Reproduction and Immunity-Driven Natural Selection in the Human *WFDC* Locus

Zélia Ferreira,*[,1,2,3] Susana Seixas,[2] Aida M. Andrés,[4] Warren W. Kretzschmar,[5] James C. Mullikin,[6] Praveen F. Cherukuri,[6] Pedro Cruz,[6] Willie J. Swanson,[7] NISC Comparative Sequencing Program,[1,6] Andrew G. Clark,[8] Eric D. Green,[1] and Belen Hurle[1]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

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

[3]Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal

[4]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

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

[7]Department of Genome Sciences, University of Washington

[8]Department of Molecular Biology and Genetics, Cornell University

***Corresponding author:** E-mail: zferreira@ipatimup.pt.

Associate editor: Ryan Hernandez

Abstract

The whey acidic protein (WAP) four-disulfide core domain (*WFDC*) locus located on human chromosome 20q13 spans 19 genes with WAP and/or Kunitz domains. These genes participate in antimicrobial, immune, and tissue homoeostasis activities. Neighboring *SEMG* genes encode seminal proteins Semenogelin 1 and 2 (SEMG1 and SEMG2). *WFDC* and *SEMG* genes have a strikingly high rate of amino acid replacement (d_N/d_S), indicative of responses to adaptive pressures during vertebrate evolution. To better understand the selection pressures acting on *WFDC* genes in human populations, we resequenced 18 genes and 54 noncoding segments in 71 European (CEU), African (YRI), and Asian (CHB + JPT) individuals. Overall, we identified 484 single-nucleotide polymorphisms (SNPs), including 65 coding variants (of which 49 are nonsynonymous differences). Using classic neutrality tests, we confirmed the signature of short-term balancing selection on *WFDC8* in Europeans and a signature of positive selection spanning genes *PI3*, *SEMG1*, *SEMG2*, and *SLPI*. Associated with the latter signal, we identified an unusually homogeneous-derived 100-kb haplotype with a frequency of 88% in Asian populations. A putative candidate variant targeted by selection is Thr56Ser in *SEMG1*, which may alter the proteolytic profile of SEMG1 and antimicrobial activities of semen. All the well-characterized genes residing in the *WDFC* locus encode proteins that appear to have a role in immunity and/or fertility, two processes that are often associated with adaptive evolution. This study provides further evidence that the *WFDC* and *SEMG* loci have been under strong adaptive pressure within the short timescale of modern humans.

***Key words:* WFDC, semenogelins, natural selection, innate immunity, serine protease inhibitors, reproduction.**

Article

Introduction

The whey acidic protein (WAP) four-disulfide core domain (*WFDC*) gene locus on human chromosome 20q13 spans 19 genes with WAP and/or Kunitz domains that confer serine protease inhibitor activity (Clauss et al. 2005, 2011; Lundwall 2007; Lundwall and Clauss 2011). *WFDC* genes exhibit core functions involving reproduction, antimicrobial, immune, and tissue homoeostasis activities that in most cases remain poorly understood (Yenugu et al. 2004; Bouchard et al. 2006; Bingle and Vyakarnam 2008; Lundwall and Clauss 2011). The *WFDC* locus includes genes encoding the seminal proteins Semenogelin 1 and 2 (SEMG1 and -2) (Peter et al. 1998; de Lamirande 2007; Lundwall 2007). The *WFDC* and *SEMG* genes stand out for reports of striking signatures of adaptive evolution, reflecting effects of natural selection during mammalian evolution (Dorus et al. 2004; Hurle et al. 2007).

Most evolutionary and functional studies on the *WFDC* gene family have focused on genes located within the centromeric sublocus of the large gene cluster (fig. 1*A*). This small but dynamic genome region (hereafter referred to as *WFDC*-CEN) has a notably complex evolutionary history resulting in rapid interspecies divergence of both coding and noncoding sequences (Hurle et al. 2007). Proteins encoded by the genes in *WFDC*-CEN include the well-studied peptidase inhibitor 3 (PI3, also known as elafin) and secretory leucocyte proteinase inhibitor (SLPI), which are pleiotropic molecules engaged in the surveillance against microbial infections at mucosal surfaces (Williams et al. 2006; Weldon and Taggart 2007; Weldon et al. 2007; McKiernan et al. 2011). Also well characterized are the *SEMG1* and *SEMG2* genes encoding seminal plasma proteins with roles in semen clotting and in antimicrobial protection for the spermatozoa in the female reproductive tract

