Comparative Population Genomics of the Ejaculate in Humans and the Great Apes

Jeffrey M. Good,[*,1,2] Victor Wiebe,[1] Frank W. Albert,[‡,1] Hernán A. Burbano,[§,1] Martin Kircher,[‖,1] Richard E. Green,[¶,1] Michel Halbwax,[#,1] Claudine André,[3] Rebeca Atencia,[4] Anne Fischer,[‡‡,1] and Svante Pääbo[1]

[1]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2]Division of Biological Sciences, University of Montana

[3]Lola ya Bonobo Sanctuary, Kinshasa, Democratic Republic of Congo

[4]Réserve Naturelle Sanctuaire à Chimpanzés de Tchimpounga, Jane Goodall Institute, Pointe-Noire, Republic of Congo

[‡]Present address: Princeton University, Lewis Sigler Institute for Integrative Genomics, Princeton, NJ

[§]Present address: Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tüebingen, Germany

[‖]Present address: Department of Genome Sciences, University of Washington School of Medicine

[¶]Present address: Department of Biomolecular Engineering, University of California at Santa Cruz

[#]Present address: EcoHealth Alliance, New York, NY

[‡‡]Present address: International Center for Insect Physiology and Ecology, Nairobi, Kenya

***Corresponding author:** E-mail: jeffrey.good@mso.umt.edu.

Associate editor: Willie Swanson

All Illumina data have been deposited in the Sequence Read Archive (accession no. ERP002136).

Abstract

The rapid molecular evolution of reproductive genes is nearly ubiquitous across animals, yet the selective forces and functional targets underlying this divergence remain poorly understood. Humans and closely related species of great apes show strongly divergent mating systems, providing a powerful system to investigate the influence of sperm competition on the evolution of reproductive genes. This is complemented by detailed information on male reproductive biology and unparalleled genomic resources in humans. Here, we have used custom microarrays to capture and sequence 285 genes encoding proteins present in the ejaculate as well as 101 randomly selected control genes in 21 gorillas, 20 chimpanzees, 20 bonobos, and 20 humans. In total, we have generated >25× average genomic coverage per individual for over 1 million target base pairs. Our analyses indicate high levels of evolutionary constraint across much of the ejaculate combined with more rapid evolution of genes involved in immune defense and proteolysis. We do not find evidence for appreciably more positive selection along the lineage leading to bonobos and chimpanzees, although this would be predicted given more intense sperm competition in these species. Rather, the extent of positive and negative selection depended more on the effective population sizes of the species. Thus, general patterns of male reproductive protein evolution among apes and humans depend strongly on gene function but not on inferred differences in the intensity of sperm competition among extant species.

Key words: **molecular evolution, sexual selection, male reproduction, sperm competition.**

Article

Introduction

The rapid divergence of reproductive genes is a general evolutionary pattern in animals (Coulthart and Singh 1988; Vacquier et al.1997; Begun et al. 2000; Wyckoff et al. 2000; Swanson et al. 2003; Gibbs et al. 2004; Clark and Swanson 2005; Andres et al. 2006). Several different forms of sexual and natural selection likely contribute to this, including sperm competition among males (Parker 1970), antagonistic sexual conflict between males and females (Rice 1996; Holland and Rice 1999), cryptic female choice (Eberhard 1996), and immune defense of sexually transmitted pathogens (Nunn et al. 2000). However, the relative contribution and functional consequences of these diverse evolutionary pressures remain unclear.

Most models of sexual selection predict that the intensity of positive selection should depend strongly on the mating system of a given species, with the most intense selection predicted to occur in species where females commonly mate with multiple males over a short time. Consistent with this, the evolution of numerous male and female reproductive phenotypes has been tied to mating behavior (Andersson 1994; Eberhard 1996; Dixson 1998; Markow 2002) and the degree of female promiscuity across species is positively correlated with the evolution of several male reproductive phenotypes (Harcourt et al. 1981; Dixson 1987; Harcourt et al. 1995; Anderson et al. 2005). Despite this phenotypic trend, there are relatively few examples where the molecular evolution of the underlying reproductive

genes is correlated with variation in the presumed intensity of sexual selection across species (Dorus et al. 2004; Herlyn and Zischler 2007; Ramm et al. 2008; O'Connor and Mundy 2009; Wong 2010; but see Hamm et al. 2007). Thus, the general role of mating behavior or mating system in determining the intensity of selection at the molecular level remains unclear (Wong 2011).

Genes involved in reproduction span a broad array of functions, and it is likely that the intensity of sexual selection and other evolutionary pressures varies across different aspects of reproduction. Function-dependent heterogeneity in the evolution of reproductive genes has been well documented in *Drosophila*. Ejaculated seminal fluid proteins (SFPs) produced by male accessory glands play an integral role in female reproductive physiology, behavior, and immune response (Wolfner 1997; Chapman 2001) and directly influence the fertilization success of competing males (Clark et al. 1995; Clark and Begun 1998). Genes encoding SFPs are among the most rapidly evolving genes in *Drosophila* genomes (Tsaur et al. 1998; Aguade 1999; Begun et al. 2000; Mueller et al. 2005), while genes directly involved in sperm development tend to be under stronger functional constraint and have much slower rates of gene evolution (Dorus et al. 2006).

In contrast, most molecular data on the rapid evolution of reproduction in mammals come from genes involved in spermatogenesis (Wyckoff et al. 2000; Good and Nachman 2005; Wong 2010) and fertilization (Torgerson et al. 2002; Swanson et al. 2003; Torgerson et al. 2005). The molecular evolution of SFPs is less well known. Mammalian seminal proteins are secreted from several tissues including the epididymides, seminal vesicles, the prostate, coagulating glands, and bulbourethral glands. Once ejaculated, this diverse set of proteins participates in a suite of reproductive processes including sperm motility and viability (Henault and Killian 1996), semen coagulation (Shivaji et al. 1990), and mediation of the female immune response (Robertson 2005, 2007). Many of these functional classes parallel rapidly evolving sets of *Drosophila* SFPs (Mueller et al. 2004), and some seminal proteins show evidence for pervasive positive selection in primates (Clark and Swanson 2005) and mice (Dean et al. 2009). However, many proteins in the mouse ejaculate are highly conserved and subject to strong functional constraint (Dean et al. 2008, 2009). Thus, the form and intensity of selection may vary considerably across species and different functional classes of mammalian reproductive genes (see also Jensen-Seaman and Li 2003; Kingan et al. 2003; Dorus et al. 2004, 2010).

Human fertility has been extensively studied at the molecular level, providing detailed insights into which genes are involved in various reproductive processes. Moreover, ape species closely related to humans show striking divergence in mating systems (fig. 1; Dixson 1998), presenting a compelling framework for testing basic predictions of sexual selection theory. We designed a custom microarray to capture and sequence the exons of 386 genes (including 285 SFPs; supplementary table S1, Supplementary Material online) from 20 humans, 21 gorillas, 20 chimpanzees, and 20 bonobos

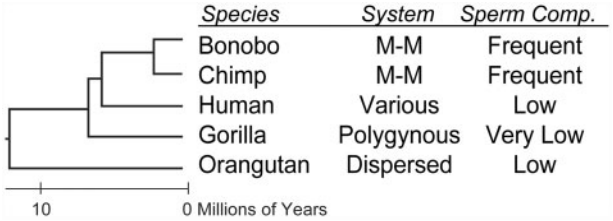

FIG. 1. Phylogeny, mating systems, and relative frequency of sperm competition in the great apes (M-M = multiple male – multiple female). The relative frequency of sperm competition is based on the estimated number of males that females from each species mate with during the fertile phase of a single ovulatory cycle (Dixson 1998).

(supplementary table S2, Supplementary Material online). We use these population genomic data to address two complementary questions. First, we test if the direction and intensity of genic selection varies across the male ejaculate in order to identify functional components that are the most frequent targets of positive selection. Second, we test if SFPs have experienced more positive selection in species with high rates of female remating and, consequently, sperm competition (i.e., bonobos and chimpanzees).

Results and Discussion

Exon Captures

We used data from a proteomic study on the human ejaculate (Pilch and Mann 2006) to identify a representative set of 285 SFPs suitable for high-throughput targeted resequencing in humans and the great apes. This set included most of the highly abundant SFPs that comprise the core functional components of the ejaculate (see supplementary fig. S1 and supplementary table S3 for details on SFP gene selection, Supplementary Material online). As controls we also randomly selected 101 genes not present in the ejaculate proteome. Following Hodges et al. (2009), we then designed a custom Agilent SureSelect 244K capture array to target all exons for each of the 386 genes (total target size of 1,016,257 bp).

