Comparative sequence analyses reveal rapid and divergent evolutionary changes of the *WFDC* locus in the primate lineage

Belen Hurle, Willie Swanson, NISC Comparative Sequencing Program, et al.

Genome Res. published online January 31, 2007
Access the most recent version at doi:10.1101/gr.6004607

Letter

Comparative sequence analyses reveal rapid and divergent evolutionary changes of the *WFDC* locus in the primate lineage

Belen Hurle,[1] Willie Swanson,[2] NISC Comparative Sequencing Program,[1,3] and Eric D. Green[1,3,4]

[1]*Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA;* [2]*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;* [3]*NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland 20852, USA*

The initial comparison of the human and chimpanzee genome sequences revealed 16 genomic regions with an unusually high density of rapidly evolving genes. One such region is the whey acidic protein (WAP) four-disulfide core domain locus (or *WFDC* locus), which contains 14 *WFDC* genes organized in two subloci on human chromosome 20q13. WAP protease inhibitors have roles in innate immunity and/or the regulation of a group of endogenous proteolytic enzymes called kallikreins. In human, the centromeric *WFDC* sublocus also contains the rapidly evolving seminal genes, semenogelin 1 and 2 (*SEMG1* and *SEMG2*). The rate of *SEMG2* evolution in primates has been proposed to correlate with female promiscuity and semen coagulation, perhaps related to post-copulatory sperm competition. We mapped and sequenced the centromeric *WFDC* sublocus in 12 primate species that collectively represent four different mating systems. Our analyses reveal a 130-kb region with a notably complex evolutionary history that has included nested duplications, deletions, and significant interspecies divergence of both coding and noncoding sequences; together, this has led to striking differences of this region among primates and between primates and rodents. Further, this region contains six closely linked genes (*WFDC12*, *PI3*, *SEMG1*, *SEMG2*, *SLPI*, and *MATN4*) that show strong patterns of adaptive selection, although an unambiguous correlation between gene mutation rates and mating systems could not be established.

[Supplemental material is available online at www.genome.org. Genomic sequences reported in this manuscript have been submitted to GenBank under accession numbers DP000036 to DP000048.]

The recently generated draft sequence of the chimpanzee genome (Chimpanzee Sequencing and Analysis Consortium 2005) provides an exciting resource for better understanding primate biology and evolution. Human–chimpanzee comparative sequence analyses raise numerous intriguing questions about the genetic basis for the myriad differences that distinguish *Homo sapiens* from other primates. Among closely related primate species, distinct phenotypic features can reflect differences in gene content and/or gene expression. In some cases, positive Darwinian selection (i.e., rapid or adaptive evolution) of specific genes or genomic regions is thought to play an important role in these differences. Positive selection for amino acid diversification results in the rate of nonsynonymous substitutions (d_N) exceeding that of synonymous substitutions (d_S). In the absence of selection (i.e., neutral evolution), the ratio d_N/d_S is expected to be one, with values less than one and significantly greater than one indicating purifying selection and positive selection, respectively.

By the above criteria, a number of genes have been shown to be under positive selection in one or more primate lineages (for reviews, see Vallender and Lahn 2004; Sabeti et al. 2006). For example, the recent comparative analysis of the human and chimpanzee genome sequences, which included calculating the median d_N/d_S ratio for sliding windows (of 10 orthologous genes each) across the aligned genomes, revealed 16 regions with a notably high density of rapidly evolving genes (Chimpanzee Sequencing and Analysis Consortium 2005). One of these regions contains genes that encode whey acidic protein (WAP) domain protease inhibitors. This 683-kb region, which resides on human chromosome 20q13, is also called the WAP four-disulfide core domain locus (or *WFDC* locus) (Clauss et al. 2002).

The human *WFDC* locus contains the genes encoding 14 *WFDC*-type protease inhibitors and is organized into two subloci separated by a 215-kb segment (Clauss et al. 2002). The typical *WFDC* gene contains a promoter region devoid of a TATA-box, a 5′ exon encoding a signal peptide, one or more exons that each encodes a *WAP* domain, and a 3′ exon with limited or no coding sequence that contains the polyadenylation signal. Several of the *WFDC* genes are expressed ubiquitously, but in most cases, expression is predominantly in the epididymis, testis, and trachea. The *WFDC*-encoded proteins are thought to play a role in innate immunity and/or in regulating endogenous kallikreins (*KLK*) (Borgono et al. 2004).

In human, the centromeric *WFDC* sublocus also contains the genes encoding the seminal proteins semenogelin 1 and 2 (*SEMG1* and *SEMG2*, respectively) (Peter et al. 1998) and elafin (or protease inhibitor 3 [*PI3*]). *PI3* is a chimeric-like gene in the trappin family, and was presumably derived from the shuffling of exons between ancestral *SEMG*-like and *WFDC* genes (Schalkwijk et al. 1999). The semenogelin and trappin genes are collectively referred to as the Rapidly Evolving Substrates for Transaminases

[4]Corresponding author.
E-mail egreen@nhgri.nih.gov; fax (301) 402-2040.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6004607.

(*REST*) family. *REST* family members have a characteristic three-exon structure, with the second exon containing the entire open reading frame; the first and last exons of all *REST* genes are highly conserved among mammals, while the second exon is typically quite diverged, such that the encoded proteins are not similar in their primary structure (Lundwall and Lazure 1995; Lundwall and Ulvsback 1996). The rapid evolution of the *SEMG* genes appears to have involved internal expansion of the second exon, leading to a highly repetitive primary structure of the encoded SEMG proteins (Ulvsback and Lundwall 1997; Jensen-Seaman and Li 2003) or, in some cases, alternative splicing of transcripts (Hagstrom et al. 1996).

The *SEMG* genes are highly expressed in the seminal vesicles, with the encoded SEMG proteins accounting for nearly half of all the protein in human ejaculate. After ejaculation, the SEMG proteins undergo cross-linking to become the principal structural component of semen coagulum, entrapping the ejaculated spermatozoa in the reproductive tract of recipient females (Robert and Gagnon 1999). Later, the prostate-specific antigen (PSA) protease breaks down the cross-linked matrix into smaller peptides, allowing the spermatozoa to regain their mobility. In primates with monoandrous mating systems (monogamy and polygyny—with each female mating with one male during a given periovulatory period), the ejaculate is viscous in texture but does not readily solidify. In primates with polyandrous mating systems (dispersed and multimale–multifemale—with each female mating with multiple males during a given periovulatory period), the ejaculate forms a conspicuous or even rigid copulatory plug (Dixson and Anderson 2002, 2004). Interestingly, female promiscuity and semen coagulation in primates are thought to correlate with the rate of *SEMG2* evolution, perhaps related to post-copulatory sperm competition (Dorus et al. 2004). In contrast, *PI3* has no known role in primate semen physiology; the encoded protein has anti-protease and antibiotic activities and is produced at mucosa and epithelial sites (e.g., cervix) and wounds, under inflammatory conditions (e.g., psoriasis and lung disease), and in some skin cancers (Lundwall and Ulvsback 1996; Williams et al. 2006).

The *WFDC* locus thus represents an interesting genomic region associated with important physiological functions (Table 1) and rapid evolutionary change. To understand better the evolutionary history of this region and its functional consequences, we sought to sequence and study the centromeric *WFDC* sublocus in a large set of primates. The comparative sequence analyses reported here provide important insights about the adaptive evolutionary changes that have uniquely sculpted this genomic region in the primate lineage.

Results

Comparative sequence data set

The centromeric *WFDC* sublocus spans 145 kb in the human genome, containing the *WFDC5* and secretory leukocyte peptidase inhibitor *(SLPI*) genes at the centromeric and telomeric ends, respectively (Fig. 1A). We isolated (Thomas et al. 2002) and sequenced (Thomas et al. 2003) sets of bacterial artificial chromosome (BAC) clones spanning this genomic interval in 12 primate species, including three great apes, four Old World monkeys, three New World monkeys, and two prosimians (for a listing of the specific species, see Table 2, Fig. 1B). Complete BAC-based coverage of the centromeric *WFDC* sublocus was obtained for all species, and in most cases, the clone and sequence coverage extended for >100 kb on each side of the centromeric *WFDC* sublocus (Table 2).

Gene content and genomic architecture of the centromeric *WFDC* sublocus

Detailed comparative analyses of the generated sequence revealed as many as 12 functional genes in the examined genomic region (Fig. 1B; Table 1), including genes encoding small serine protease inhibitors (*WFDC5*, *WFDC12*, and *SLPI*), trappin (*PI3*), and seminal vesicle-secreted proteins (*SEMG1* and *SEMG2*). Flanking the *WFDC/SEMG* gene cluster are a number of genes that are not functionally related to either family (serine/threonine kinase 4 [*STK4*] and potassium voltage-gated channel, member 1 [*KCNS1*] on the centromeric side; matrilin 4 [*MATN4*], recombining binding protein L [*RBPSUHL*], syndecan 4 [*SDC4*], and dysbindin [dystrobrevin-binding protein 1] domain containing 2 [*DBNDD2*] on the telomeric side [note that *SDC4* and *DBNDD2* are not shown in Fig. 1B]). We also identified pseudogenes in some species (Ψ*WFDC15b* in all species except owl monkey; Ψ*SEMG3* in lemur and galago; Ψ*PI3* and Ψ*WFDC15c* in all great apes and Old World monkeys; and Ψ*WFDC15d*, previously named *LOC149709* [Hagiwara et al. 2003], in all species [Fig. 1B]).