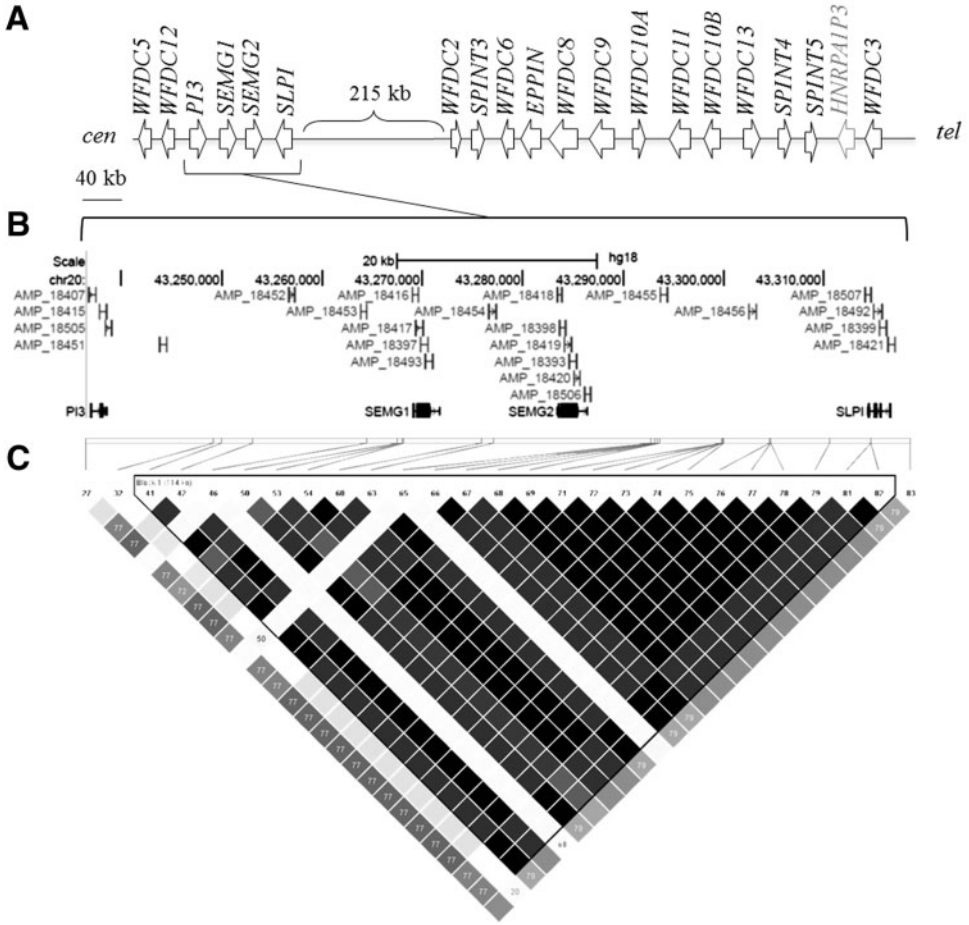

Fig. 1. Schematic representation of the 20q13 *WFDC* gene cluster. (*A*) Diagram showing the relative positions of the *WFDC* genes. As depicted, the *WFDC* cluster spans 700 kb and its genes are organized into two subloci (centromeric and telomeric; *WFDC*-CEN and *WFDC*-TEL, respectively), separated by 215 kb of unrelated sequence. *HNRPA1P3* pseudogene is indicated in light gray. (*B*) Strategy for the resequencing effort across the *WFDC* locus. One hundred thirty 700-bp-long amplicons were designed to include all exonic regions and a selection of noncoding sequences evenly spaced every 10 kb across the two *WFDC* subloci. (*C*) Linkage disequilibrium in the *PI3-SEMG1-SEMG2-SLPI* region in Asians (calculated using the resequenced data and displayed with Haploview; Haplotype blocks defined by Gabriel et al. 2002).

(Lundwall et al. 2002; Bourgeon et al. 2004; Edstrom et al. 2008; Martellini et al. 2009).

Comparative genomics and phylogenetic analysis indicate that *SLPI*, *PI3*, *SEMG1*, and *SEMG2* have evolved rapidly since the separation of the primate and murine lineages (Hurle et al. 2007). In particular, multiple studies show their accelerated molecular evolution as measured by their high d_N/d_S values (Dorus et al. 2004; Hurle et al. 2007; Ramm et al. 2008).

The *WFDC* telomeric sublocus (hereafter referred to as *WFDC*-TEL) is physically separated from *WFDC*-CEN by 215 kb of unrelated genomic sequence. The best-characterized gene within *WFDC*-TEL encodes the epididymal protease inhibitor (EPPIN, also known as *SPINLW1*). At ejaculation, EPPIN coats the surface of human spermatozoa, binds to SEMG1, and helps to modulate the activity of prostate-specific antigen (PSA) while providing antimicrobial protection for spermatozoa (Robert and Gagnon 1999; Bourgeon et al. 2004; Wang et al. 2005; Edstrom et al. 2008; Zhao et al. 2008). The functions of most other *WFDC* genes remain poorly characterized.

Surprisingly, despite the strong signatures of positive selection revealed by excess nonsynonymous (NS) divergence among species, few studies have used intraspecific polymorphism data to examine the selective pressures acting on *WFDC* and *SEMG* genes within populations. Most of these focused on *WFDC*-CEN, where the *SEMGs* have been identified as genes under adaptive evolution, specifically, by correlating their single-nucleotide polymorphisms (SNPs) and copy-number variants to the different mating systems of various primate species (Jensen-Seaman and Li 2003; Kingan et al. 2003; Dorus et al. 2004; Carnahan and Jensen-Seaman 2008).

The only study examining the selective pressures occurring in *WFDC*-TEL focused on *WFDC8*, a proposed target of recent balancing selection in Europeans based on the intermediate frequency of the haplotypes containing the candidate variant [-44(G/A)] and *SPINT4*, which was linked to a rapid increase in frequency of the advantageous allele (Ser73), associated with a long-range haplotype (LRH) and low-frequency variants—thus appearing to evolve under an incomplete selective sweep in Africans (Ferreira et al. 2011).