Initial array captures of single genomic libraries prepared from 15 central chimpanzees and 1 western gorilla resulted in between 43% and 60% of mapped single-end reads (average 54%) overlapping with the 1,016,257 positions that were targeted on the array, representing an ~1,300-fold enrichment. Over 99% of targets were covered by at least one Illumina read. Capture performance for the single gorilla was within the range of results observed among the 15 chimpanzee samples (supplementary fig. S2, Supplementary Material online). These results verify that the sequence divergence between these species was not an impediment to the performance of the capture array, consistent with previous studies (George et al. 2011; Bi et al. 2012). We then proceeded to combine multiple samples (2–20 individuals) onto single capture arrays (multiplexed captures). Initial experiments resulted in a ~3-fold drop in capture efficiency (i.e., 15–20% of reads in target; supplementary fig. S2, Supplementary Material online). This is likely due in part to the use of

longer adapter sequences used in paired-end sequencing, which may increase cross-hybridization to nontarget molecules (Hodges et al. 2009). We optimized aspects of the library preparation and hybridization protocol (see Meyer and Kircher 2010) so that multiplexed captures of up to 20 individuals had on-target efficiencies of 55–59%, similar to the capture experiments on single individuals (supplementary fig. S2, Supplementary Material online). The final dataset included 81 individuals across the four species. All individuals were represented by at least 26× average sequence coverage (range 26–116×), with an average individual coverage per species of 53× or greater (supplementary table S4, Supplementary Material online). Unless otherwise noted, we focus on the subset of genes for which there is a 1:1 ortholog in each of the three species with annotated genomes (gorilla, human, and chimpanzee) to facilitate comparisons.

Genetic Variation in Humans and Three Species of Great Apes

When all 1:1 orthologous genes are analyzed together, central chimpanzees were found to have the highest levels of diversity (table 1), approaching 0.3% (θ_w) at 4-fold degenerate positions. In contrast, bonobos were the least variable but similar to humans. Assuming an autosomal mutation rate of 2.5×10^{-8} per site per generation (Nachman and Crowell 2000), effective population sizes vary between less than 9,000 in bonobos to nearly 28,000 in central chimpanzees. Thus, the overall patterns of diversity are similar to previous results in humans (Wall et al. 2008) and captive-born apes (central chimpanzees and bonobos, Fischer et al. 2011; western gorillas, Fischer et al. 2006; Thalmann et al. 2007).

All four populations show negative estimates of Tajima's *D* (table 1), indicating that the site frequency spectra are skewed towards rare alleles. In principle, this could reflect positive and negative natural selection and/or nonequilibrium demography (e.g., cryptic population subdivision or a population expansion). The most striking pattern was found in central chimpanzees, where numerous previous studies have also indicated a strong signature of a past population expansion (Caswell et al. 2008; Fischer et al. 2011; Hvilsom et al. 2012). Our estimates of nucleotide variability and Tajima's *D* in the human samples are consistent with previous genic data in Yorubans (Plagnol and Wall 2006). Tajima's *D* for bonobos and central chimpanzees was also similar to intergenic data from the same individuals, while our estimates of nucleotide variability are marginally lower (Fischer et al. 2011). One possible explanation for this is that purifying selection at linked sites is expected to slightly reduce the overall effective population size of genic regions relative to less constrained intergenic regions.

We analyzed each of the four species using STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) to test for cryptic population structure. We present the primary findings of these analyses here with an expanded discussion of the results available in the supplementary analyses, Supplementary Material online. All four species showed some deviations from an equilibrium population that could be caused by subdivision or admixture (supplementary table S5, Supplementary Material online). We found statistical evidence for multiple populations under a model of admixture within our sample of Yorubans; however, this signature does not appear to reflect true subdivision within the sample based on several common metrics (see supplementary analyses, Supplementary Material online). We found stronger evidence for subdivision within chimpanzees, gorillas, and bonobos. In chimpanzees, the signal for structure derived primarily from four individuals that we had previously determined to have a significant portion of ancestry from adjacent (and closely related) eastern chimpanzee populations (*Pan troglodytes schweinfurthii*; Fischer et al. 2011). In gorillas, the signature of subdivision stems largely from a single individual derived from the Cross River area of Cameroon. Cross River gorillas (*Gorilla gorilla diehli*) are a critically endangered population that appears to have become isolated from other western gorilla populations quite recently (~18K years; Thalmann et al. 2011). Thus, cryptic structure within our samples of chimpanzees and gorillas derive from previously identified geographic partitions.

Genome-wide patterns of genetic structure within natural populations of bonobos have not been documented previously. We found a consistent signature of population subdivision within our sample of 20 bonobos with strong support for models with three distinct populations (supplementary fig. S3 and supplementary table S6, Supplementary Material online). Population genetic subdivision within bonobos is noteworthy given the relatively small and contiguous nature of their current distribution within the Congo basin. Multiple mitochondrial (mtDNA) lineages have been documented within bonobos (Zsurka et al. 2010; Fischer et al.

Table 1. Nucleotide Diversity.

Species	Sites[a]	$\theta_{w\,All}$[b] (%)	$\theta_{\pi\,All}$[b] (%)	$\theta_{w\,Silent}$[c] (%)	$\theta_{\pi\,Silent}$[c] (%)	Tajima's *D*[d] (95% CI)	N_e[e]
Bonobo	**471,534**	**0.046**	**0.037**	**0.088**	**0.074**	**−0.441 (−0.15, −0.74)**	**8,800**
Chimpanzee	**455,634**	**0.125**	**0.079**	**0.279**	**0.188**	**−0.962 (−0.81, −1.15)**	**27,900**
Gorilla	**466,224**	**0.079**	**0.069**	**0.155**	**0.143**	**−0.344 (−0.11, −0.55)**	**15,500**
Human	**470,193**	**0.058**	**0.044**	**0.104**	**0.088**	**−0.645 (−0.40, −0.90)**	**10,400**

[a]Total number of protein coding positions in 319 1:1 orthologous autosomal genes.
[b]Estimated from all protein-coding autosomal positions.
[c]Estimated from 4-fold degenerate autosomal positions.
[d]Estimated from 4-fold degenerate autosomal positions using a bootstrap procedure (see text).
[e]Approximate effective population size assuming $\theta_{w\,Silent} = 4N_e\mu$ and a mutation rate of 2.5×10^{-8} (per site per generation).

2011), though this variation does not show strong geographic structuring (Eriksson et al. 2004). Our data suggest that significant structure exists across the nuclear genome in contemporary populations of bonobos. It is unclear if or how this structure partitions across the current geographic range of bonobos because our samples derive from unknown locations in the Democratic Republic of Congo.

Cryptic subdivision and/or admixture could impact our analyses that rely on estimates of polymorphism within each species. To account for this, we discuss results from all individuals as well as from subsets of the data where structure has been removed. For chimpanzees and gorillas, this involved removing the five individuals inferred to derive from closely related populations. For bonobos, we analyzed the subset of individuals ($N = 9$) comprising the largest single group under the best-fit model inferred from structure ($K = 3$).

Molecular Evolution of the Ejaculate

To examine rates of protein evolution across the four great apes (including orangutan) and humans, we first used parsimony to reconstruct an ancestral sequence from the population data within each of the four focal species in order to reduce the contribution of polymorphic positions to sequence divergence. We then used a maximum likelihood (ML) framework (Yang 2007) to estimate rates of protein evolution (dN/dS) and to test for positive directional selection (Yang and Nielsen 2000; Yang and Swanson 2002) for 298 genes (excluding genes without an annotated or intact 1:1 ortholog in the orangutan). Median rates of protein evolution were similar between the randomly selected set of control genes and the SFP reproductive genes (table 2). The incidence of positive selection was more than twice as frequent among the ejaculate proteins as among the control genes (10.0% vs. 3.9%, $\alpha = 0.05$, not corrected for multiple tests). Although only marginally significant (Fisher's exact test [FET] $P = 0.07$), this suggests a trend toward more positive selection on SFPs. As is common for genomic studies of molecular evolution, only two genes (both SFPs) remained significant following a conservative Bonferroni correction for multiple tests.

