana G, Abrescia P, De Santis A, et al. (1984) Transglutaminase-mediated modifications of the rat sperm surface in vitro. *Science* 226: 852–855.
39. Kamada M, Mori H, Maeda N, Yamamoto S, Kunimi K, et al. (1998) Beta-microseminoprotein/prostatic secretory protein is a member of immunoglobulin binding factor family. *Biochim Biophys Acta* 1388: 101–110.
40. Hirano M, Kamada M, Maeda N, Yamamoto S, Aono T, et al. (1996) Presence of immunoglobulin binding factor on human sperm surface as sperm coating antigen. *Arch Androl* 37: 163–170.
41. Weiber H, Andersson C, Murne A, Rannevik G, Lindstrom C, et al. (1990) Beta microseminoprotein is not a prostate-specific protein. Its identification in mucous glands and secretions. *Am J Pathol* 137: 593–603.
42. Schenkels LC, Walgreen-Weterings E, Oomen LC, Boltscher JG, Veerman EC, et al. (1997) In vivo binding of the salivary glycoprotein EP-GP (identical to GCDFP-15) to oral and non-oral bacteria detection and identification of EP-GP binding species. *Biol Chem* 378: 83–88.
43. Gaubin M, Autiero M, Basmaciogullari S, Metivier D, Mis hal Z, et al. (1999) Potent inhibition of CD4/TCR-mediated T cell apoptosis by a CD4-binding glycoprotein secreted from breast tumor and seminal vesicle cells. *J Immunol* 162: 2631–2638.
44. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, et al. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302: 1960–1963.
45. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, et al. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3: e170.
46. Huttley GA, Easteele S, Southey MC, Tesoriero A, Giles GG, et al. (2000) Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. *Australian Breast Cancer Family Study*. *Nat Genet* 25: 410–413.
47. Zhang J, Rosenberg HF (2002) Diversifying selection of the tumor-growth promoter angiogenin in primate evolution. *Mol Biol Evol* 19: 438–445.
48. Pavlicek A, Noskov VN, Kouprina N, Barrett JC, Jurka J, et al. (2004) Evolution of the tumor suppressor BRCA1 locus in primates: Implications for cancer predisposition. *Hum Mol Genet* 13: 2737–2751.
49. Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, Ostrander EA (2003) Understanding missense mutations in the BRCA1 gene: An evolutionary approach. *Proc Natl Acad Sci U S A* 100: 1151–1156.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
51. Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng* 10: 1–6.
52. Gordon D, Abajian C, Green P (1998) Consed: A graphical tool for sequence finishing. *Genome Res* 8: 195–202.
53. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175–185.
54. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31: 3497–3500.
55. Purvis A (1995) A composite estimate of primate phylogeny. *Philos Trans R Soc Lond B Biol Sci* 348: 405–421.
56. Muller S, Hollatz M, Wienberg J (2003) Chromosomal phylogeny and evolution of gibbons (Hylobatidae). *Hum Genet* 113: 493–501.
57. Hayasaka K, Fujii K, Horai S (1996) Molecular phylogeny of macaques: Implications of nucleotide sequences from an 896-base pair region of mitochondrial DNA. *Mol Biol Evol* 13: 1044–1053.
58. Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20: 18–20.
59. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 31: 3381–3385.
60. Milner-White RSaEJ (1995) RasMol: Biomolecular graphics for all. *Trends Biochem Sci* 20: 374.
61. Franzkiewicz R, Braun W (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comput Chem* 19: 319–333.
62. Wang I, Lou YC, Wu KP, Wu SH, Chang WC, et al. (2005) Novel solution structure of porcine beta-microseminoprotein. *J Mol Biol* 346: 1071–1082.