To gain a better understanding of the more recent selective pressures shaping genetic variation in the human *WFDC* locus, we systematically resequenced 18 genes of the locus plus 54 evenly spaced noncoding segments in 71 humans

from European (CEU), African (YRI), and Asian (CHB + JPT) HapMap populations. A set of 47 autosomal, unlinked, and neutrally evolving loci were also surveyed to assess baseline (neutral) genomic diversity. Using classic neutrality tests (Tajima's *D* and Fay and Wu's *H*), we confirmed the signature of short-term balancing selection on *WFDC8* in the CEU population; and we further pinpointed a signature of positive selection spanning *PI3*, *SEMG1*, *SEMG2*, and *SLPI*. The best candidate variant for the latter selective footprint in Asians was allele Ser56 in SEMG1. This variant potentially modifies the likelihood of PSA-mediated hydrolysis of SEMG1, simultaneously altering the peptide profile and antimicrobial activities of semen.

This study is the first to provide systematic and comprehensive population genomics-based evidence that a number of *WFDC* and *SEMG* genes are under strong adaptive pressures within the recent timescale of modern humans.

Results

To gain a better understanding of the selective pressures shaping the genetic variation within *WFDC* genes, we designed 130 (~700 bp) amplicons across the *WFDC* locus. These amplicons were amplified from a panel of 71 HapMap Phase I/II individuals (21 CEU, 25 YRI, and 25 CHB + JPT) and Sanger sequenced (supplementary tables S1 and S2, Supplementary Material online). In this study, a total of 8.1 Mb of targeted genomic regions were sequenced, 20% of which corresponds to exonic regions and the rest accounts for intronic and putative *cis*-regulatory regions (52%) and intergenic regions (28%) (supplementary table S3, Supplementary Material online).

Genetic Variation in WFDC Genes

Overall, 484 SNPs were identified, of which 65 resided in coding regions. Forty-nine of the coding SNPs were NS, of which 67% were present at very low frequencies in all populations ($f \leq 0.08$) (fig. 2; supplementary table S3a, Supplementary Material online). Such a pattern of allele frequencies is consistent with mildly deleterious effects of most NS variants, although it does not depart from a strictly neutral site frequency spectrum (SFS; 1,000 coalescent simulations; $S = 49$; χ^2 test; $P = 0.47$). Seven NS-SNPs were predicted to affect protein function by SIFT and PolyPhen v2 where only rs6017667 (Gly73Ser in SPINT4) occurs at an intermediate frequency $f = 0.44$.

Twenty-four insertions/deletions (indels) were found, 21 of which were located in intronic and intergenic regions. The three remaining indels were in untranslated coding regions of *WFDC9* and *WFDC13*. Because indels might have a distinct mutation rate compared with SNPs and their genomic localization does not seem to affect protein function or expression, they were excluded from the following analyses. Additionally, we found 456 fixed human–chimpanzee differences, of which only 19 were within coding regions and human specific. The Polyphen v2 and SIFT analysis show that the functional impact of most of the NS fixed differences was classified as benign (supplementary table S3b, Supplementary Material online).

Deviations of Allele Frequency Spectra from Neutral Expectations

Figure 2 depicts the distribution of folded SFS for all the surveyed genes. In *WFDC*-CEN, there is a skew toward low-frequency variants (fig. 2*A* and *B*), whereas in *WFDC*-TEL, there is a shift toward intermediate-frequency variants (fig. 2*C* and *D*), following the trend observed for coding SNPs. The significance of deviations from neutral expectations of the SFS within each population was tested using the summary statistics π, θ_w, Tajima's *D*, and Fay and Wu's *H* (Tajima 1989; Fu 1996; Fay and Wu 2000; Zeng et al. 2006). We controlled for demography effects by using the demographic model developed by Gutenkunst et al. (2009) and determined nominal *P* values for each statistic (table 1 and supplementary table S4, Supplementary Material online). Individual genes in the *WFDC* locus present summary statistics that have moderate *P* values and whose significance is marginal after multiple test correction (Benjamini and Hochberg 1995; Storey 2002; Storey and Tibshirani 2003; Storey et al. 2004). However, the nominal *P* values clearly show that there is a trend toward lower nominal *P* values for Tajima's *D*, Fu and Li's *D*, Fay and Wu's *H*, and Mann-Whitney U $(MWU)_{high}$ (supplementary fig. S2, Supplementary Material online), which suggested the need for further testing.

The tail probability of test statistics of the *WDFC* region was assessed by using simulations based on fits of demographic models to the neutral regions. One evaluation of the validity of this approach is to determine the corresponding tail probabilities for the control regions in the study. We calculated the levels of nucleotide diversity (π) and Tajima's *D* for the 47 control regions in each population and created an empirical distribution of the obtained values. Because the control regions have been subject to the same demographic history as the *WFDC* locus, an outlier value (2.5 or 97.5 percentile) would suggest a non-neutral evolution event (supplementary fig. S3, Supplementary Material online).