Next, we contrasted the ratio of polymorphism to divergence at amino acid changing (nonsynonymous) and silent (synonymous) positions across the entire phylogeny (chimpanzees, bonobos, humans, and gorillas). These phylogeny-wide analyses incorporate polymorphism information from all four species and divergence from all of the branches in the four species phylogeny (orangutan was not included in this contrast). They thus describe the overall mode of molecular evolution of genes and are not informative regarding lineage-specific patterns of evolution. Under neutrality, the ratio of polymorphism to divergence should be equivalent between these two site classes (McDonald and Kreitman 1991). Deviations from this expectation can be quantified using the neutrality index (NI), which is the ratio of polymorphism to divergence between nonsynonymous (pN/dN) and synonymous (pS/dS) sites, respectively. Assuming that silent positions reflect the neutral equilibrium condition, positive selection should result in an excess of amino acid divergence ($NI < 1$), while weak purifying selection should generate an excess of amino acid polymorphism ($NI > 1$). An excess of nonsynonymous divergence (dN) relative to polymorphism (pN) is considered evidence for positive directional selection, whereas an excess of nonsynonymous polymorphism (pN) suggests an excess of slightly deleterious polymorphisms. Here, we use an unbiased estimator of the NI for all analyses (NI_{TG}; Stoletzki and Eyre-Walker 2011). We observed a significant excess of nonsynonymous polymorphisms relative to divergence across all genes ($NI_{TG} = 1.31$ [95% CI = 1.18–1.47]; pooled $pN = 2{,}817$, $dN = 1{,}559$, $pS = 3{,}622$, $dS = 2{,}644$; FET $P \ll 0.0001$), consistent with a dominant influence of purifying selection shaping protein-coding variation. Similar results were obtained when individuals contributing to population substructure were removed ($NI_{TG} = 1.30$ [95% CI = 1.17–1.45]). Previous studies on humans have found that most genes show an excess of nonsynonymous polymorphism relative to divergence between species when compared with the ratio of polymorphism to divergence at synonymous sites (Fay et al. 2001; Bustamante et al. 2005). This pattern presumably reflects the occurrence of slightly deleterious mutations segregating within species that are eventually removed by purifying selection and thus do not contribute to divergence.

Both the control genes and the reproductive genes showed an excess of nonsynonymous polymorphism relative to nonsynonymous divergence between species (table 2). However, this tended on average to be more pronounced in the control genes ($NI_{TG} = 1.40$) when compared with reproductive genes ($NI_{TG} = 1.29$). In principle, this could reflect differences in the intensity of positive and/or negative selection in these two sets of genes. Several researchers have

Table 2. Polymorphism and Divergence.

Divergence only				
	N	*dN/dS* (SE)	PosSel[a]	PosSel (corr[b])
All genes	298[c]	0.31 (0.03)	25 (8.4%)	2 (<1%)
Control	77	0.28 (0.06)	3 (3.9%)	0 (0%)
Reproduction	221	0.32 (0.03)	22 (10.0%)	2 (<1%)
Polymorphism and divergence				
	N	Pooled NI (95% CI)	PosSel[a]	NegSel
All sites				
All genes	327	1.31 (1.18–1.47)	6 (1.8%)	10 (3.1%)
Control	86	1.40 (1.11–1.81)	1 (1.2%)	4 (4.7%)
Reproduction	241	1.29 (1.15–1.45)	5 (2.1%)	6 (2.5%)
No singletons[d]				
All genes	327	1.15 (1.03–1.28)	4 (1.2%)	7 (2.1%)
Control	86	1.15 (0.93–1.50)	1 (1.2%)	2 (2.3%)
Reproduction	241	1.16 (1.02–1.31)	3 (1.2%)	5 (2.1%)

[a]All tests based on $\alpha = 0.05$.
[b]Number of significant genes after a Bonferroni correction.
[c]The inclusion of orangutan led to the removal of several genes due to uncertain orthology or other quality filters. Differences in total counts reflect this added level of filtering.
[d]Excluding all single polymorphisms.

proposed excluding low-frequency variants as a means to overcome the strong influence of purifying selection on tests that rely on contrasts of polymorphism and divergence because most slightly deleterious mutations are at low frequencies within a population (Eyre-Walker 2006). Consistent with this, we observed a reduction in the NI when all single polymorphisms (i.e., those SNPs sampled as heterozygous in a single individual) were removed from our data (table 2). Moreover, the slight difference between reproductive and control genes disappeared when single polymorphisms were excluded, indicating that the trend toward a higher NI_{TG} at control genes primarily reflects relatively more deleterious polymorphisms segregating in these genes.

Despite a global signature of purifying selection, only 10 (6 reproductive) genes show a significant excess of nonsynonymous polymorphism relative to divergence in gene-by-gene McDonald–Kreitman tests ($\alpha = 0.05$). The per-gene signal for positive selection was even weaker, with only six (five reproductive) genes showing an excess of protein divergence relative to polymorphism. However, no genes were significant after correcting for multiple tests. While removal of low-frequency variants (singletons) does indeed lead to a decrease in the overall NI, we still detect essentially no signal for positive directional selection using the McDonald–Kreitman test. These findings are consistent with the fact that this framework has little power to detect positive selection when levels of polymorphism and divergence are low (Li et al. 2008), as is the case for great apes and humans (table 1).

These general patterns differ somewhat from several divergence-based genomic studies in mammals that have found both a higher average rate of protein evolution and incidence of positive selection for testis-expressed genes (Torgerson et al. 2002; Good and Nachman 2005; Torgerson et al. 2005; Wong 2010). Mammalian SFPs are secreted from diverse tissues and are involved in various processes including sperm motility and viability (Henault and Killian 1996), semen coagulation (Shivaji et al. 1990), signal transduction and mediation of the female immune response (Robertson 2005, 2007), and defense against pathogens. Given this diversity in gene function, it is plausible that the intensity of positive selection is highly heterogeneous across mammalian SFPs. To determine if SFP molecular evolution is influenced by gene function, we identified functional groups of genes based on Gene Ontology terms. Of the 22 genes with some evidence for recurrent positive selection based on *dN*/*dS* analyses (table 2, divergence only), 13 are involved in protein binding and/or are peptidases (*ALB*, *ANPEP*, *CD44*, *CD59*, *CTSF*, *CTSG*, *FN1*, *LTF*, *PIP*, *SMPDL3A*, *TF*, *TPP1*, *VTN*), 6 are involved in signal transduction (*CD44*, *CD59*, *GRN*, *HPX*, *TF*, *VTN*), and 4 are involved in immune responses and defense against pathogens (*CTSG*, *HPX*, *LTF*, *VTN*).

To further explore the influence of gene function on overall patterns of molecular evolution, we used gene ontology annotation to identify five groups of genes that represent major functions of the ejaculate (immune response, cell adhesion, peptidase, peptidase inhibition, and signal transduction). Note that these functions are not mutually exclusive and a gene can be included in more than one of these groups.

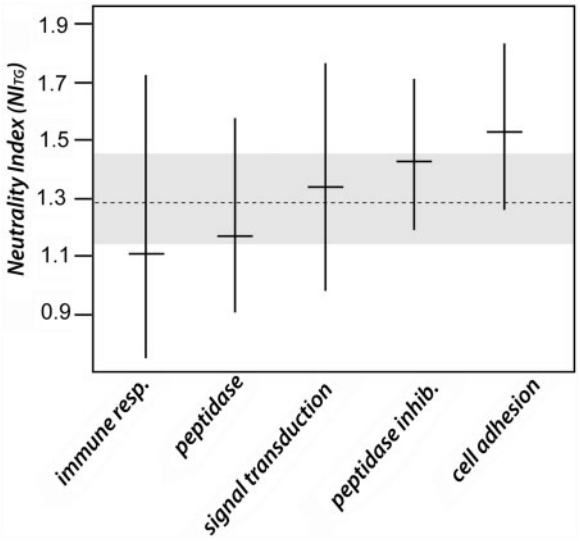

Fig. 2. Neutrality index (NI_{TG}) for genes associated with five different gene ontology categories. Groups are not mutually exclusive and error bars represent 95% confidence intervals. The mean (dotted line) and 95% confidence interval (gray shading) of NI_{TG} for all reproductive genes are indicated.

We then calculated the neutrality index (NI_{TG}) for each of the functional categories and found that patterns of polymorphism to divergence differed among the groups (fig. 2). Genes involved in immunity or peptidase activity showed the lowest NI_{TG} values, while genes involved in peptidase inhibition and cell adhesion tended towards higher NI_{TG} values. These patterns are broadly consistent with our gene-by-gene *dN*/*dS* analyses. That is, the categories with several rapidly evolving genes also showed the lowest NI_{TG} values.

Overall, these data indicate that while overall rates of evolution at ejaculated proteins are similar to the genome average, a subset of SFPs involved in immunity and the breakdown of proteins are more rapidly evolving. Interestingly, many of these positively selected functional groups parallel rapidly evolving sets of *Drosophila* SFPs (Mueller et al. 2004). Heterogeneity in the degree of functional constraint and/or positive selection across different functional groups is also found in genes expressed in the mouse epididymis (Dean et al. 2008) and the mouse sperm proteome (Dorus et al. 2010), with the most rapidly evolving genes in both studies tending to be those involved in immune or proteolytic functions. It is unclear if the subset of rapidly evolving SFPs derives primarily from specific male accessory tissues in humans and apes, but SFPs under positive selection in mice are primarily secreted from the seminal vesicles (Dean et al. 2009).