Mammalian genomes are mosaics of discrete regions with different G + C contents (Eyre-Walker and Hurst 2001). Analysis of the generated sequences encompassing the centromeric

Table 1. Genes in the human centromeric *WFDC* sublocus

Gene	Encoded protein	Tissue expression	Function	References
WFDC5	WAP four-disulfide core domain 5	Epididymis, skin	Proteinase inhibitor (proposed)	Clauss et al. 2002
WFDC12	WAP four-disulfide core domain 12	Prostate, skin, lung, esophagus	Antibacterial	Hagiwara et al. 2003
PI3	Elafin (protease inhibitor 3)	Skin, endometrium, large intestine, bronchial secretions, seminal plasma	Innate immunity, antibacterial, antiviral, anti-inflammatory, tissue repair	Williams et al. 2006
SEMG1	Semenogelin 1	Seminal vesicles, vas deferens, prostate, epididymis, trachea	Semen clot formation, antibacterial, sperm motility inhibitor	Robert and Gagnon 1999
SEMG2	Semenogelin 2	Seminal vesicles, vas deferens, prostate, epididymis, trachea	Semen clot formation	Dorus et al. 2004
SLPI	Secretory leukocyte peptidase inhibitor	Seminal plasma, cervical mucus, endometrium, bronchial secretions	Innate immunity, antibacterial, antifungal, anti-inflammatory, tissue repair	Williams et al. 2006

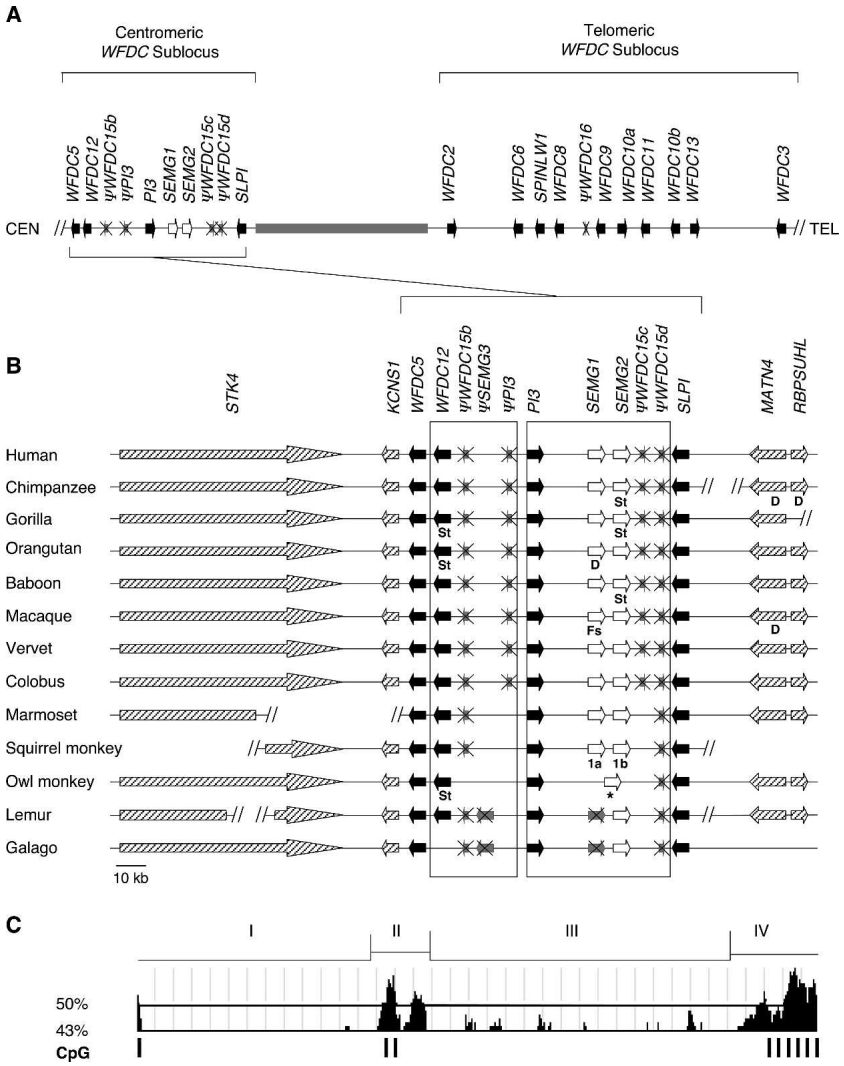

Figure 1. The *WFDC* locus in human and other primates. (*A*) The long-range organization of the *WFDC* locus on human chromosome 20q13 is depicted (oriented relative to the centromere [CEN] and telomere [TEL]) and consists of the two indicated subloci separated by a 215-kb region containing unrelated genes (gray rectangle). The 145-kb centromeric *WFDC* sublocus contains the indicated six genes and four pseudogenes. The 322-kb telomeric *WFDC* sublocus contains the indicated 10 genes and one pseudogene. (*B*) The long-range organization of the centromeric *WFDC* sublocus in human and 12 nonhuman primates is depicted, as deduced by comparative analyses of the orthologous genomic sequences (see Table 2). Also shown are the two genes immediately flanking this sublocus on each side. Solid arrows represent small serine protease inhibitor genes and trappin (*WFDC* and *PI3*, respectively); open arrows represent semenogelin genes (*SEMG*); hatched arrows represent nonrelated neighboring genes; gray arrows with an X represent *WFDC*, trappin, and *SEMG* pseudogenes. The direction of all arrows reflects the gene's transcriptional orientation. //, Gap between sequenced BACs (clone gaps in Table 2). "St" *under* a gene indicates the presence of a premature stop codon; "Fs" *under* a gene indicates the presence of a frameshift; "D" *under* a gene indicates that the corresponding sequence was retrieved from public databases rather than generated as part of this study (specifically, for chimpanzee, macaque, and orangutan). Note that squirrel monkey has two *SEMG1* genes, labeled as "1a" and "1b" *below* each, and that owl monkey has a single unique *SEMG* gene (a *SEMG1-SEMG2* chimera, which is labeled *below* the gene with an *). The two boxed areas reflect the *WFDC12*-to-Ψ*PI3* and *PI3-to-*Ψ*WFCD15d* paralogous duplicons, respectively. (*C*) The G + C content of the human centromeric *WFDC* sublocus is shown. The *Y*-axis is scaled to reflect a minimum of 43% and maximum of 70% G + C content. Note the alternating genomic segments of low (I and III) and high (II and IV) G + C content. Also shown are the positions of CpG islands (defined as >200-bp stretches of sequence with a G + C content of >50% and an observed CpG to expected CpG ratio of >0.6). Note that panel *C* is drawn to scale relative to panel *B*.

WFDC sublocus in different primates revealed a conserved pattern of four regions with alternating low versus high G + C content, as depicted for the human sequence in Figure 1C (see also Supplemental Materials). A similar pattern is seen with the orthologous mouse and rat genomic regions (Supplemental Fig. S1C); while the genomic sequences are less refined at present, it appears that a similar pattern also exists with the orthologous dog and cow genomic regions (data not shown).

Evolutionary history of the centromeric *WFDC* sublocus

Comparative analyses of the generated sequences (both self–self and interspecies pairwise comparisons) indicate that the centromeric *WFDC* sublocus is the product of an ancient duplication that yielded two adjacent segments in a head-to-head configuration (reflected by the two boxed regions in Fig. 1B). Examination of both primate (Fig. 1B) and rodent (Supplemental Fig. S1B) sequences reveals evidence for this duplication event, suggesting that it preceded the divergence of these two lineages roughly 75 million yr ago. The generated sequence data do not, however, allow the boundaries of these duplicated segments to be precisely defined in all species (see Supplemental Materials).

Within a given species, these duplicons have retained detectable coding and noncoding sequence homology (Fig. 2A). However, interspecies sequence comparisons reveal a history of rapid divergence of these genomic segments, with the lengths of corresponding duplicons and the extent of sequence homology varying greatly from one species to another (Supplemental Fig. S2). Likewise, pairwise sequence comparisons of the entire centromeric *WFDC* sublocus show rapid evolutionary divergence among primates, as well as clear evidence of gene deletions and conversions into pseudogenes (see below). Interestingly, different portions of the sublocus have diverged in an asymmetric fashion. The longer and more conserved portion is the *PI3*-to-*SLPI* interval, which spans ~100 kb in human. The *WFDC12*-to-Ψ*PI3* interval (45 kb in human) is drastically smaller and less conserved than its paralogous *PI3*-to-*SLPI* counterpart in all primates. The regions flanking the centromeric *WFDC* sublocus show more-typical patterns of sequence conservation (Supplemental Fig. S3).