At the population level, the lowest π levels were found mainly in the Asian population, followed by the CEU and YRI populations (supplementary table S4, Supplementary Material online), as expected under the out-of-Africa model for human populations (Schaffner et al. 2005; Voight et al. 2005; Gutenkunst et al. 2009). At the gene level, the genes that display the most unusual π values (supplementary fig. S3*A* and table S4, Supplementary Material online) are *SEMG1* and *SEMG2*, with low nucleotide diversity values ($\pi_{SEMG1} = 0.761063 \times 10^{-4}$; $\pi_{SEMG2} = 0.933816 \times 10^{-4}$) in the Asian population, and *WFDC3*, with high nucleotide diversity in Europeans and Africans ($\pi_{WFDC3} = 11.473 \times 10^{-4}$ and $\pi_{WFDC3} = 14.0656 \times 10^{-4}$ for each population, respectively). The generated empirical distribution of Tajima's *D* values compared with each gene suggests that *PI3* and *SEMG2* are outliers in the Asian population (supplementary fig. S3*B*, Supplementary Material online). The overall levels of diversity

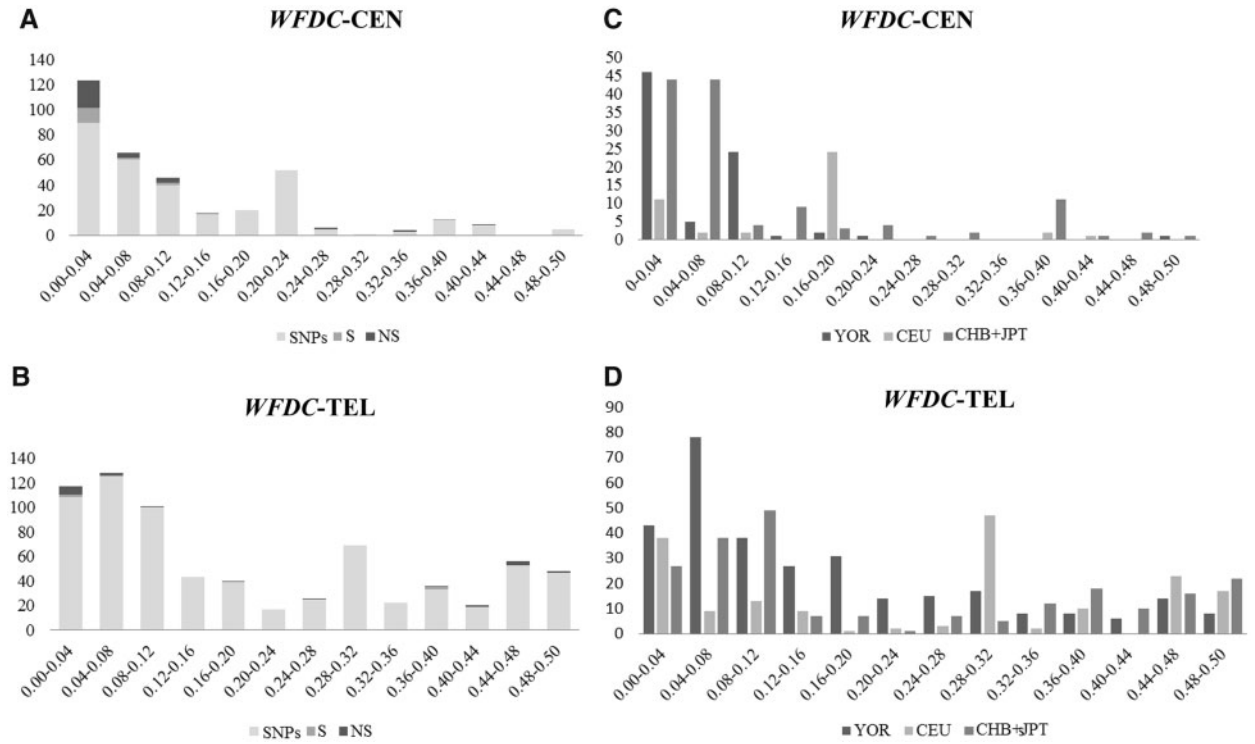

Fig. 2. Folded site frequency spectrum (folded SFS) for the *WFDC* in all populations resequenced. The *x* axis depicts the frequency of the allele frequency bin in the generated data set, whereas the *y* axis represents the number of alleles found within each frequency bin. S, synonymous changes; NS, nonsynonymous changes. (*A*) and (*B*), folded SFS in *WFDC*-CEN; (*C*) and (*D*), folded SFS in *WFDC*-TEL.

Table 1. Significant Summary Statistics at the *WFDC* Locus.