Lineage and Population-Specific Ejaculate Evolution

Levels of sperm competition differ greatly among great apes (fig. 1), with very frequent female remating in chimpanzees and bonobos compared with gorillas or humans. If sperm competition has been a major driver of positive selection at genes coding for ejaculated proteins, then one would predict more rapid protein divergence along the lineage leading to

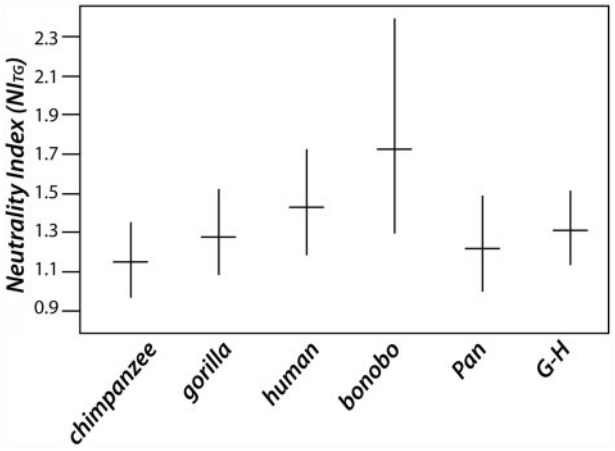

FIG. 3. Species and lineage-specific estimates of the neutrality index (NI_{TG}) for reproductive (SFP) genes. Species divergence estimates are based on pairwise comparisons to an outgroup (gorilla for human, bonobo, chimpanzee; human for gorilla). Polymorphism was estimated from the largest subset of individuals within chimpanzees, gorillas, and bonobos for which there was no evidence of population substructure. *Pan* incorporates divergence along the lineage leading to chimpanzees and bonobos and polymorphism from both species. *G-H* incorporates pairwise divergence between gorillas and humans and polymorphism from both species.

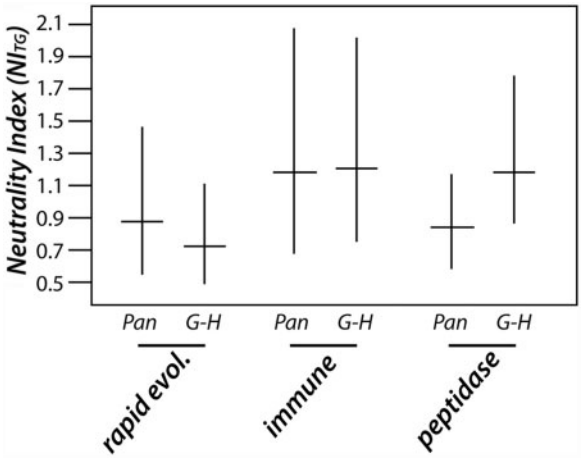

FIG. 4. Lineage-specific estimates and 95% confidence intervals for the neutrality index (NI_{TG}) of rapidly evolving or select functional subsets of reproductive (SFP) genes. The rapidly evolving gene set includes the 22 SFPs with some evidence for positive selection based on maximum likelihood analysis of *dN*/*dS* (see table 2). Genes involved in immune defense ($n = 22$) or peptidases ($n = 38$) were defined based on gene ontology annotation.

chimpanzees and bonobos. To test this, we examined patterns of polymorphism and divergence (NI_{TG}) for each of the four species. All four species showed overall NI_{TG} values that were above 1 (fig. 3, within species substructure removed) for SFPs, again indicative of an excess of slightly deleterious variation. Indeed, only estimates from chimpanzee were not significantly greater than 1 (NI_{TG} 95% CI = 0.97–1.36). If there was more positive selection in bonobos and chimpanzees, then one might expect these two species to show lower NI_{TG} values for SFPs when compared with humans and gorillas. Chimpanzees followed this general prediction with the lowest average NI_{TG}; however, bonobos showed the highest NI_{TG} of the four species. We found the same general rank order for the mean of NI_{TG} of each species in the pooled set of all genes (supplementary fig. S4, substructure removed, Supplementary Material online) and when inferred substructure is ignored within each species. The control genes were less informative when analyzed separately because of the high variance of this statistic when estimated from smaller gene sets.

Analyzing each of the four species in isolation is expected to have reduced power given their close evolutionary relationships. Moreover, high levels of promiscuity in chimpanzees and bonobos likely evolved in their common ancestor and thus the two species are not independent with respect to mating system. To provide a more direct test for intrinsic differences in patterns of evolution associated with mating system in great apes and humans, we calculated NI_{TG} of SFPs for chimpanzees and bonobos as a group (*Pan*) and gorillas and humans as a group (*G-H*). These lineage-specific estimates for SFPs yielded very similar values of NI between the two groups (fig. 3; see also supplementary analyses and supplementary fig. S5, Supplementary Material online). *Pan* tended toward lower values when considering all genes (supplementary fig. S4, Supplementary Material online), but this appears to be driven primarily by higher overall estimates of NI_{TG} for the control genes in *G-H* (*G-H* control genes $NI_{TG} = 1.61$, 95% CI = 1.27–2.18 vs. *Pan* control genes $NI_{TG} = 1.20$, 95% CI = 0.84–1.78). Thus, there is not a simple correlation between NI_{TG} and mating system.

Analysis of SFPs as a single group might obscure lineage-specific patterns that involve one or a few functional gene categories, especially given the observation of heterogeneity in rates of evolution across different functional subsets of the male ejaculate in mice (Dean et al. 2008, 2009) and the great apes (fig. 2). To examine this in more detail, we evaluated lineage-specific patterns for select subsets of genes that might reasonably be expected to be recurrent targets of positive selection. Genes involved in immune response (the most rapidly evolving functional class overall, fig. 2) appear to evolve similarly between *Pan* and *G-H* (fig. 4). Likewise, the 22 genes with the strongest evidence for positive selection based on *dN*/*dS* also appear to be rapidly evolving in both groups when incorporating variation within species. This relationship is not surprising given that the *dN*/*dS* framework we used to test for positive selection should have the most power to detect genes that are rapidly evolving on multiple branches in the phylogeny. One gene set that does tend toward more rapid evolution in *Pan* when compared with gorillas and humans are the peptidases (fig. 4; *Pan* $NI_{TG} = 0.84$, *G-H* $NI_{TG} = 1.19$). We emphasize that this trend toward lower *Pan*-specific NI_{TG} for proteases is not significant (fig. 4). Nonetheless, a trend in this direction is intriguing given this set is comprised mostly of proteases that could play an important role in the outcomes of sperm competition (Clark and Swanson 2005; Dean et al. 2009).

Table 3. Proportion (α) Amino Acids Fixed by Positive Selection.

Partition	Proportion Mutations in N_eS Categories			α (95% CI)
	0–1	1–10	>10	
Bonobo	0.30	0.04	0.66	−0.39 (−0.72, −0.07)
Bonobo (no structure)	0.30	0.04	0.66	−0.37 (−0.81, 0.19)
Chimpanzee	0.14	0.14	0.72	0.52 (0.23, 0.69)
Chimpanzee (no structure)	0.17	0.08	0.75	0.33 (0.05, 0.54)
Gorilla	0.27	0.07	0.66	−0.09 (−0.44, 0.20)
Gorilla (no structure)	0.28	0.06	0.66	−0.14 (−0.50, 0.36)
Human	0.23	0.16	0.61	0.08 (−0.39, 0.45)

Among the possible functions of the ejaculate that may be influenced by mating system, semen coagulation and copulatory plug formation have received the most attention given its role as a putative defense against sperm competition (Dixson and Anderson 2002). Multiple genes that we targeted participate in plug formation; however, this group of genes is too small to analyze with the NI_{TG} framework given the high variance of this statistic when levels of variation and/or divergence are low. Among these coagulation genes, the molecular evolution of *SEMG2* appears to show the strongest association with lineage-specific variation in sperm competition based on phylogenetic comparisons in primates (Dorus et al. 2004; O'Connor and Mundy 2009). Our data generally support the observation that *SEMG2* is rapidly evolving across humans and the great apes species when within species polymorphisms were removed ($dN/dS = 0.80$), though the support for positive selection was marginally nonsignificant (*M8* vs. *M8a*; $P = 0.084$). Interestingly, the evolution of *SEMG2* appears unexceptional within *Pan* when we consider both polymorphism and divergence (NI = 0.48; FET $P = 0.56$).