Adjacent to the centromeric *WFDC* sublocus, there is no evidence for gene deletion or the presence of pseudogenes among the primates examined. In contrast, within the sublocus,

Table 2. General features of comparative sequence data set

Species	Name	Lineage	No. of BACs	Clone gaps[a]	Sequence gaps[b]	Total sequence[c]	GenBank accession no.
Chimpanzee	*Pan troglodytes*	Great ape	2	0	4	302,717	DP000037
Gorilla	*Gorilla gorilla*	Great ape	2	0	6	355,851	DP000041
Orangutan	*Pongo pygmaeus*	Great ape	3	0	20	644,818	DP000045
Baboon	*Papio anubis*	Old World monkey	4	0	9	492,848	DP000036
Rhesus macaque	*Macaca mulatta*	Old World monkey	3	0	7	373,358	DP000043
Vervet	*Cercopithecus aethiops*	Old World monkey	4	0	23	583,711	DP000048
Black and white colobus	*Colobus guereza*	Old World monkey	3	0	12	481,828	DP000038
Marmoset	*Callithrix jacchus*	New World monkey	3	1	18	574,594	DP000044
Squirrel monkey	*Saimiri boliviensis*	New World monkey	1	0	7	211,360	DP000047
Owl monkey	*Aotus nancymaae*	New World monkey	2	0	7	387,541	DP000046
Ring-tailed lemur	*Lemur catta*	Prosimian	2	1	7	452,647	DP000042
Galago	*Otolemur garnettii*	Prosimian	2	0	7	277,620	DP000040

[a]Number of sequence gaps due to the lack of BAC coverage across an interval.
[b]Number of gaps within assembled sequences of individual BACs.
[c]Total nonredundant nucleotides in the assembled sequences of all BACs from that species (As, Gs, Cs, and Ts; not Ns).

there is considerable evidence for genomic rearrangements and alterations, including the loss of gene function due to deletions, frameshift mutations, and the introduction of premature stop codons. In some cases, a given gene appears to have been inactivated in all primates, whereas in other cases, this inactivation appears to be species-specific. The specific findings for the *WFDC,* trappin, and *SEMG* gene families are discussed separately below.

WFDC gene family

Several *WFDC* genes appear to be nonfunctional in all primates examined and may represent the products of older pseudogenization events. For example, there is no intact *WFDC15* gene in human, but there are two *Wfdc15* genes (*Wfdc15a* and *Wfdc15b*) residing side-by-side at one end of the *Wfdc* locus in rodents. The orthologous pseudogene in human has been proposed to reside between *WFDC12* and *PI3* (Clauss et al. 2005) or between *SEMG2* and *SLPI* (Hagiwara et al. 2003). Depending on the primate species, our analyses revealed up to three *WFDC15*-like sequences (hereafter designated Ψ*WFDC15b*, Ψ*WFDC15c*, and Ψ*WFDC15d*), organized in two clusters (Fig. 2A). Each *WFDC15*-like sequence consists of a ~1.2-kb segment that includes two corrupted exons and corresponding flanking regions (Fig. 2B, panels 4). The number, organization, and relative orientation of the rodent *Wfdc15* genes and primate Ψ*WFDC15* sequences suggest that the duplicated segment in the ancestral mammalian genome contained two copies of *Wfdc15* at one end. In contrast to the pseudogenization seen with *WFDC15* in all primates, there is species-specific loss of function seen with *WFDC12* (via deletion in galago and premature stop codons in gorilla [Tyr91], orangutan [Tyr91], and owl monkey [Glu93]).

Trappin gene family

In contrast to New World monkeys and prosimians, great apes and Old World monkeys appear to contain a trappin pseudogene (Ψ*PI3*) in addition to *PI3*. *PI3* and Ψ*PI3* sit near the boundaries of the paralogous duplicons, as best seen in the great apes and Old World monkeys (Fig. 1B).

Our findings also suggest that *WFDC12* and *PI3* may have derived from the alternative splicing of a common ancestral gene. As seen in Figure 2B (panel 2), the 5′ region and first exon of *WFDC12* and *PI3* are homologous, but sequences homologous to exons 2 and 3 of human *WFDC12* reside within the first intron of *PI3*.

SEMG gene family

In human, the two *SEMG* genes have been reported to reside within adjacent, duplicated 9-kb blocks (Ulvsback et al. 1992); these blocks are labeled *Sgb1* and *Sgb2* (for Semenogelin genomic block) in Figure 2A. Unexpectedly, sequence comparisons of either *Sgb1* or *Sgb2* and the entire centromeric *WFDC* sublocus revealed additional *SEMG*-related sequences in all primates, hereafter referred to as *Sgb3* and *Sgb4* (Fig. 2C). *Sgb4* is a truncated duplicon spanning 2 kb of noncoding sequence that is located 10 kb distal to (and in the same orientation as) *SEMG2*; *Sgb4* is only present in the great apes and New World monkeys (Supplemental Fig. S4). *Sgb3*, also truncated and devoid of *SEMG*-coding sequence, resides within the *WFDC12*-to-Ψ*PI3* duplicon in all primates except colobus and owl monkey (Supplemental Fig. S4). Among the primates studied, the prosimians provide the best insight about the ancestral architecture of the *SEMG* gene cluster (Supplemental Fig. S4). For instance, *Sgb3* spans a portion of a *SEMG* pseudogene in galago and lemur (Ψ*SEMG3*); furthermore, in lemur, the *Sgb3*- and *Sgb1*-containing segments are 5 kb longer (at their 5′ ends) compared with other primates. In higher primates, those extended *Sgb* segments are fragmented or missing. The presence of multiple *Sgb* sequences in two oppositely oriented clusters resembles the spatial organization of *Svs* genes in rodents (Supplemental Fig. S1B), and suggests that a number of *SEMG* genes may have been deleted during primate evolution.

The *SEMG* gene family appears to have been particularly dynamic in New World monkeys, among which only marmoset has the above-described genomic structure with four *Sgb*-containing duplicons (Supplemental Fig. S4). The two semenogelin genes in squirrel monkey (*SEMG1a* and *SEMG1b*) are highly similar at a sequence level and cluster in a single monophyletic group within the *SEMG1* phylogenetic tree; the same result was obtained when exonic or intronic sequences were examined (data not shown). This pattern is consistent with a recent genetic exchange that homogenized the *SEMG* genes in squirrel monkey. In owl monkey, a deletion (likely triggered by an *Alu* insertion in intron 1 of *SEMG1*) appears to have yielded a single chimeric *SEMG* gene, which consists of a *SEMG1*-like exon 1 and intron 1 as well as *SEMG2*-like exon 2, intron 2, and exon 3. Of note, in