Gene	Population	S[a]	θ_w[b]	π[c]	Tajima's *D*[d]	*H*[e]
PI3	YRI	34	7.60101	4.69663	−1.28656	−1.70946**
	CHB + JPT	17	3.81646	1.67945	−1.75823*	−3.79418**
SEMG2	YRI	18	4.02479	2.81504	−0.947299	−1.68143**
	CHB + JPT	11	2.46906	0.933816	−1.82029*	−3.54599**
SLPI	CEU	10	2.56518	1.86032	−0.884379	−2.60125*
	CHB + JPT	16	3.59903	2.76312	−0.724534	−2.85197*
SPINT3	YRI	11	2.45787	1.32809	−1.33954	−1.60077**
WFDC8	CEU	27	6.88872	10.7402	2.02506*	−0.250797
	CHB + JPT	31	7.01214	5.01087	−0.964878	−3.7064**

[a]S, number of segregating sites.
[b]Watterson's estimator of θ ($4N_e\mu$) (Watterson 1975) per base pair ($\times 10^{-4}$).
[c]Nucleotide diversity per base pair ($\times 10^{-4}$).
[d]Tajima's *D* statistic (Tajima 1989).
[e]Fay and Wu's *H* test (Fay and Wu 2000; Zeng et al. 2006).
*$P < 0.05$ and **$P < 0.025$.

in the *WFDC* locus suggest a non-neutral evolution of these genes.

Footprints of Recent Positive Selection in Asians

Summary statistics suggest that the *PI3-SEMG1-SEMG2-SLPI* region at *WFDC*-CEN has evolved under non-neutral evolution in the Asian population (table 1 and supplementary table S4, Supplementary Material online). Considering the physical clustering of these genes (fig. 1), their low levels of intrapopulation nucleotide diversity, and outlier Tajima's *D* values both in the empirical and simulated comparisons, we looked for possible signatures of positive selection in this region.

Considering each gene individually, these have borderline significant Tajima's *D* and Fay and Wu's *H* *P* values (table 1). Additionally, a number of SNPs in *SEMG1* and *SLPI* presented elevated F_{ST} values, with *P* values ranging from 0.01 to 0.05 (supplementary fig. S1*A*–*F*, Supplementary Material online). Coincidentally, the *PI3-SEMG1-SEMG2-SLPI* (fig. 1*A*) gene array forms a single linkage disequilibrium (LD) block ($D' = 1$; $r^2 = 0.8$) of approximately 100 kb (fig. 1*B* and *C*) in the Asian population. This LD block is longer than the average Asian LD extension of approximately 44 kb and longer than the flanking regions (Gabriel et al. 2002).

To determine whether the low levels of diversity and elevated haplotype extension could be due to natural selection,

we used a sliding window of Tajima's *D* with the aim to identify peaks of aberrant Tajima's *D* values across this region (fig. 3). Two peaks of very low Tajima's *D* were identified: one between *PI3* and *SEMG1* (−2.17) and another one in *SEMG2* (−1.86). Downstream of *SEMG2*, the levels of diversity variation recover rapidly, and *SLPI* no longer has low levels of diversity, with higher Tajima's *D*, π, and haplotype diversity values than its neighboring genes. These results suggest that *SLPI* is not a gene under selective pressure.

Taking into account the strong LD among these genes, we sought for signals of recent selection using the following haplotype tests: Hudson's haplotype test (Hudson et al. 1994), derived intra-allelic nucleotide diversity (DIND; Barreiro et al. 2009), and extended haplotype homozygosity (EHH) and relative EHH (REHH) (Sabeti et al. 2002). We started by applying the Hudson's haplotype test because the LD block shown in figure 1*C* could be attributed to a single haplotype with a 100-kb extension present in 88% of the in Asian population. With such test, we examined whether the common haplotype contained fewer segregating sites than expected under neutrality given its frequency. We obtained significant results for the 100 kb region, encompassing 8 variable positions out of 44 segregating sites. The tests were based on 10,000 simulations (ms, Hudson 2002) of the constant neutral model ($P = 0.0023$) and on the Gutenkunst model for Asian populations (without migration) ($P = 0.0205$) (Gutenkunst et al. 2009).

In spite of the low nucleotide diversity associated with the extended haplotype, the DIND test showed that none of the *PI3*, *SEMG1*, or *SEMG2* variants was significant under the Gutenkunst demographic model. Nonetheless, a number of variants (rs13042431, rs2267864, rs2301366, and rs2071651; fig. 4) located in the region of interest display borderline nonsignificance under the Gutenkunst model and present significant values under the constant model of demography. From the latter SNPs, only one is a NS variant (rs2301366 A→T), reflecting an amino acid change of Thr56Ser in SEMG1 (fig. 4).

To further evaluate whether this haplotype structure could result from the action of positive selection, we calculated the EHH statistic proposed by Sabeti et al. (2002). We started by centering our analysis in the only NS variant that presented borderline *P* values from the DIND test, as a possible candidate variant of selection. Specifically, we measured the decay of LD around a three-SNP core haplotype centered in Thr56Ser. The bifurcation plot associated with Ser56 shows a frequent haplotype that extends for more than 60 kb in both directions from Thr56Ser (fig. 5*A*). We determined whether the EHH for the Ser56 core haplotype in the Asian sample was unusual by comparing its frequency and REHH at the largest distance where non−T/A haplotypes had nonzero values of EHH (80 kb distal and proximal) against null distributions. The deviations from simulated null distributions were significant for the haplotype associated with Ser56 in Asians (fig. 5*B*) but not in the other populations (results not shown).

A better understanding of the evolutionary history of *SEMG1* was obtained by the analysis of haplotype genealogies.