Collectively, our results indicate that differences in inferred levels of sperm competition between humans and closely related species of great apes have had a relatively small impact on the protein evolution of SFPs as a group. Even smaller subsets of SFPs that we expect are likely to play a direct role in sperm competition show weak or no differentiation related to mating system (fig. 4). This is not to imply that sperm competition does not influence the evolution of a handful of SFPs but that the overall contribution appears to be rather limited. These findings reinforce and extend previous studies that have demonstrated considerable heterogeneity in patterns of evolution across different functional components of the mammalian male reproductive system (Dean et al. 2008, 2009; Dorus et al. 2010).

Strikingly, NI_{TG} appears to scale much more closely with inferred effective population size (table 1 and fig. 1). To investigate this pattern in more detail, we used an ML approach (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009) to estimate the distribution of fitness effects of new amino acid mutations and the proportion of adaptive substitutions (α) for each species across all genes. For each species, we used the folded site frequency spectrum of 0-fold and 4-fold positions and divergence to an outgroup—a single gorilla for humans, chimpanzees, and bonobos and a single human for gorillas. This approach incorporates a simple demographic model allowing for a single step change in ancestral population size. Note that under this framework, negative values of α are possible; they imply that there is an excess of slightly deleterious variation and that rates of adaptive evolution are near 0. Only chimpanzees generated estimates of α that were significantly greater than 0 (table 3). Previous work in humans has found that the proportion of adaptive substitutions is low and near 0 (Halligan et al. 2010; but see Fay et al. 2001; Boyko et al. 2008) and our data from humans and gorillas show the same general pattern. Bonobos show strongly negative estimates of α that are in agreement with their high excess of nonsynonymous variation (fig. 3).

These general trends could reflect differences in the efficacy of natural selection across these four species, but they are also expected to be sensitive to deviations in the assumed model of population history. The simple demographic models estimated from each of the four species reflected historic growth, as expected given an excess of rare variants in the site frequency spectrum (table 1). The general pattern of population growth and moderate to no appreciable adaptive evolution persists when cryptic substructure was accounted for within chimpanzees and gorillas, respectively (table 3). Interestingly, the signal of population growth in bonobos does appear to be influenced by cryptic population subdivision. In fact, we find support for a model of population decline and a significantly positive skew in the allele frequency spectrum (Tajima's $D = 0.35$ [95% CI = 0.05–0.96]) when restricting our analyses to the largest subset of bonobos ($N = 9$) for which there is no evidence for substructure. Estimation of the contribution of adaptive evolution within this subset of bonobos is difficult given the reduced sample size and low overall levels of genetic variation, but the point estimate for α remains more negative than in the other three species (table 3).

A growing body of evidence suggests that the rate of adaptive evolution is fundamentally constrained by effective population size. Numerous studies have shown that species characterized by small populations (e.g., humans, *Arabidopsis*; Bustamante et al. 2002, 2005; Zhang and Li 2005) show much lower rates of adaptive evolution than species with very large N_e (e.g., flies, rabbits, and mice; Smith and Eyre-Walker 2002; Andolfatto 2007; Halligan et al. 2010; Carneiro et al. 2012; Phifer-Rixey et al. 2012). At face value, our results agree well with this generalization. However, there is one important caveat. Our estimates of α rely upon independent estimates

of frequency spectra but partially shared estimates of divergence. For the chimpanzees and bonobos, most of the divergence to the gorilla is shared. Given this, it is difficult to reconcile our estimates of ~33% of all substitutions driven to fixation by positive selection in chimpanzees (after removing structure) with essentially no adaptive substitutions in bonobos (table 3). It is possible that nonstationary demography in one or both species or underlying biases in our estimate of the site frequency spectra have influenced our results. However, similar results have been found in mice, where very closely related subspecies show very different estimates of the rate of adaptive substitutions in pairwise comparisons to a common outgroup (Phifer-Rixey et al. 2012). The efficacy of selection should be higher in larger populations, resulting in the more efficient removal of slightly deleterious variants as well as more effective positive selection. Therefore, large differences in α between closely related populations may simply reflect more efficient purifying selection (Phifer-Rixey et al. 2012). Under this interpretation, central chimpanzees would not necessarily differ relative to the other three species in the degree of positive selection but rather in that they carry less deleterious genetic variation.

The apparent lack of agreement between simple predictions of sexual selection and the evolution of SFPs raise several interesting questions regarding our ability to predict long-term patterns of molecular evolution from complex aspects of life history. Female chimpanzees and bonobos remate frequently (several matings per fertile cycle), whereas humans and gorillas have comparatively low rates of female promiscuity (1–2 matings per fertile cycle). Binning species into discrete categories undoubtedly masks the behavioral complexity of primate reproduction. However, the likelihood of more intense sperm competition in chimpanzees and bonobos appears overwhelming. Moreover, there is a strong correlation between mating system and primate phenotypic evolution (Clutton-Brock et al. 1977; Harcourt et al. 1981, 1995; Dixson and Anderson 2002). In apes, male chimpanzees and bonobos have large relative testis sizes (Harcourt et al. 1981, 1995) and form postcopulatory plugs (a putative defense against sperm competition; Dixson and Anderson 2002). In contrast, although gorillas show pronounced sexual dimorphism (a form of precopulatory sexual selection; Clutton-Brock et al. 1977), males have small testis and do not form copulatory plugs (Dixson and Anderson 2002). Designating a mating system for humans remains controversial but several male reproductive phenotypes appear intermediate relative to chimpanzees/bonobos and gorillas, indicating that sperm competition has been relatively infrequent during human evolution (Dixson 1998). Given all of this, the lack of a strong signal for more rapid evolution of SFPs in chimpanzees and bonobos is somewhat surprising. However, if long-term patterns of protein evolution are dominated by variation in effective population size, then comparing patterns of molecular evolution between species with different population histories may be relatively ineffective for understanding the extent to which ecological or life history variables influence genome evolution.

Another contributing factor could be that sperm competition may act primarily on the regulation of SFPs (Ramm et al. 2009; Claydon et al. 2012). For example, it is possible that the rate of production or the relative abundances of different SFPs in the male ejaculate are more directly relevant to sperm competition. Consistent with this, the protein composition of the ejaculate has been shown to be highly diverse among closely related rodent species (Ramm et al. 2009). While the relative abundance of major SFPs has been associated with human fertility and associated reproductive diseases, no comparative data are currently available on the composition of primate ejaculates. In addition, sexual selection may act more directly on phenotypes associated with sperm form, function, and production. Rates of protein evolution are higher in chimpanzees when compared with humans for testis-expressed genes (Wong 2010), and chimpanzees show a number of physiological and histological changes indicative of higher sperm production when compared with humans (Dixson 1998). Ultimately, understanding the genetic basis of sexually selected phenotypes in humans and the great apes will require the study of diverse metrics of molecular evolution.

Materials and Methods

Biological Samples

This research was approved by the European Commission (233297, TWOPAN) and was conducted following international guidelines. Veterinarians collected all blood samples from western gorillas, central chimpanzees, and bonobos during routine medical examinations, and permission for use of these samples was obtained from the Ministries of Environment and the Ministère de la Recherche Scientifique (Democratic Republic of Congo, DRC) and the Ministère de l'Enseignement Supérieur et de la Recherche Scientifique (Republic of Congo). International transportation of samples was approved following the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES).

While many previous genetic studies on great apes have relied upon samples from captive individuals, we primarily used DNA resources collected from wild-born chimpanzees, gorillas, and bonobos living in African sanctuaries (see supplementary table S2, Supplementary Material online; Fischer et al. 2011, and Thalmann et al. 2007 for more details). We collected samples from 20 bonobos (*Pan paniscus*) in the Lola ya Bonobo sanctuary near Kinshasa, DRC; 20 common central chimpanzees (*Pan troglodytes troglodytes*) from the Tchimpounga rehabilitation center near Pointe Noire, Republic of Congo; 16 wild-born western gorillas (*Gorilla gorilla*) derived primarily from the Limbe Wildlife Centre, Limbe, Cameroon, as well as 5 captive-born western gorillas from European zoos. For humans, we used 20 unrelated Yoruba from the Corriell DNA repository. DNAs from chimpanzees, bonobos, and humans were derived from Epstein-Barr virus-transformed cell lines. The 16 wild-born gorilla DNAs came from blood samples and the 5 captive-born came from postmortem tissue. Relatively low amounts of total

DNA were available for several gorilla samples so multiple displacement amplification (QIAGEN) was used prior to library construction, as previously described (Albert et al. 2007).