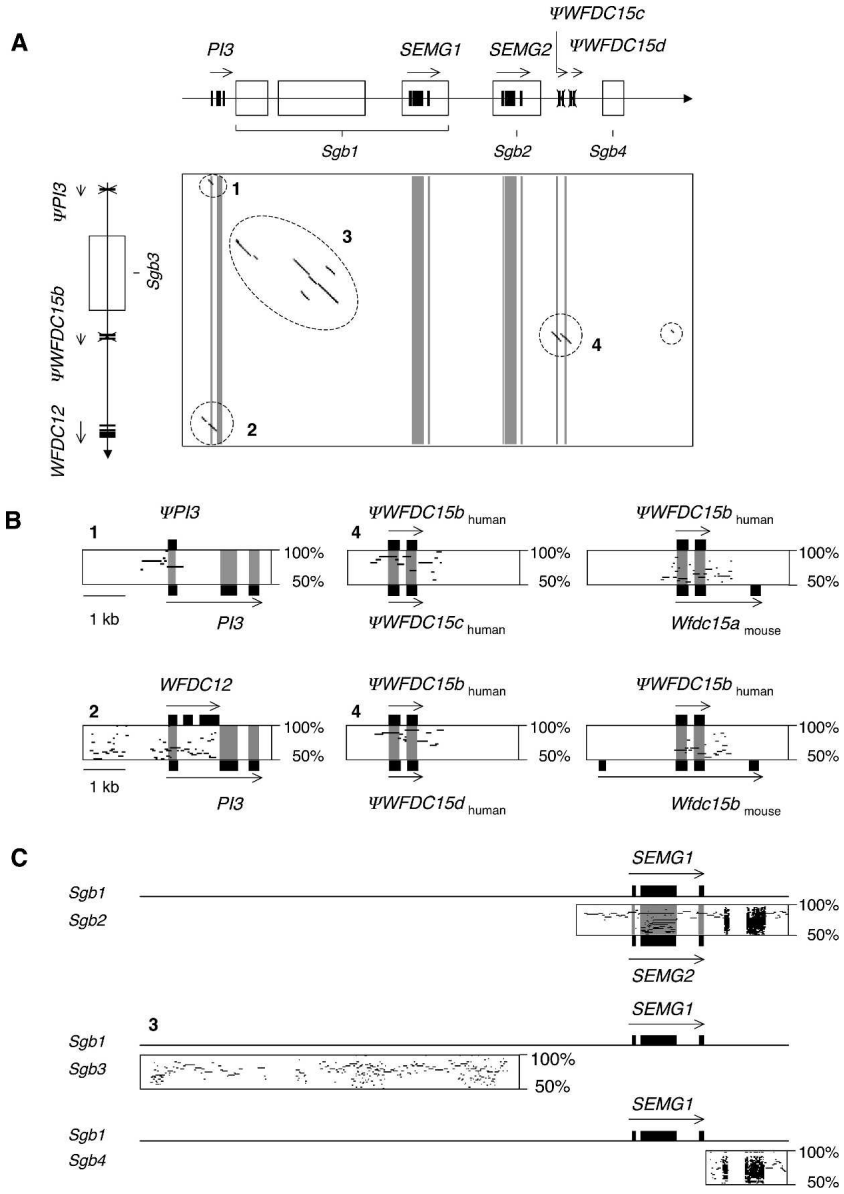

Figure 2. Sequence conservation between paralogous segments in the centromeric *WFDC* sublocus. (*A*) A dot plot depicts sequence conservation between the two paralogous segments that reside within the human centromeric *WFDC* sublocus. The *X*- and *Y*-axes represent the *PI3*-to-Ψ*WFCD15d* and *WFDC12*-to-Ψ*PI3* duplicons, respectively (boxed regions in Fig. 1B). A schematic representation of the exon-based gene structures and pseudogenes is shown for each, with the positions of exons along the *X*-axis highlighted in the dot plot. The four *SEMG* genomic blocks (*Sgb*) are boxed and labeled. *Sgb1* is larger than previously described (Ulvsback et al. 1992) and is fragmented into three pieces, as determined by comparison with the lemur sequence. The two paralogous segments have significant coding and noncoding sequence conservation (highlighted by dashed circles, four of which are labeled 1–4 and detailed in *B* and *C*). (*B*) PipMaker plots of three paralogous regions highlighted in *A*. (*Left* column, panel *1*) Ψ*PI3* resides within a genomic segment that is 79% similar to a 774-bp region of *PI3*, including exon 1; (*left* column, panel *2*) there is considerable conservation between *PI3* and *WFDC12*, including the proximal 1495 bp of the 5′region (58% identical), the first exon (79% identical at the nucleotide level; 10/22 residues identical at the amino acid level), and the first *PI3* intron with *WFDC12* exons 2 and 3 (55% identical); (*center* column, panels *4*) the three human *WFDC15* pseudogenes show >79% sequence identity with each other as well as lower but significant identity with a 1.2-kb segment of mouse *Wfdc15a* and *Wfdc15b* (*right* column). (*C*) PipMaker plots showing the sequence conservation among the four paralogous human *SEMG* genomic blocks (*Sgb1* to *Sgb4*). In all cases, the reference sequence is *Sgb1*. Newly identified *Sgb3* and *Sgb4* are truncated blocks that have lost all *SEMG*-coding sequences. The *Sgb1*/*Sgb3* plot corresponds to the highlighted region 3 in panel *A*. Panels *A* and *C* are not to scale.

the cotton-top tamarin (another New World monkey), *SEMG2* has been replaced by a truncated LINE1 element (Lundwall and Olsson 2001).

The C-terminal half of mature SEMG proteins consists of multiple transglutaminase domains, each 60 amino acids in length. Upon ejaculation, the cross-linking of SEMG proteins by a prostate-derived transglutaminase results in semen coagulation. Consistent with previous reports (Ulvsback and Lundwall 1997; Jensen-Seaman and Li 2003), the length of the *SEMG1* and *SEMG2* coding regions, which dictates the number of transglutaminase domains in the corresponding SEMG proteins, varies substantially among primates. As shown in Figure 3, there is a general (albeit imperfect) trend between the number of repeated transglutaminase domains in the SEMG1 and SEMG2 proteins and the relative amount of both female promiscuity and semen coagulation. Note that a certain amount of SEMG1 and SEMG2 size variation has been described in individual hominoid species, although the smaller allele was always much less frequent than the larger (Jensen-Seaman and Li 2003). In that regard, the sequence of two overlapping macaque BACs (apparently derived from different haplotypes) revealed polymorphic forms of SEMG2 that differed in length by one transglutaminase domain. Finally, galago SEMG2 is strikingly distinct in its structure due to a unique ~3-kb insertion in exon 2 that encodes 77 glutamine-rich repetitive domains, each 13 amino acids in length (Fig. 3; data not shown).

Additionally, we found evidence for disrupted open reading frames in at least one *SEMG* gene in a number of primates (e.g., frameshift in exon 2 of macaque *SEMG1* [Gln371], premature stop codon in exon 2 of *SEMG2* [Gln409 in gorilla, Gly384 in baboon, and Tyr409 in chimpanzee], and a near-complete deletion of *SEMG1* [and Ψ*SEMG3*] in lemur and galago [Fig. 1B]). The presence of premature stop codons in primate *SEMG* genes has been reported, including gorilla (multiple polymorphic, premature stop codons in *SEMG1* and *SEMG2*), gibbon (*SEMG1*), chimpanzee (*SEMG2*), and bonobo (*SEMG2*) (Kingan et al. 2003). In aggregate, only five of the 16 primates studied (i.e., human, orangutan, vervet, colobus, and marmoset) appear to contain functional, full-length copies of both *SEMG1* and *SEMG2*.

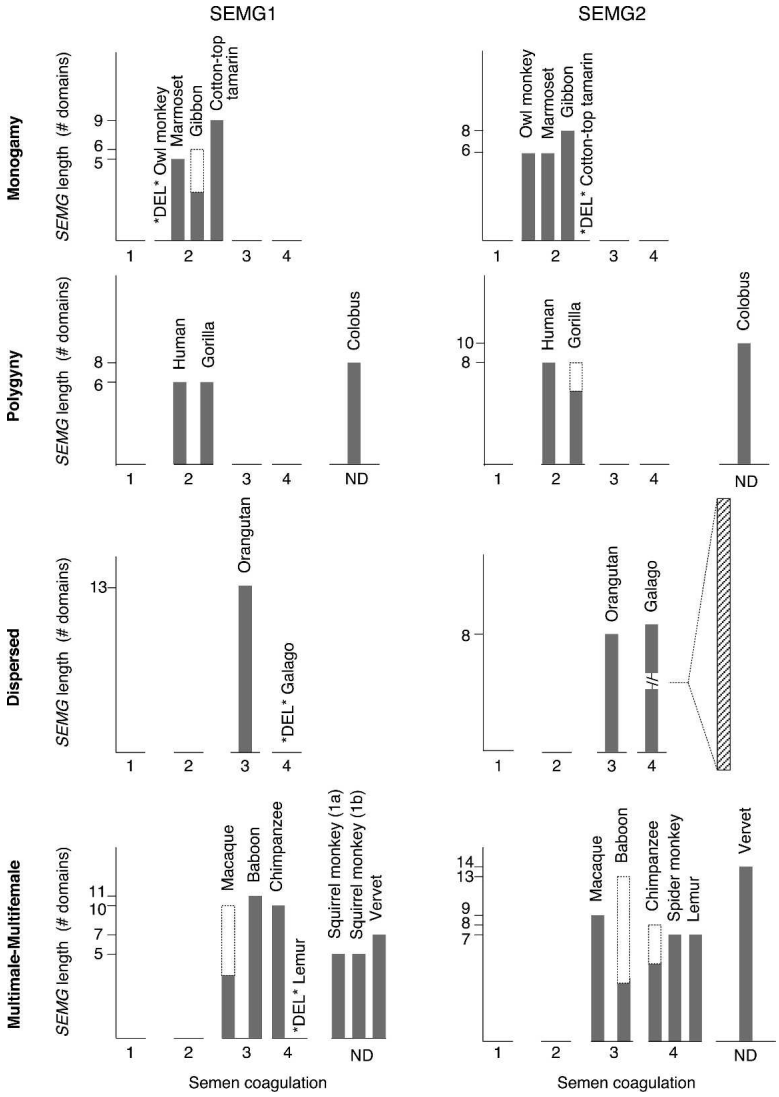

Figure 3. Relationships of semen coagulation, SEMG protein length, and mating system among primates. Primates are grouped according to their established mating system (Dixson 1997): monogamy, polygyny, dispersed, and multimale–multifemale. Semen coagulation is rated on a four-point scale (Dixson and Anderson 2002), with 1 reflecting no coagulation and 4 reflecting the production of a solid copulatory plug (ND, semen coagulation data not available). The indicated gray bars reflect the number of repetitive domains (each 60 amino acids in length) per SEMG protein, indicated for SEMG1 (*left*) and SEMG2 (*right*). Open bars represent predicted truncated proteins due to a premature stop codon or frameshift. *SEMG* genes that are deleted or nearly deleted are indicated by *DEL* before the species name. The hatched bar for galago SEMG2 indicates a ~3-kb galago-specific inserted sequence encoding 77 additional domains (each 13 amino acids in length). The sequences for gibbon SEMG1 and SEMG2, cotton-top tamarin SEMG2, and spider monkey SEMG2 were obtained from the public databases (see Methods). Squirrel monkey has no *SEMG2* gene, but contains two copies of a *SEMG1* gene that may be the product of a gene-conversion event (yielding *SEMG1a* and *SEMG1b*, respectively).