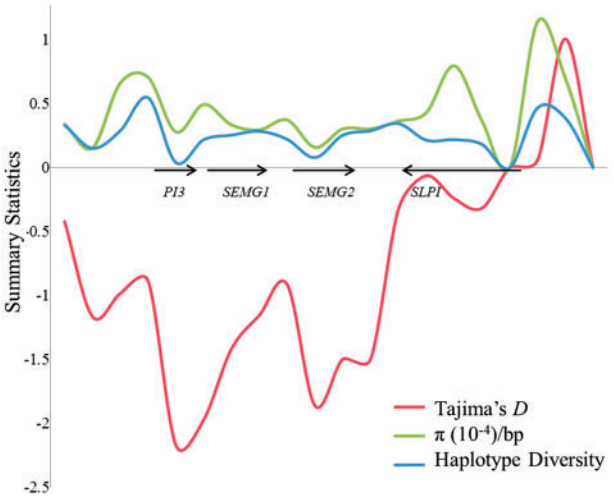

Fig. 3. Sliding window of Tajima's *D*, π, and Haplotype diversity (red, green, and blue lines, respectively) in the *PI3-SEMG1-SEMG2-SLPI* region in Asians. *PI3* and *SEMG1* region shows lower values than the rest of *WFDC*-CEN. Window size,1,000 bp; increment, 500 bp.

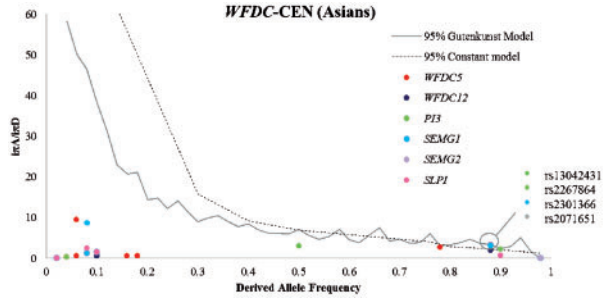

Fig. 4. Ratio of the ancestral ($i\pi_A$) alleles to the haplotypes carrying the derived ($i\pi_D$) alleles above expected, plotted as a function of Derived allele frequency. $P < 0.05$; Dashed line—5% constant model, recombination at 0.2 cM. Solid line—5% Gutenkunst model (Gutenkunst et al. 2009) in *WFDC*-CEN for Asians.

To estimate the divergence time of *SEMG1* and the age of Thr56Ser, we used a maximum likelihood coalescent analysis by GENETREE (Griffiths and Tavare 1994) and estimated the time to most recent common ancestor (T_{MRCA}). Because we suspected a selective force acting on Ser56, we estimated the β parameter and used it to determine gene trees under selection (Coop and Griffiths 2004). Using all three populations, the estimated T_{MRCA} for the entire *SEMG1* genealogy was 0.675 ± 0.103 My and for the Ser56 variant was 0.287 ± 0.05 My ($\theta_{ML} = 6.13$; $\beta = 1.70$). We then reconstructed the haplotype phylogenetic network of *SEMG1* using a median-joining algorithm (fig. 6*A*). Specifically, most Asian haplotypes cluster around a haplotype defined by Thr56Ser (rs2301366). The derived allele (T) is shared among all the descendent haplotypes, showing a star-shaped haplotype network, which is usually associated with a selective sweep or population expansion.

The haplotype tests and network phylogenetic structure suggest a non-neutral evolution of *SEMG1* (combined

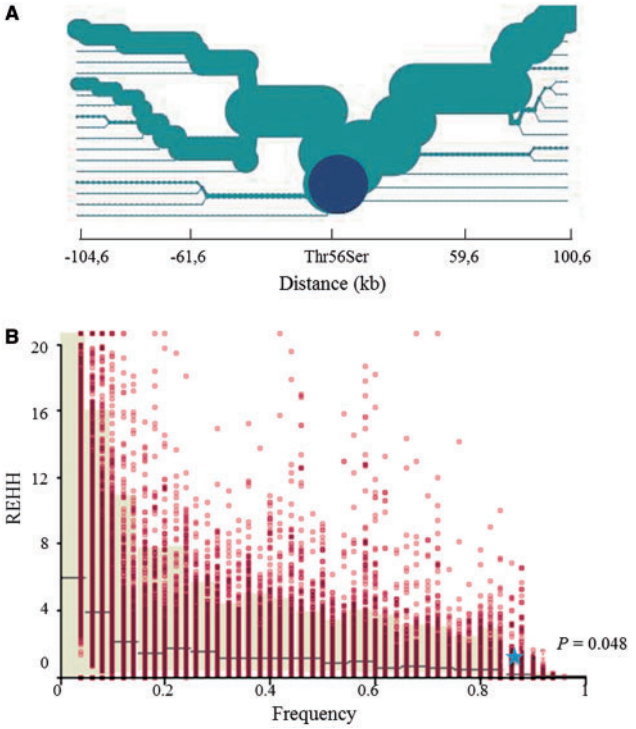

Fig. 5. (*A*) Haplotype bifurcation plot centered in position Thr56Ser of SEMG1 in Asian populations, using SWEEP. Thr56Ser is marked with a dark circle. The diameter of the circle and arm length is proportional to the number of individuals with the same LRH. Each of the additional SNPs is represented by a node from which bifurcation indicates a recombination event. (*B*) Relative expected haplotype homozygosity (REHH) deviations from simulated null distributions in the Asian population, using SWEEP software (www.broadinstitute.org/mpg/sweep, last accessed January 14, 2013). Highlighted point (star, P = 0.048) is Thr56Ser.