Choice of Targeted Genes and Microarray Design

A previous high-throughput proteomic study identified 923 high-confidence proteins in the ejaculate of a single human (Pilch and Mann 2006). Sperm were removed through centrifugation in this experiment; therefore, most of these SFPs are expected to derive from secretions of male accessory tissues, the epididymides, or to represent soluble components of the sperm cell surface (Pilch and Mann 2006). We applied a series of filters to these data to identify a core set of SFPs (supplementary fig. S1, Supplementary Material online). First, we used Ensembl annotation (release 50; www.ensembl.org, last accessed January 29, 2013) to associate the 923 proteins with 846 annotated genes. We then removed all genes without clear 1:1 orthology in chimpanzee, resulting in 746 genes. Next, we removed an additional 90 genes that had close paralogs (>80% amino acid identity) in either humans or chimpanzees. While close paralogs may be of evolutionary relevance, they are not amenable to targeted resequencing using hybridization-based approaches. These standard comparative genomic filters resulted in 656 genes suitable for comparative genomic analysis between humans, chimpanzees, bonobos, and gorillas. However, this gene set potentially contains many proteins that are not truly secreted components of the male ejaculate. Body fluids usually contain many false positives derived from epithelial shedding during secretion (Pilch and Mann 2006). To account for this, we removed genes without annotated signal peptides indicative of cellular secretion (SignalP; Emanuelsson et al. 2007) or not previously identified to encode membrane-bound vesicles secreted from the prostate (i.e., prostasomes; Utleg et al. 2003). This final filtering step yielded a core set of 285 SFPs that were enriched for more abundant proteins relative to the entire ejaculate (supplementary fig. S1, Supplementary Material online). We then tested for enrichment of Gene Ontology terms (Eden et al. 2009) and found that our filtered subset of genes had a similar functional profile as the entire proteome (supplementary table S3, Supplementary Material online; Pilch and Mann 2006). Notably, the final set of 285 SFP genes was highly enriched for proteases and included most of the highly abundant SFPs that are known to comprise the core functional components of the ejaculate (e.g., gel-forming proteins [*SEMG1*, *SEMG2*, *TGM4*, *FN1*] and kallikrein-like proteases [*KLK2*, *KLK3*, *KLK11*]). As controls we also randomly selected 101 genes from a set of 13,904 protein-coding genes whose gene products were not present in the ejaculate proteome (Pilch and Mann 2006) and had 1:1 chimpanzee orthologs with no close paralogs (>80% protein identity).

We designed a custom Agilent SureSelect 244K capture array to target all exons of each gene based on Ensembl (release 50) annotation of the chimpanzee genome (*panTro2*). The Agilent 244K platform includes 243,504 60 bp probes. We identified 3,899 individual exonic regions, collapsing exons separated by fewer than 140 bp into a single contiguous target. For each targeted exon, we included 10 bp up- and downstream of the exon boundary and evenly tiled an overlapping probe every five bases (Hodges et al. 2009). This strategy resulted in a total target size of 1,016,257 bp. To avoid targeting highly repetitive regions of the genome, we removed all probes containing 15-mers that occur more than 100 times in the chimpanzee genome (Hodges et al. 2009).

Exon Capture and Illumina Sequencing

We constructed Illumina sequencing libraries for 15 central chimpanzees and 1 western gorilla using Illumina single-end genomic library preparation kits (Illumina) and hybridized each library to a single capture array following Hodges et al. (2009). Captured products were polymerase chain reaction amplified and sequenced on a single lane of Illumina GA II, using 36, 51, or 72 bp single reads. For the remaining samples, individually barcoded paired-end libraries were prepared for each individual, and multiple individuals within a single species (up to 20) were pooled and captured on a single array, as described previously (Burbano et al. 2010). We performed multiple captures within each species, varying the number of individuals per capture and other details of the capture procedure. The final protocol has been published elsewhere (Meyer and Kircher 2010). Multiplexed capture products were sequenced with 51, 72, or 101 bp paired-end Illumina reads on the GA II platform.

Sequence Assembly and Genotype Calling

All Illumina base calls were made using Ibis (Kircher et al. 2009) and raw reads were filtered for average quality and sequence entropy. The quality of the reference genome and accompanying annotation varies considerably among sequenced apes and only humans have a reference of finished quality. To facilitate comparison across species, all quality filtered reads were mapped to the human genome sequence (*hg19*) using BWA (Li and Durbin 2009). To accommodate the mapping of divergent reads to the human reference, we increased the default maximum edit distance (*-n* = 0.02) and the maximum number of gaps (*-o* = 2). Mapped reads were filtered for uniqueness using a custom python script, and consensus genotypes were constructed for targeted regions +/− 100 bp of flanking sequence using the pileup function in SAMtools (Li et al. 2009). Only genotypes with a minimum unique coverage of 8× or more, a minimum PHRED-scaled consensus quality of 40, and an average mapping quality of at least 30 were considered. Homozygous positions that differed from the reference base were required to have a minimum SNP quality of 40 and to be at least 10 bp from insertion–deletion variants identified by the pileup routine. Putative heterozygous positions were only called if they had at least 16× coverage and a minor allele frequency of at least 0.3. We furthermore removed sites that were masked in >30% of individuals or were heterozygous in more than 75% of individuals.

Protein-coding reading frames were based on a consensus agreement between the human (*hg19*) and chimpanzee (*panTro2*) annotation. Since the onset of this project, additional genomic data (e.g., Scally et al. 2012) has become available that has further refined gene annotation between humans and the great apes. Currently, 327 (319 autosomal) of the original 386 genes are annotated as 1:1 orthologs between chimpanzee, human, and gorillas (Ensembl release 68; July 2012). Unless otherwise noted, we restrict our focus to this conservative subset of genes for all comparisons between species. For analyses including the orangutan, we retrieved homologous positions based on whole genome alignments of the UCSC genome browser. Comparisons with orangutan were based on 298 genes, excluding genes without 1:1 orthology or with premature stop codons in the orangutan genome sequence (Locke et al. 2011).

Evolutionary Analyses

Summary Statistics

All summaries of population genetic data were calculated with software implementing the libsequence library (Thornton 2003) or custom scripts written in perl or R. Standard statistics for nucleotide diversity (θ_π, Nei and Li 1979; θ_w, Watterson 1975) were estimated with compute and polydNdS, excluding positions that violated the infinite sites model (i.e., three or more nucleotide states). We calculated the skew in the site frequency spectrum with Tajima's *D* or the normalized difference between π and θ_w (Tajima 1989). Tajima's *D* assumes that the same number of chromosomes are sampled across sites; therefore, we estimated the frequency of the minor allele at each site using the randomized sampling approach proposed by Halligan et al. (2010). For each autosomal position, 20 out of 40 (chimpanzees, bonobos, and humans) or 42 (gorillas) alleles were randomly sampled without replacement.

Population Structure

We used the program STRUCTURE (Pritchard et al. 2000; Falush et al. 2003) to test for evidence of subdivision within each species. STRUCTURE uses genotype data to infer subdivision or admixture proportions based on a clustering method that assumes a defined number of populations (*K*). For each species, we randomly selected 200 four-fold degenerate synonymous SNPs and tested a strict model of no admixture and a model allowing individuals to descend from multiple populations. For each model, we considered between one to three populations (*K*), performed three replicates per *K*, and used a burn-in period of 100,000 iterations of the Markov Chain Monte Carlo (MCMC) followed by 1,000,000 MCMC iterations per replicate. All replicates were checked for convergence and repeated as necessary. For bonobos, we performed additional analyses using more SNPs and we expanded the number of potential populations to five ($K = 1–5$; see supplementary analyses, Supplementary Material online).

Divergence

To examine overall patterns of molecular evolution between species, we first used standard divergence-based approaches. For each gene, we generated four- (bonobo, chimpanzee, human, gorilla) and five- (plus orangutan, see below) species alignments derived from our mapping assemblies. We then used a ML framework (CODEML, PAML 4.0; Yang 2007) to estimate rates of protein evolution (*dN*/*dS*) and to test for positive directional selection (Yang and Nielsen 2000; Yang and Swanson 2002). We estimated *dN*/*dS* using two approaches. First, we fit data from each gene to two site-specific models of molecular evolution that allow for heterogeneity in *dN*/*dS* ratios across codons (*M8* and *M8a*; Swanson et al. 2003). *M8a* is a model of purifying selection that allows individual sites to evolve under differing levels of constraint (*dN*/*dS* 0–1), while *M8* allows for positive selection (*dN*/*dS* > 1) at an estimated proportion of sites. Positive selection was inferred if *M8* provided a significantly better fit to the data using a likelihood ratio test ($\alpha = 0.05$; $df = 1$). Second, we used branch-specific models that allow *dN*/*dS* ratios to vary across branches in the phylogeny.

Comparative *dN*/*dS* studies often use a single sequence to represent each species, assuming that the majority of nucleotide differences between any two species will reflect fixed differences rather than polymorphic variants still segregating within a population. However, this assumption is often violated between closely related species (Keightley and Eyre-Walker 2012). Moreover, there tends to be an excess of amino acid changing polymorphisms relative to divergence in species with relatively small effective population sizes, such as humans, due to the segregation of slightly deleterious variants (Fay et al. 2001; Bustamante et al. 2005; Keightley et al. 2005). To account for this, we replaced polymorphic positions within each of the four focal species with the parsimony-inferred ancestral state.