Evidence for positive selection

Each gene in the centromeric *WFDC* sublocus (*WFDC5*, *WFDC12*, *PI3*, *SEMG1*, *SEMG2*, and *SLP1*) shows an elevated rate of substitution, as measured by the tree length and represented by *s* in Table 3 (*s* = substitutions per codon). On average, this value (*s* = 1.7) is more than twice that of the flanking genes (*s* = 0.6). To determine if the divergence of this region has been driven by evolution of the coding regions or a generally higher mutation rate, we calculated the mean pairwise divergence of introns and exons for each gene. All of the introns have similar rates of divergence, ranging from 6.0%–7.5%. Interestingly, the exons in the region are, on average, more divergent, with the average exon divergence being greater than that of the corresponding introns for three (*WFDC12*, *SEMG1*, and *SEMG2*) of the six genes. Overall, exon divergence ranged from 4.7%–10.5%. While such analyses are not a test for adaptive evolution, it is rare for exons to be more divergent than introns. For a few genes in which a similar observation has been made, there is additional evidence that adaptive evolution is being driven by positive selection (Metz et al. 1998; Johnson et al. 2001).

In light of the dynamic nature of the genes in this genomic region, we tested whether their evolutionary divergence has been promoted by positive selection. Specifically, we calculated the d_N/d_S ratio (ω) averaged across all sites and species for each gene. ω is a measure of the selective pressure acting upon a gene. The neutral theory predicts that ω should be equal to one in cases where no selection is operating (e.g., in the case of a pseudogene). In rare cases, ω significantly exceeds one, and this can be accounted for by positive selection acting to drive divergence of the amino acid sequence. The results of these analyses are summarized in Table 3. While ω is greater than or equal to one in the case of four genes (*PI3*, *SEMG1*, *SEMG2*, and *SLP1*), it is less than one for the majority of genes in the region. This is not unexpected, as averaging ω across all sites is not a powerful test of adaptive evolution (Yang et al. 2000). We thus proceeded to test for evidence of site-specific adaptive evolution using likelihood ratio tests (Table 3), and found that six (of 12) genes in this region have been subjected to adaptive evolution (*WFDC12*, *PI3*, *SEMG1*, *SEMG2*, *SLP1*, and *MATN4*). The results for two genes (*WFDC12* and *MATN4*) were significant only when testing whether the extra ω class was significantly greater than one, confirming previous indications that such an approach may reflect the most powerful and robust comparison (Swanson et al. 2003). The results of all comparisons remained significant when correcting for multiple tests using a false-discovery rate of 5% (Storey 2002). All of the genes found to be subject to adaptive evolution reside in the centromeric *WFDC* sublocus, and are contiguous between *WFDC12* and *MATN4*. The proportion of amino acids subject to positive selection is relatively high for these genes; in the case of the *SEMG* genes, such sites are candidates for being in-

Table 3. Evidence for adaptive evolution of genes residing in the *WFDC* locus

Gene	M1 vs. M2	M7 vs. M8	M8 vs. M8a	Lineage	Parameter estimates	Sites under positive selection	*s*	d_N/d_S	N	lc
STK4	0.1	0.1	0.1	17.6	NA	NA	0.4	0.6	12	487
KCNS1	0.7	4.1	3.8	42.0[a]	NA	NA	0.7	0.1	11	531
WFDC5	1.1	2.1	1.6	23.8	NA	NA	1.5	0.4	13	128
WFDC12	3.6	4.3	4.0[a]	15.0	37% (d_N/d_S = 1.9)	None with $P > 0.95$	1.9	0.7	12	116
P13	12.0[a]	5.3	5.8[a]	8.0	36% (d_N/d_S = 2.1)	67K	1.4	1.0	13	126
SEMG1	34.4[b]	34.28[a]	34.3[b]	48.8[a]	26% (d_N/d_S = 2.7)	111R, 138R, 182R, 252C, 335A, 336H, 395A, 455A, 809A, 831W, 857S, 869I, 880H, 882S	1.7	1.1	13	895
SEMG2	45.2[b]	54.9[b]	45.2[b]	28.6	19% (d_N/d_S = 4.0)	85W, 252H, 294R	1.3	1.4	12	964
SLP1	24.4[b]	24.7[b]	24.2[b]	28.4	23% (d_N/d_S = 3.9)	32G, 51Y, 57Q, 72T, 85N, 93K, 109F, 112M, 124M, 135V	2.0	1.1	13	137
MATN4	0.1	5.0	4.6[a]	19.4	NA	278G	0.6	0.1	11	583
RBPSUHL	0.1	0.8	0.1	37.0	NA	NA	0.6	0.1	7	519
SDC4	0.1	0.1	0.2	42.0[a]	NA	NA	0.4	0.3	9	198
DBNDD2	0.7	0.8	0.6	10.6	NA	NA	0.3	0.6	7	260

M1 vs. M2, likelihood ratio test statistic for model M1 versus M2 (see Methods); M7 vs. M8, likelihood ratio test statistic for model M7 versus M8; M8 vs. M8a, likelihood ratio test statistic for model M8 versus M8a; lineage, likelihood ratio test statistic for variation in the d_N/d_S ratio among lineages; parameter estimates, the proportion of amino acids predicted to be under adaptive evolution (and their corresponding d_N/d_S ratio), with all data from model M8; sites under positive selection, individual amino acid positions found to be under positive selection with $P > 0.95$ (for full listing, see Supplemental Table S2); *s*, the tree length; d_N/d_S, the d_N/d_S ratio averaged across all sites and lineages; and N, number of primate species with sequences in alignment; lc, length of the alignment given as the number of codons (or amino acids); NA, not applicable since positive selection is not a significantly better fit to the data than the neutral model.
[a]Significance with $P < 0.05$.
[b]Significance with $P < 0.001$.

volved in interactions with other male- or female-specific proteins.

We also examined the genes in this region by analyzing variation in the d_N/d_S ratio between lineages to establish whether the selective pressure varied among primate lineages and their corresponding mating systems (Table 3). Indeed, other groups have reported a correlation between the divergence rate of *SEMG2* and the number of mates per ovulatory cycle (Dorus et al. 2004). We only found evidence for variation in ω among lineages for three genes (*SEMG1*, *KCNS1*, and *SDC4*), only one of which resides within the centromeric *WFDC* sublocus. Using simple regression analyses, as was done previously for *SEMG2* (Dorus et al. 2004), we were unable to detect a correlation between ω and mating system.

Discussion

Previously, the ancestral organization of the genomic region harboring the *WFDC/SEMG* gene cluster could not be readily inferred because of the limited sequence similarity of the orthologous genes (e.g., primate *SEMG* and rodent *Svs* genes), the differences among major lineages with respect to the number of genes in the region, and the fragmentary nature of available genomic sequence for species other than human, chimpanzee, mouse, and rat (see http://www.genome.ucsc.edu). The studies we report here, which included generating >5 Mb of high-quality sequence from 12 primates and detailed multi-species sequence comparisons, have provided enhanced understanding of the structure and evolutionary history of this biologically rich genomic region.

The centromeric *WFDC* sublocus has been subjected to numerous dynamic events in a relatively short evolutionary time period, with a general summary of these events cataloged in Figure 4. The general pattern of change seen with this genomic region is consistent with "birth-and-death" evolution—a model that involves the generation of new genes by duplication events, with duplicate genes then eliminated or rendered nonfunctional during subsequent speciation (Nei and Rooney 2005). Eventually, different lineages may not truly share orthologous genes at a particular locus; rather, the remaining orthology can reflect a gene in one species that corresponds to a pseudogene or a deleted gene in another species.

Based on our sequence comparisons, it is most parsimonious to conclude that the ancestral mammalian locus experienced a number of duplications that led to a cluster of ancestral *SEMG/Svs* genes and exon shuffling between an ancestral *SEMG/Svs* gene and an ancestral *WFDC* gene that led to the original trappin gene (Schalkwijk et al. 1999). The precise timing and order of these events cannot be pinpointed, although the birth of an ancestral trappin gene is thought to have occurred prior to the split of primates, rodents, artiodactyls, and carnivores (Schalkwijk et al. 1999; Furutani et al. 2005). Later, a larger duplication event affecting the *SEMG/Svs* gene cluster, ancestral trappin gene, and a number of neighboring *WFDC* genes yielded the ancestral centromeric *WFDC* sublocus; this event preceded the split of rodents and primates that occurred ~75 million yr ago (see Fig. 4). Of note, *PI3* and *ΨPI3* sequences have been identified in the orthologous region of the dog genome (Clauss et al. 2005), suggesting the presence of a similar duplicated segment in dog. After this major duplication event, the seminal genes rapidly diverged and lost homology at the primary sequence level, eventually yielding the *SEMG* family in primates and *Svs* family in rodents. Various additional gene-duplication and gene-loss events occurred in different lineages, producing partially overlapping inventories of genes, even in the case of closely related species. The above changes have been particularly dramatic in the primate genomes. Among the primates examined, all four *WFDC15* genes and one trappin gene have been converted to pseudogenes or deleted, and there has been a progressive reduction in the size of the *SEMG* gene family. The evolutionary history of *SEMG* genes in primates has also been punctuated by numerous species-specific events, including gene deletions, gene homogenizations, long-range ge-