z-weighted *P* values = 0.002) (Whitlock 2005). We hypothesize that Ser56 is likely be under the influence of a selective sweep representing an advantageous allele that was swept to higher frequency in the Asian population (88%), simultaneously lowering the overall levels of nucleotide diversity and increasing the haplotype homozygosity in 160 kb of the surrounding regions (*PI3-SEMG1-SEMG2-SLPI*). However, because of the less recent age of the candidate variant and its presence in the other sequenced populations at somewhat elevated frequencies (Europeans 80% and Africans 38%), one cannot rule out the possibility of an event of selection on standing variation of Ser56 (Przeworski et al. 2005; Pritchard et al. 2010; Hernandez et al. 2011).

We set to compare our summary statistics (based on Sanger validated data from the *WFDC* locus) with summary statistics generated from data obtained from human genetic variation in public reference projects. The goal was to evaluate whether the different sequencing methods used had an effect in detecting genomic outliers, potentially affected by non-neutral evolution. Specifically, we performed a principal component and SFS analyses for the 1000 Genomes Project (supplementary figs. S4 and S6*B* and *D*, Supplementary Material online; Patterson, Price, et al. 2006) and generated SFS for the Complete Genomics Diversity Panel (Complete Genomics Assembly v1.3; Drmanac et al. 2010). However, given the substantial differences in SNP distribution in the latter data set due to the low sample size per population (supplementary fig. S6*A* and *C*, Supplementary Material online), we chose in the analysis that follows to directly compare the variants found in the 1000 Genomes Project and our sequencing survey, restricting our attention only to the sample that was sequenced in both projects.

For the *WFDC* Locus, our Sanger-based sequencing strategy detected 80% of the SNPs gathered by the 1000 Genomes Project. Conversely, the 1000 Genomes data contains 75% of the SNPs present in the data set generated for this study. The bulk of the discrepancies lie in low-frequency variants for which the 1000 Genomes data set presents lower singleton, doubleton, and tripleton frequencies (supplementary fig. S7, Supplementary Material online). These findings show that the publicly available genomes are very useful to detect genomic outliers, even though they do not yet completely replace deeper coverage and high-quality sequencing data. Despite the differences between SFS for the *WFDC* locus, the summary statistics present very similar and relatable values (supplementary tables S5 and S6, Supplementary Material online) suggesting that both approaches lead to the same results in this region of the genome. Specifically, for *SEMG1* in the Asian population, the summary statistic values ($\pi_{SEMG1} = 0.805 \times 10^{-4}$; Tajima's $D = -2.07$; and Fu and Li's $D = -3.9752$) are consistent with a non-neutral evolution of this gene.

Footprint of Short-Term Balancing Selection in Europeans

A previous study indicated that *WFDC8* is under short-term balancing selection in the CEU population (Ferreira et al. 2011). Sequencing the entire *WFDC* locus in three HapMap populations provided an opportunity to test in a larger data set the selective signal centered on *WFDC8*. The resulting sequence data confirmed that *WFDC8* has a positive Tajima's *D* (2.02) and elevated π values (10.7×10^{-4}) in the CEU population (table 1). The folded SFS for *WFDC8* shows an excess of polymorphic sites with intermediate frequency (fig. 2*C* and *D* and supplementary fig. S6, Supplementary Material online), which is significant in the CEU population based on MWU_{high} test ($P = 0.0089$) (Nielsen et al. 2009; Andrés et al. 2010) (supplementary table S4, Supplementary Material online). The haplotype network of *WFDC8* is structured into two highly differentiated haplotypes: "Haplotype A" and "Haplotype B" both several mutations away from the ancestral state (fig. 6*B*). Furthermore, the analysis of the 1000 Genomes data set confirms the elevated Tajima's *D* value (2.11) of *WFDC8* in the CEU population (supplementary table S5, Supplementary Material online).

In combination, these results confirm that Haplotypes A and B differ at SNP rs7273669 (A/G), which is located 44 bp upstream the translation start site (hereafter, we refer to this SNP as -44(A/G) for simplicity) and presents elevated F_{ST} values in the European/Asian comparison ($F_{ST} = 0.52$; $P = 0.0026$; supplementary fig. S1*B*, Supplementary Material online). This SNP, situated in the 5′-region of *WFDC8*,