The overall species relationships among the great apes and humans are unambiguous. However, the internal branching structure of individual genealogies may differ due to incomplete lineage sorting. An unrooted tree grouping chimpanzee and bonobo was assumed for all *dN*/*dS* analyses based on a four species alignment. For five-taxon alignments (including the orangutan), we used CODEML to estimate the likelihood of alternative phylogenies that were probable under a model of incomplete lineage sorting. We restricted our focus to the three alternative trees with different relationships between *Gorilla*, *Homo*, and *Pan* and chose the tree with the highest likelihood for subsequent analyses. Other possible trees are relatively rare in the genome, especially in exons (Caswell et al. 2008; Prüfer et al. 2012; Scally et al. 2012).

Polymorphism and Divergence

We used the McDonald–Kreitman test (McDonald and Kreitman 1991) to contrast the ratio of polymorphism to divergence at amino acid changing (nonsynonymous) and silent (synonymous) positions. Under a neutral model of molecular evolution, the ratio of polymorphism to divergence should be equivalent between these two site classes. We also summarized global patterns of polymorphism to divergence using the NI (Rand and Kann 1998). NI is the ratio of polymorphism to divergence between nonsynonymous and synonymous sites derived from the 2×2

McDonald–Kreitman test contingency table. Assuming that silent positions reflect the neutral equilibrium condition, positive selection should result in an excess of amino acid divergence ($NI < 1$), while weak purifying selection should generate an excess of amino acid polymorphism ($NI > 1$). We report an unbiased estimator of the NI for all analyses (NI_{TG}; Stoletzki and Eyre-Walker 2011).

The McDonald–Kreitman test is most often applied to pairwise comparisons with polymorphism data collected from one of the two species, but it is valid for any partition of a neutral genealogy (McDonald and Kreitman 1991). We examined patterns of polymorphism and divergence for the entire phylogeny, pairwise species comparisons, and specific lineages. Polymorphism counts were generated with polydNdS (Thornton 2003). We extrapolated phylogeny-wide and branch-specific divergence from our branch-specific estimates of *dN* and *dS*.

Finally, we estimated the distribution of fitness effects (DFE) of new amino acid mutations and the proportion of adaptive substitutions (α) for each species using an ML approach (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009) as implemented through the DFE webserver (http://lanner.cap.ed.ac.uk/~eang33/dfe-alpha-server.html, last accessed January 29, 2013). For each species, we used the folded site frequency spectrum of 0-fold and 4-fold positions to estimate DFE and α under a simple demographic model allowing for a single-step change in ancestral population size. For these pairwise analyses on population samples of chimpanzee, bonobo, or human, we used the gorilla with the highest overall sequence coverage as a reference and calculated the number of fixed differences. A single high-coverage human sequence was used as a reference for gorilla populations. 95% confidence intervals for DFE and α were estimated using 120 bootstrap replicates by site.

Supplementary Material

Supplementary analyses, supplementary figures S1–S5, and supplementary tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Emily Hodges, Matthias Meyer, Adrian Briggs, Johannes Krause, Michael Knapp, and MPI-EVA sequencing group for technical advice. Miguel Carniero, Thomas Giger, Peter Keightley, and Adam Eyre-Walker provided useful advice on analyses. The authors also thank Olaf Thalmann and Linda Vigilant for access to gorilla DNA samples and for useful discussions. They are grateful to the Yerkes Regional Primate Research Center, the Lola ya Bonobo Sanctuary (Democratic Republic of Congo), the Tchimpounga Chimpanzee Rehabilitation Center (Republic of Congo), F. Lankester and the Limbe Wildlife Centre (Cameroon), the Le Ministère de l'Environnement et des Forêts du Cameroun, and the Ministry of Forest of the Republic of Congo for their support and for granting access to samples. At Lola ya Bonobo, the authors thank Dominique Morel, Fanny Mehl, and Pierrot Mbonzo for their support in collaboration with the Ministry of Research and the Ministry of Environment in the Democratic Republic of Congo for supporting their research (research permit: MIN.RS/SG/004/2009). This work was supported by the European Research Council (grant number 233297, TWOPAN to S.P.) and a National Science Foundation international postdoctoral fellowship (OISE-0754461 to J.M.G).

References

Aguade M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics* 152:543–551.

Albert TJ, Molla MN, Muzny DM, et al. (12 co-authors). 2007. Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 4:903–905.

Anderson MJ, Nyholt J, Dixson AF. 2005. Sperm competition and the evolution of sperm midpiece volume in mammals. *J Zool.* 267: 135–142.

Andersson M. 1994. Sexual selection. Princeton (NJ): Princeton University Press.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* 17:1755–1762.

Andres JA, Maroja LS, Bogdanowicz SM, Swanson WJ, Harrison RG. 2006. Molecular evolution of seminal proteins in field crickets. *Mol Biol Evol.* 23:1574–1584.

Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. *Genetics* 156:1879–1888.

Bi K, Vanderpool D, Singhai S, Linderoth T, Moritz C, Good JM. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics* 13:403.

Boyko AR, Williamson SH, Indap AR, et al. (14 co-authors). 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 4:e1000083.

Burbano HA, Hodges E, Green RE, et al. (20 co-authors). 2010. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science* 328:723–725.

Bustamante CD, Fledel-Alon A, Williamson S, et al. (14 co-authors). 2005. Natural selection on protein-coding genes in the human genome. *Nature* 437:1153–1157.

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416:531–534.

Carneiro M, Albert FW, Melo-Ferreira J, Galtier N, Gayral P, Blanco-Aguiar JA, Villafuerte R, Nachman MW, Ferrand N. 2012. Evidence for widespread positive and purifying selection across the european rabbit (*Oryctolagus cuniculus*) genome. *Mol Biol Evol.* 29:1837–1849.

Caswell JL, Mallick S, Richter DJ, Neubauer J, Schirmer C, Gnerre S, Reich D. 2008. Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genet.* 4:e1000057.

Chapman T. 2001. Seminal fluid-mediated fitness traits in *Drosophila*. *Heredity* 87:511–521.

Clark AG, Aguade M, Prout T, Harshman LG, Langley CH. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. *Genetics* 139:189–201.

Clark AG, Begun DJ. 1998. Female genotypes affect sperm displacement in *Drosophila*. *Genetics* 149:1487–1493.

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

Claydon AJ, Ramm SA, Pennington A, Hurst JL, Stockley P, Beynon R. 2012. Heterogenous turnover of sperm and seminal vesicle proteins in the mouse revealed by dynamic metabolic labeling. *Mol Cell Proteomics.* 11:M111.014993.

Clutton-Brock TH, Harvey PH, Rudder B. 1977. Sexual dimorphism, socionomic sex ratio and body weight in primates. *Nature* 269: 797–800.