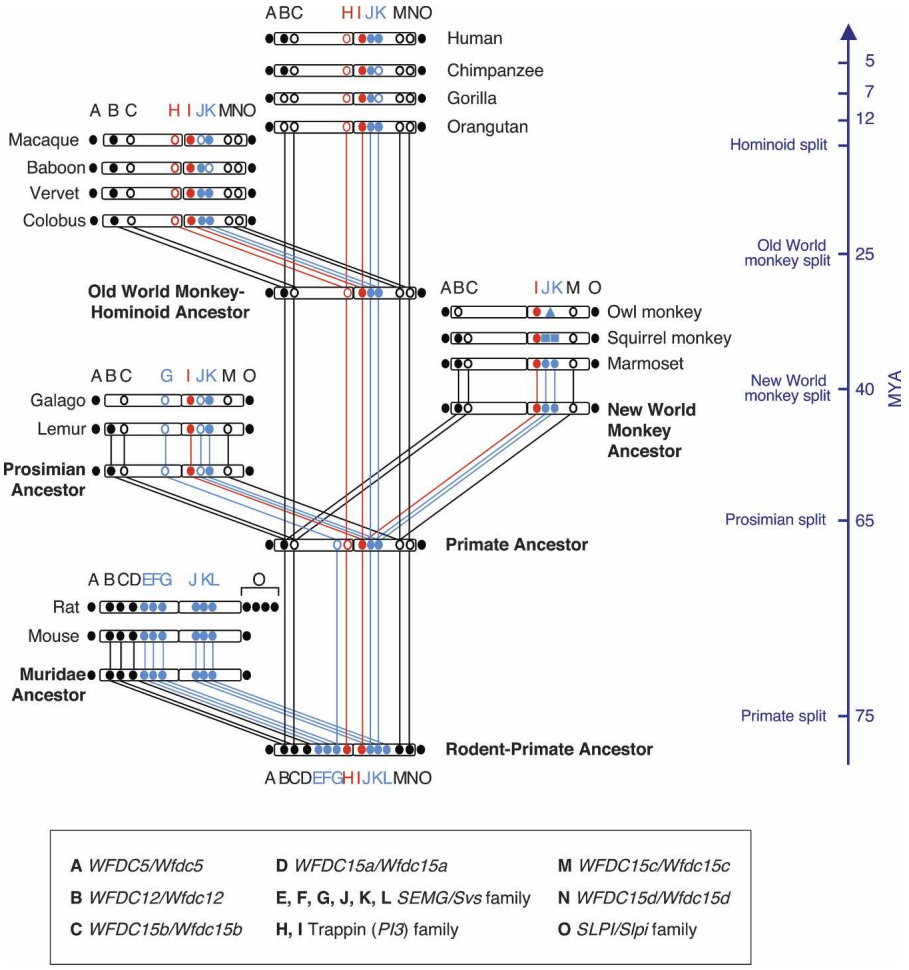

Figure 4. Evolutionary history of the centromeric *WFDC* sublocus. The major duplication, deletion, rearrangement, and pseudogenization events involved in the evolution of the centromeric *WFDC* sublocus since the split of rodents and primates are schematically cataloged. In each case, the two paralogous duplicons (configured in a head-to-head fashion) within the sublocus are represented by two horizontal rectangles. Solid and open circles indicate functional genes and nonfunctional pseudogenes, respectively; black, red, and blue circles represent *WFDC*, trappin, and *SEMG/Svs* genes, respectively. Genes are labeled A through O, with the specific names provided in the key at the *bottom*. The blue squares depict the two *SEMG1* genes in squirrel monkey, which are the result of a gene-conversion event. A blue triangle depicts the single *SEMG* gene in owl monkey (a *SEMG1-SEMG2* chimera). The evolutionary history of each gene can be traced along the indicated lineages by the colored lines. Divergence times are provided on the *right* (MYA, million years ago), as indicated by Goodman et al. (1998); note that this evolutionary timeline is not drawn to scale.

nomic rearrangements, and open reading frame expansions (see Results; Fig. 4; Supplemental Fig. S4). Furthermore, in half of the extant primates examined (eight of 16), the *SEMG1* and/or *SEMG2* genes have truncated open reading frames, suggesting an ongoing evolutionary trend toward pseudogenization.

In contrast to primates, mouse and rat contain a large cluster of six *Svs* genes. Although structurally belonging to the general *Svs* gene family, three *Svs* genes (*Svs4*, *Svs5*, and *Svs6*) do not contain transglutaminase substrate domains and, therefore, are not likely to encode semen-clotting activity (Lin et al. 2005). The mouse and rat genomes also contain two functional *Wfdc15* genes, but have apparently lost all trappin genes. In the 12 million yr since the split of the mouse and rat lineages, the *Slpi* gene family in rat has expanded to include four members. Of note, the orthologous region in the guinea pig genome seems to differ from that in both mouse and rat, containing at least two trappin genes (including a unique gene encoding caltrin II, which is not present in any other mammalian genome studied to date) (Furutani et al. 2005), and only one seminal vesicle protein gene (*GPIG*) (Hagstrom et al. 1996). At present, the structure of the centromeric *WFDC* sublocus in other mammals is largely uncharacterized, but the available data suggest that birth-and-death evolutionary processes are occurring in other lineages as well. For example, no *SEMG* gene has been identified in the dog genome to date (Clauss et al. 2005), while the pig genome contains at least six trappin genes (Furutani et al. 1998).

Analyses of single-nucleotide polymorphism (SNP) frequencies across the human genome have revealed the presence of several extended regions with a striking deficiency of variation, which is suggestive of a selective sweep (Schwartz et al. 2003b; Hinds et al. 2005) Some of these regions are particularly large, encompassing as much as a megabase of DNA and containing up to 16 genes (Carlson et al. 2005). While these regions are likely to contain at least one gene that has been subjected to adaptive evolution, it is nearly impossible to establish the exact target of selection due to "hitchhiking" effects. The *WFDC* locus studied here does not stand out as particularly unusual in this regard, based on HapMap (International HapMap Consortium 2003) or Perlegen (Hinds et al. 2005) polymorphism data and either the Tajima D test statistic (Carlson et al. 2005) or long-range linkage disequilibrium decay analyses (Wang et al. 2006). In contrast to such intra-species polymorphism surveys, the interspecies divergence data we report here demonstrate that several genes in the centromeric *WFDC* sublocus have been subjected to adaptive evolution. Hitchhiking effects will not lead to significant results when using estimates of d_N/d_S (ω), but they do make it difficult for successive selective sweeps of tightly linked loci (Barton 1995; Kim and Stephan 2003) as found in the present study. The basic idea is that if there are two linked selected genes on separate haplotypes, they will cause interference with one another during fixation, resulting in a reduced fixation probability. It is unusual to find six tightly linked genes that all show robust patterns of adaptive evolution, making this a promising genomic region to further investigate the population genetics of interference selection (Barton 1995; Kim and Stephan 2003). There are other cases where genes involved in reproduction are tightly linked and subject to strong selection; for example, the self-incompatibility genes encoding components involved in pollen–pistil interaction are physically linked in order

to maintain self-incompatibility (Schopfer et al. 1999). We currently do not know if there is a selective advantage to having the genes in the centromeric *WFDC* sublocus tightly linked, or if they have become linked by chance due to the dynamics of the region.

When testing for variation in the d_N/d_S ratio between lineages, significant results were only obtained for *SEMG1*, *KCNS1*, and *SDC4*. Nonetheless, these results indicate that selective pressure does vary between lineages. As previously observed for *SEMG2* (Dorus et al. 2004), there appears to also be a trend of increased rates of *SEMG1* evolution with increased levels of mating. For example, the trend of an increasing d_N/d_S ratio for *SEMG1* among hominoids (gorilla $d_N/d_S = 0.6$, human $d_N/d_S = 1.4$, and chimpanzee $d_N/d_S = 1.8$) parallels the increased number of mates from gorilla to human to chimpanzee, suggesting increases in the selective pressures that potentially relate to enhanced sperm competition. However, the overall correlation coefficient is only 0.1 for this relationship, suggesting that other factors may play a role as well. One potential problem with this correlation is that mating systems change over time; thus, mating-system classifications are contemporaneous, whereas evolutionary rates reflect a historical record. Another potential problem is that binary binning of mating behaviors into promiscuous or monogamous can be arbitrary to a certain point; most species may actually possess somewhat flexible mating behavior, which does not necessarily fit into a single "system." Lastly, the majority of primate-mating systems are classified based on observation rather than genetics, which alone can result in misclassification. For example, birds were often thought to be monogamous until molecular approaches demonstrated that clutches are typically sired by multiple males (Gowaty and Karlin 1984). Additionally, correction for the phylogenetic relatedness of the species using an independent contrasts test (Felsenstein 1985) would reduce the significance of any correlation. One possible reason that we do not observe the same pattern for *SEMG2* as reported previously (Dorus et al. 2004) is that we have examined more taxa, which increases the number of degrees of freedom in the statistical tests and may affect the overall power of the analyses.

All of the well-characterized genes residing in the *WDFC* locus encode proteins that appear to have a role at the interface between immunity and fertility, two processes that are often associated with adaptive evolution (Swanson et al. 2003). SEMG proteins serve various functions in the preparation of spermatozoa for fertilization. Besides their well-established role in semen coagulation and spermatozoa entrapment, the N-terminal peptides produced by cleavage of SEMG1 also have antimicrobial activity that contributes to the survival of spermatozoa in the female reproductive tract (Bourgeon et al. 2004). SLPI and PI3 are serine proteinase inhibitors produced at mucosal sites (i.e., upper respiratory tract, oral cavity, skin, genitals, and gastrointestinal tract) and wounds, where they promote early eradication of invading pathogens and protect the host against proteolytic destruction that can follow neutrophil recruitment. In the male reproductive tract, SLPI has a local protective function against proteolytic tissue degradation during inflammation. In the female reproductive tract, SLPI and PI3 contribute to the innate defenses that prevent infection, which can compromise both implantation and pregnancy (Williams et al. 2006). The function, expression pattern, and evolutionary dynamics of the serine proteinase inhibitors encoded by genes in this locus are reminiscent of genes encoding a number of endogenous antimicrobial peptides (e.g., β-defensins and α-defensins), which are also found in mucosal secretions, participate in the first line of defense against invading microorganisms, and have evolved by successive rounds of duplication followed by substantial divergence involving positive selection (Williams et al. 2006). Interestingly, a correlation between immune response (using white blood cell count as an indicator) and primate mating systems has been suggested (Nunn et al. 2000; Anderson et al. 2004). Although the risk of sexually transmitted infections may be greater in species with mating systems that typically involve multiple partners, other factors (e.g., group size, population density, and terrestrial vs. arboreal habitats) can also affect the extent of exposure to infectious agents via the skin or mucosa. Also to be considered, ejaculation elicits immune system-mediated (both cellular and humoral) destruction of sperm post-coitus (Denison et al. 1999). The anti- and pro-inflammatory responses that take place in the uterus and cervix in an attempt to promote sperm survival (male) and/or removal (female) could also impact the evolutionary dynamics of the *WFDC/SEMG* locus. In short, the evolutionary forces that have driven the rapid diversification of *WFDC* and *SEMG* genes may be related to the different mating systems, the dynamics of host–pathogen interactions, and male attempts to counteract the female immune response. The challenge remains to establish the relative contributions of these (or other) selection pressures to the overall evolutionary process, with the sequence and analyses reported here providing a step in that direction.

Methods

Comparative genome sequencing

BAC clones were isolated from the following 12 libraries (see http://bacpac.chori.org), as described (Thomas et al. 2002, 2003), common chimpanzee (*Pan troglodytes*; CHORI-251), gorilla (*Gorilla gorilla*; CHORI-255), orangutan (*Pongo pygmaeus*; CHORI-253), baboon *(Papio anubis*; RPCI-41), rhesus macaque (*Macaca mulatta*; CHORI-250), black and white colobus (*Colobus guereza*; CHORI-272), vervet (*Cercopithecus aethiops*; CHORI-252), marmoset (*Callithrix jacchus*; CHORI-259), squirrel monkey (*Saimiri boliviensis*; CHORI-254), owl monkey (*Aotus nancymaae*; CHORI-258), galago (*Otolemur garnettii*; CHORI-256), and ring-tailed lemur (*Lemur catta*; LBNL-2). Specifically, each library was screened using pooled sets of oligonucleotide-based probes designed from the established sequence of the human centromeric *WFDC* sublocus (probe sequences are available upon request). After isolation and mapping, a total of 31 BACs were shotgun sequenced and subjected to sequence finishing, as described (Blakesley et al. 2004). For each species, a single nonredundant sequence was generated from the individual BAC sequences (i.e., a multi-BAC sequence assembly) using the program TPF Processor (http://www.ncbi.nlm.nih.gov/projects/zoo_seq). The resulting assemblies were manually verified and submitted to GenBank under accession numbers DP000036–DP000048 (Table 2).

The following additional sequences, each orthologous to the centromeric *WFDC* sublocus, were obtained from the UCSC Genome Browser (see http://www.genome.ucsc.edu): (1) human reference sequence (NCBI human genome sequence build 36.1, March 2006; chr20:42,777,545–43,472,662); (2) gap-filling sequences for chimpanzee (NCBI chimpanzee genome sequence build 2.1, November 2006; chr20:42,310,742–42,765,304), orangutan (GenBank AY256473), and macaque (NCBI macaque genome sequence build 1.0, January 2006; chr10:19,085,240–19,450,295); and (3) mouse (NCBI mouse genome sequence build 36, February 2006; chr2:163,765,619–164,181,967), rat (NCBI rat genome sequence build 3.4, November 2004; chr3:155,014,767–

155,490,839), and dog (NCBI dog genome sequence build 2.0, May 2005; chr24:35,352,539–35,642,063) sequences. Additionally, *SEMG* gene sequences were obtained from the following primate species: spider monkey (*Ateles geoffroyi SEMG2* [GenBank AY781393]); gibbon (*Hylobates lar SEMG2* [GenBank AY781389]; *Hylobates klossii SEMG1* and *SEMG2* [GenBank AY256474 and AY259291, respectively]); and cotton-top tamarin (*Saguinus oedipus SEMG1* [GenBank AJ002153]).

Sequence annotation and comparative analyses

The assembled sequences were annotated for gene content based on alignments to human RefSeq mRNA (or species-specific mRNA, if available) sequences using Spidey (http://www.ncbi.nlm.nih.gov/spidey). Known repetitive sequences were detected by RepeatMasker (http://www.repeatmasker.org) using appropriate repeat libraries for each species. Pairwise and multi-species sequence comparisons were performed using MultiPipMaker (Schwartz et al. 2003a). Pseudogene annotation required manual inspection of the sequence alignments aided by MultiPipMaker and blast (Altschul et al. 1990). In addition, Sequin (http://www.ncbi.nlm.nih.gov/Sequin) was used to import and confirm all annotations, including verifying splice-site consensus sequences, exon structure, and predicted protein sequences. Multi-sequence alignments were generated with each gene's coding sequence and each encoded protein using ClustalW (Chenna et al. 2003). The protein sequence was aligned first, with the coding DNA sequences then aligned according to the protein alignment. The close relationship among the studied primates allowed for the generation of high-confidence multi-sequence alignments with few gaps. In the case of multi-domain proteins, domains with the highest percentage of DNA identity were aligned.

Likelihood ratio tests for positive selection

A Kimura 2-parameter model, as implemented in MEGA3 (Kumar et al. 2004), was used to calculate mean pairwise divergence of introns and exons. For calculating the d_N/d_S ratio (ω) at sites or lineages (defined as all branches in the phylogeny, both terminal species nodes and internodes), secretion signal sequences and sequences associated with species-specific premature stop codons or frameshifts were removed. We tested for positive selection by comparing models of codon evolution that allow for variation in ω between sites (Bielawski et al. 2000; Yang et al. 2000). First, the likelihood of a nearly neutral model M1 was compared with that of a selection model M2. M1 allows for two ω ratios, a ratio fixed at 1 and another estimated between 0 and 1. M2 allows for an additional class of sites freely estimated from the data that can take on a value greater than or less than 1. Next, the likelihood of a more flexible neutral model M7 was compared with that of a selection model M8. M7 allows ω to vary between 0 and 1 in the form of a β distribution, while M8 allows for one additional ω class that can be greater than or equal to 1. Our last analysis determined if the additional class in M8 was significantly greater than 1; this was accomplished by comparing M8 to M8a, a model where the additional class had ω fixed at 1 (Swanson et al. 2003). In all cases, we used a likelihood ratio test to determine if the selection model (L_1) was a better fit to the data than the neutral model (L_0) by comparing the negative of twice the difference in the likelihoods between the two models ($-2[\log(L_0) - \log(L_1)]$) with the χ^2-distribution. While the significance of M8 to M8a could be compared to a 50:50 mixture of the χ^2-distribution and a point mass at 0, it is thought that a more conservative approach that uses only the χ^2-distribution is advisable to avoid false-positive results. Degrees of freedom were equal to the difference in the number of parameters estimated between the two models. To control for false-positive results due to multiple testing, the false-discovery rate was calculated using the Qvalue program (Storey 2002). In all cases, we checked for convergence by performing the analysis with at least three initial ω values (0.3, 1, 3). To test for variation in ω among lineages, the likelihood of a model with one ω value for all lineages was compared to the likelihood of a model with each lineage having a separate ω value using a likelihood ratio test (Nielsen and Yang 1998; Yang and Nielsen 1998). Likelihood calculations were carried out using PAML version 3.14.

Primate mating systems and seminal coagulation

Primate species were classified according to their primary mating system, although one or more secondary mating systems can occur within a species (Dixson 1991, 1997). Species with mating systems that promote low post-copulatory sperm competition are owl monkey and marmoset (monogamous) as well as gorilla and colobus (polygynous). Species with mating systems that promote high post-copulatory sperm competition are macaque, baboon, chimpanzee, lemur, vervet, and squirrel monkey (multimale–multifemale) as well as orangutan and galago (dispersed). Comparative ratings of seminal coagulation have been reported for all the primate species studied here (Dixson and Anderson 2002), with the exception of vervet, squirrel monkey, and colobus (see Fig. 3).