Coulthart MB, Singh RS. 1988. High level of divergence of male reproductive tract proteins between *Drosophila melanogaster* and its sibling species, *Drosophila simulans*. *Mol Biol Evol.* 5:182–191.
Dean MD, Clark NL, Findlay GD, Karn RC, Yi X, Swanson WJ, MacCoss MJ, Nachman MW. 2009. Proteomics and comparative genomic investigations reveal heterogeneity in evolutionary rate of male reproductive proteins in mice (*Mus domesticus*). *Mol Biol Evol.* 26: 1733–1743.
Dean MD, Good JM, Nachman MW. 2008. Adaptive evolution of proteins secreted during sperm maturation: an analysis of the mouse epididymal transcriptome. *Mol Biol Evol.* 25:383–392.
Dixson AF. 1987. Observations on the evolution of the genitalia and copulatory behavior in male primates. *J Zool.* 213:423–443.
Dixson AF. 1998. Primate sexuality. New York: Oxford University Press.
Dixson AF, Anderson MJ. 2002. Sexual selection, seminal coagulation and copulatory plug formation in primates. *Folia Primatol.* 73:63–69.
Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. 2006. Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nat Genet.* 38:1440–1445.
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.
Dorus S, Wasbrough ER, Busby J, Wilkin EC, Karr TL. 2010. Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Mol Biol Evol.* 27:1235–1246.
Eberhard WG. 1996. Female control: sexual selection by cryptic female choice. Princeton (NJ): Princeton University Press.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.
Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat Protoc.* 2:953–971.
Eriksson J, Hohmann G, Boesch C, Vigilant L. 2004. Rivers influence the population genetic structure of bonobos (*Pan paniscus*). *Mol Ecol.* 13:3425–3435.
Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends Ecol Evol.* 21:569–575.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol.* 26:2097–2108.
Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.
Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.
Fischer A, Pollack J, Thalmann O, Nickel B, Pääbo S. 2006. Demographic history and genetic differentiation in apes. *Curr Biol.* 16:1133–1138.
Fischer A, Prüfer K, Good JM, Halbwax M, Wiebe V, André C, Atencia R, Mugisha L, Ptak SE, Pääbo S. 2011. Bonobos fall within the genomic variation of chimpanzees. *PLoS One* 6:e21605.
George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH. 2011. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res.* 21:1686–1694.
Gibbs RA, Weinstock GM, Metzker ML, et al. (230 co-authors). 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521.
Good JM, Nachman MW. 2005. Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. *Mol Biol Evol.* 22:1044–1052.
Halligan D, Oliver F, Eyre-Walker A, Harr B, Keightley P. 2010. Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet.* 6: e1000825.
Hamm D, Mautz BS, Wolfner MF, Aquadro CF, Swanson WJ. 2007. Evidence of amino acid diversity-enhancing selection within humans and among primates at the candidate sperm-receptor gene *PKDREJ*. *Am J Hum Genet.* 81:44–52.
Harcourt AH, Harvey PH, Larson SG, Short RV. 1981. Testis weight, body weight and breeding system in primates. *Nature* 293:55–57.
Harcourt AH, Purvis A, Liles L. 1995. Sperm competition: mating system, not breeding season, affects testes size of primates. *Funct Ecol.* 9: 468–476.
Henault MA, Killian GJ. 1996. Effect of homologous and heterologous seminal plasma on the fertilizing ability of ejaculated bull spermatozoa assessed by penetration of zona-free bovine oocytes. *J Reprod Fertil.* 108:199–204.
Herlyn H, Zischler H. 2007. Sequence evolution of the sperm ligand zonadhesin correlates negatively with body weight dimorphism in primates. *Evolution* 61:289–298.
Hodges E, Rooks M, Xuan ZY, Bhattacharjee A, Gordon DB, Brizuela L, McCombie WR, Hannon GJ. 2009. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc.* 4:960–974.
Holland B, Rice WR. 1999. Experimental removal of sexual selection reverses intersexual antagonistic coevolution and removes a reproductive load. *Proc Natl Acad Sci U S A.* 96:5083–5088.
Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T. 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proc Natl Acad Sci U S A.* 109:2054–2059.
Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *J Mol Evol.* 57: 261–270.
Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251–2261.
Keightley PD, Eyre-Walker A. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J Mol Evol.* 74:61–68.
Keightley PD, Lercher MJ, Eyre-Walker A. 2005. Evidence for widespread degradation of gene control regions in hominid genomes. *PLoS Biol.* 3:e42 41–47.
Kingan SB, Tatar M, Rand DM. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. *J Mol Evol.* 57:159–169.
Kircher M, Stenzel U, Kelso J. 2009. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.* 10:R83.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (Genome Project Data P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.
Li YF, Costello JC, Holloway AK, Hahn MW. 2008. "Reverse ecology" and the power of population genomics. *Evolution* 62:2984–2994.
Locke DP, Hillier LW, Warren WC, et al. (101 co-authors). 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* 469:529–533.
Markow TA. 2002. Perspective: female remating, operational sex ratio, and the arena of sexual selection in *Drosophila* species. *Evolution* 56: 1725–1734.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654.
Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc.* 2010:pdb.prot5448.
Mueller JL, Ram KR, McGraw LA, Qazi MCB, Siggia ED, Clark AG, Aquadro CF, Wolfner MF. 2005. Cross-species comparison of *Drosophila* male accessory gland protein genes. *Genetics* 171: 131–143.
Mueller JL, Ripoll DR, Aquadro CF, Wolfner MF. 2004. Comparative structural modeling and inference of conserved protein classes in *Drosophila* seminal fluid. *Proc Natl Acad Sci U S A.* 101:13542–13547.
Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156:297–304.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A.* 76: 5269–5273.
Nunn CL, Gittleman JL, Antonovics J. 2000. Promiscuity and the primate immune system. *Science* 290:1168–1170.
O'Connor TD, Mundy NI. 2009. Genotype–phenotype associations: substitution models to detect evolutionary associations between phenotypic variables and genotypic evolutionary rate. *Bioinformatics* 25: i94–i100.
Parker GA. 1970. Sperm competition and its evolutionary consequences in the insects. *Biol Rev.* 45:525–567.
Phifer-Rixey M, Bonhomme F, Boursot P, Churchill GA, Piálek J, Tucker PK, Nachman MW. 2012. Adaptive evolution and effective population size in wild house mice. *Mol Biol Evol.* 9:2949–2955.
Pilch B, Mann M. 2006. Large-scale and high-confidence proteomic analysis of human seminal plasma. *Genome Biol.* 7:R40.
Plagnol V, Wall JD. 2006. Possible ancestral structure in human populations. *PLoS Genet.* 2: e105.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.
Prüfer K, Munch K, Hellmann I, et al. (41 co-authors). 2012. The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527–531.
Ramm SA, McDonald L, Hurst JL, Beynon RJ, Stockley P. 2009. Comparative proteomics reveals evidence for evolutionary diversification of rodent seminal fluid and its functional significance in sperm competition. *Mol Biol Evol.* 26:189–198.
Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. *Mol Biol Evol.* 25:207–219.
Rand DM, Kann LM. 1998. Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. *Genetica* 103:393–407.
Rice WR. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* 381:232–234.
Robertson SA. 2005. Seminal plasma and male factor signalling in the female reproductive tract. *Cell Tissue Res.* 322:43–52.
Robertson SA. 2007. Seminal fluid signaling in the female reproductive tract: lessons from rodents and pigs. *J Anim Sci.* 85:E36–E44.
Scally A, Dutheil JY, Hillier LW, et al. (71 co-authors). 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483: 169–175.
Shivaji S, Scheit K-H, Bhargava PM. 1990. Proteins of seminal plasma. New York: John Wiley & Sons.
Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.
Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol.* 28:63–70.
Swanson WJ, Nielsen R, Yang QF. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol.* 20:18–20.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
Thalmann O, Fischer AH, Lankester FH, Pääbo SH, Vigilant LH. 2007. The complex evolutionary history of gorillas: insights from genomic data. *Mol Biol Evol.* 24:146–158.
Thalmann O, Wegmann D, Spitzner M, Arandjelovic M, Guschanski K, Leuenberger C, Bergl RA, Vigilant L. 2011. Historical sampling reveals dramatic demographic changes in western gorilla populations. *BMC Evol Biol.* 11:85.
Thornton K. 2003. libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.
Torgerson DG, Kulathinal RJ, Singh RS. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. *Mol Biol Evol.* 19:1973–1980.
Torgerson DG, Whitty BR, Singh RS. 2005. Sex-specific functional specialization and the evolutionary rates of essential fertility genes. *J Mol Evol.* 61:650–658.
Tsaur SC, Ting CT, Wu CI. 1998. Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of *Drosophila*: II. Divergence versus polymorphism. *Mol Biol Evol.* 15:1040–1046.
Utleg AG, Yi EC, Xie T, Shannon P, White JT, Goodlett DR, Hood L, Lin BY. 2003. Proteomic analysis of human prostasomes. *Prostate* 56: 150–161.
Vacquier VD, Swanson WJ, Lee YH. 1997. Positive Darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J Mol Evol.* 44:S15–S22.
Wall JD, Cox MP, Mendez FL, Woerner A, Severson T, Hammer MF. 2008. A novel DNA sequence database for analyzing human demographic history. *Genome Res.* 18:1354–1361.
Watterson GA. 1975. Number of segregating sites in genetic models without recombination. *Theor Popul Biol.* 7:256–276.
Wolfner MF. 1997. Tokens of love: functions and regulation of *Drosophila* male accessory gland products. *Insect Biochem Mol Biol.* 27:179–192.
Wong A. 2010. Testing the effects of mating system variation on rate of molecular evolution in primates. *Evolution* 64:2779–2785.
Wong A. 2011. The molecular evolution of animal reproductive tract proteins: what have we learned from mating-system comparisons? *Int J Evol Biol.* 2011:1–9.
Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309.
Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol.* 17:32–43.
Yang Z, Swanson WJ. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol.* 19:49–57.
Yang ZH. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Zhang LQ, Li WH. 2005. Human SNPs reveal no evidence of frequent positive selection. *Mol Biol Evol.* 22:2504–2507.
Zsurka G, Kudina T, Peeva V, Hallmann K, Elger CE, Khrapko K, Kunz WS. 2010. Distinct patterns of mitochondrial genome diversity in bonobos (*Pan paniscus*) and humans. *BMC Evol Biol.* 10:270.