Acknowledgments

We thank Phil Green, Evan Eichler, Michael Zody, and Pascal Gagneux for helpful comments about this manuscript. This research was supported in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Anderson, M.J., Hessel, J.K., and Dixson, A.F. 2004. Primate mating systems and the evolution of immune response. *J. Reprod. Immunol.* **61:** 31–38.

Barton, N.H. 1995. Linkage and the limits to natural selection. *Genetics* **140:** 821–841.

Bielawski, J.P., Dunn, K.A., and Yang, Z. 2000. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. *Genetics* **156:** 1299–1308.

Blakesley, R.W., Hansen, N.F., Mullikin, J.C., Thomas, P.J., McDowell, J.C., Maskeri, B., Young, A.C., Benjamin, B., Brooks, S.Y., Coleman, B.I., et al. 2004. An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res.* **14:** 2235–2244.

Borgono, C.A., Michael, I.P., and Diamandis, E.P. 2004. Human tissue kallikreins: Physiologic roles and applications in cancer. *Mol. Cancer Res.* **2:** 257–280.

Bourgeon, F., Evrard, B., Brillard-Bourdet, M., Colleu, D., Jegou, B., and Pineau, C. 2004. Involvement of semenogelin-derived peptides in the antibacterial activity of human seminal plasma. *Biol. Reprod.* **70:** 768–774.

Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J., and Nickerson, D.A. 2005. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* **15:** 1553–1565.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., and Thompson, J.D. 2003. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31:** 3497–3500.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Clauss, A., Lilja, H., and Lundwall, A. 2002. A locus on human chromosome 20 contains several genes expressing protease inhibitor

domains with homology to whey acidic protein. *Biochem. J.* **368:** 233–242.
Clauss, A., Lilja, H., and Lundwall, A. 2005. The evolution of a genetic locus encoding small serine proteinase inhibitors. *Biochem. Biophys. Res. Commun.* **333:** 383–389.
Denison, F.C., Grant, V.E., Calder, A.A., and Kelly, R.W. 1999. Seminal plasma components stimulate interleukin-8 and interleukin-10 release. *Mol. Hum. Reprod.* **5:** 220–226.
Dixson, A.F. 1991. Sexual selection, natural selection and copulatory patterns in male primates. *Folia Primatol. (Basel)* **57:** 96–101.
Dixson, A.F. 1997. Evolutionary perspectives on primate mating systems and behavior. *Ann. N.Y. Acad. Sci.* **807:** 42–61.
Dixson, A.F. and Anderson, M.J. 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. *Folia Primatol. (Basel)* **73:** 63–69.
Dixson, A.F. and Anderson, M.J. 2004. Sexual behavior, reproductive physiology and sperm competition in male mammals. *Physiol. Behav.* **83:** 361–371.
Dorus, S., Evans, P.D., Wyckoff, G.J., Choi, S.S., and Lahn, B.T. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nat. Genet.* **36:** 1326–1329.
Eyre-Walker, A. and Hurst, L.D. 2001. The evolution of isochores. *Nat. Rev. Genet.* **2:** 549–555.
Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.* **125:** 1–15.
Furutani, Y., Kato, A., Yasue, H., Alexander, L.J., Beattie, C.W., and Hirose, S. 1998. Evolution of the trappin multigene family in the Suidae. *J. Biochem.* **124:** 491–502.
Furutani, Y., Kato, A., Fibriani, A., Hirata, T., Kawai, R., Jeon, J.H., Fujii, Y., Kim, I.G., Kojima, S., and Hirose, S. 2005. Identification, evolution, and regulation of expression of Guinea pig trappin with an unusually long transglutaminase substrate domain. *J. Biol. Chem.* **280:** 20204–20215.
Goodman, M., Porter, C.A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C.P. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* **9:** 585–598.
Gowaty, P.A. and Karlin, A.A. 1984. Multiple maternity and paternity in single broods of apparently monogamous eastern bluebirds. *Behav. Ecol. Sociobiol.* **15:** 91–95.
Hagiwara, K., Kikuchi, T., Endo, Y., Huqun, Usui, K., Takahashi, M., Shibata, N., Kusakabe, T., Xin, H., Hoshi, S., et al. 2003. Mouse SWAM1 and SWAM2 are antibacterial proteins composed of a single whey acidic protein motif. *J. Immunol.* **170:** 1973–1979.
Hagstrom, J.E., Fautsch, M.P., Perdok, M., Vrabel, A., and Wieben, E.D. 1996. Exons lost and found. Unusual evolution of a seminal vesicle transglutaminase substrate. *J. Biol. Chem.* **271:** 21114–21119.
Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* **307:** 1072–1079.
International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.
Jensen-Seaman, M.I. and Li, W.H. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *J. Mol. Evol.* **57:** 261–270.
Johnson, M.E., Viggiano, L., Bailey, J.A., Abdul-Rauf, M., Goodwin, G., Rocchi, M., and Eichler, E.E. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413:** 514–519.
Kim, Y. and Stephan, W. 2003. Selective sweeps in the presence of interference among partially linked loci. *Genetics* **164:** 389–398.
Kingan, S.B., Tatar, M., and Rand, D.M. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. *J. Mol. Evol.* **57:** 159–169.
Kumar, S., Tamura, K., and Nei, M. 2004. MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* **5:** 150–163.
Lin, H.J., Lee, C.M., Luo, C.W., and Chen, Y.H. 2005. Functional preservation of duplicated pair for RSVS III gene in the REST locus of rat 3q42. *Biochem. Biophys. Res. Commun.* **326:** 355–363.
Lundwall, A. and Lazure, C. 1995. A novel gene family encoding proteins with highly differing structure because of a rapidly evolving exon. *FEBS Lett.* **374:** 53–56.
Lundwall, A. and Olsson, A.Y. 2001. Semenogelin II gene is replaced by a truncated line 1 repeat in the cotton-top tamarin. *Biol. Reprod.* **65:** 420–425.
Lundwall, A. and Ulvsback, M. 1996. The gene of the protease inhibitor SKALP/elafin is a member of the REST gene family. *Biochem. Biophys. Res. Commun.* **221:** 323–327.
Metz, E.C., Robles-Sikisaka, R., and Vacquier, V.D. 1998. Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. *Proc. Natl. Acad. Sci.* **95:** 10676–10681.
Nei, M. and Rooney, A.P. 2005. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* **39:** 121–152.
Nielsen, R. and Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148:** 929–936.
Nunn, C.L., Gittleman, J.L., and Antonovics, J. 2000. Promiscuity and the primate immune system. *Science* **290:** 1168–1170.
Peter, A., Lilja, H., Lundwall, A., and Malm, J. 1998. Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase. *Eur. J. Biochem.* **252:** 216–221.
Robert, M. and Gagnon, C. 1999. Semenogelin I: A coagulum forming, multifunctional seminal vesicle protein. *Cell. Mol. Life Sci.* **55:** 944–960.
Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. 2006. Positive natural selection in the human lineage. *Science* **312:** 1614–1620.
Schalkwijk, J., Wiedow, O., and Hirose, S. 1999. The trappin gene family: Proteins defined by an N-terminal transglutaminase substrate domain and a C-terminal four-disulphide core. *Biochem. J.* **340:** 569–577.
Schopfer, C.R., Nasrallah, M.E., and Nasrallah, J.B. 1999. The male determinant of self-incompatibility in Brassica. *Science* **286:** 1697–1700.
Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., Green, E.D., Hardison, R.C., and Miller, W. 2003a. MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31:** 3518–3524.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. 2003b. Human-mouse alignments with BLASTZ. *Genome Res.* **13:** 103–107.
Storey, J.D. 2002. A direct approach to false discovery rates. *J. R. Stat. Soc. [Ser B]* **64:** 479–498.
Swanson, W.J., Nielsen, R., and Yang, Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol. Biol. Evol.* **20:** 18–20.
Thomas, J.W., Prasad, A.B., Summers, T.J., Lee-Lin, S.Q., Maduro, V.V., Idol, J.R., Ryan, J.F., Thomas, P.J., McDowell, J.C., and Green, E.D. 2002. Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* **12:** 1277–1285.
Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.
Ulvsback, M. and Lundwall, A. 1997. Cloning of the semenogelin II gene of the rhesus monkey. Duplications of 360 bp extend the coding region in man, rhesus monkey and baboon. *Eur. J. Biochem.* **245:** 25–31.
Ulvsback, M., Lazure, C., Lilja, H., Spurr, N.K., Rao, V.V., Loffler, C., Hansmann, I., and Lundwall, A. 1992. Gene structure of semenogelin I and II. The predominant proteins in human semen are encoded by two homologous genes on chromosome 20. *J. Biol. Chem.* **267:** 18080–18084.
Vallender, E.J. and Lahn, B.T. 2004. Positive selection on the human genome. *Hum. Mol. Genet.* **13 (Spec. No. 2):** R245–R254.
Wang, E.T., Kodama, G., Baldi, P., and Moyzis, R.K. 2006. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl. Acad. Sci.* **103:** 135–140.
Williams, S.E., Brown, T.I., Roghanian, A., and Sallenave, J.M. 2006. SLPI and elafin: One glove, many fingers. *Clin. Sci. (Lond.)* **110:** 21–35.
Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46:** 409–418.
Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155:** 431–449.

Received September 30, 2006; accepted in revised form December 7, 2006.