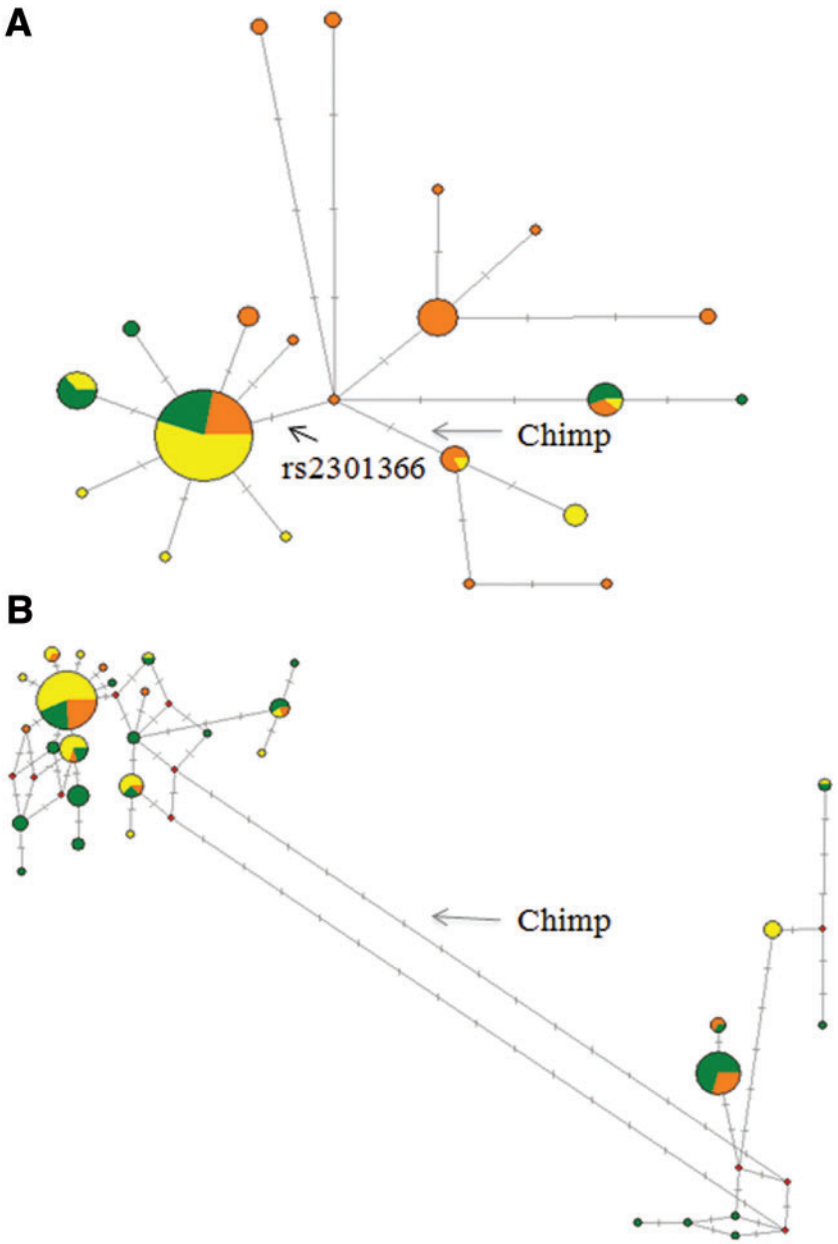

Fig. 6. Examples of inferred network haplotypes at the *WFDC* locus. Each circle represents a unique haplotype, and its area is proportional to its frequency. Within each circle, YRI, CEU, and Asian populations are labeled in orange, green, and yellow, respectively. The mutations that differentiate each haplotype are shown along each branch. The inferred network haplotype of *SEMG1* (*A*) and *WFDC8* (*B*) show a star-like structure (characteristic of a population expansion or recent selective sweep) and two highly differentiated haplotypes (characteristic of population structure or balancing selection), respectively.

potentially affects *cis*-regulatory elements that regulate *WFDC8* expression and has been proposed to affect the binding of two transcription factors (Ferreira et al. 2011).

Footprint of Incomplete Selective Sweep in Africans

We found other regions that stood out as possibly being under selection. A natural target of focused interest was *SPINT4* in *WFDC*-TEL, as this has been previously described as a candidate gene under selection in YRI populations based on the integrated Haplotype Score (Voight et al. 2006) and limited sequencing studies (Ferreira et al. 2011). *SPINT4* and the neighbor gene *WFDC3* do not show significant departure from the patterns expected under neutral evolution in the summary statistics, either in our sequencing study or the 1000 Genomes project data set (supplementary tables S4 and S5, Supplementary Material online), but they display elevated levels of F_{ST} (supplementary fig. S1*B* and *D*, Supplementary Material online) in African/non-African comparisons. Furthermore, variant rs6017667, a NS change that codes for Gly73Ser in *SPINT4*, stands out as the only NS-SNP with intermediate frequency (0.44) (supplementary table S3, Supplementary Material online) and elevated F_{ST} (F_{ST} (Af/As) = 0.26 [$P = 0.04$]; F_{ST} (Af/Eu) = 0.46 [$P = 0.04$]) (supplementary fig. S1*D* and *F*, Supplementary Material online), a signature that is typical of a variant under non-neutral evolution. *SPINT4* was previously thoroughly studied in YRI,

addressing significant EHH/REHH levels and a "star"-shaped genealogy typical of an ongoing positive selection event (Ferreira et al. 2011). Moreover, the Gly73Ser change, previously identified as the candidate variant of the incomplete selective sweep (Ferreira et al. 2011), is encoded by sequences within the second exon of *SPINT4*, which codes for the Kunitz domain and is, therefore, responsible for its serine protease inhibitor activity. Gly73Ser has been identified as a modification that affects stability of SPINT4 and has potential functional repercussions (supplementary table S3, Supplementary Material online).

Discussion

Previous efforts to identify targets of natural selection in the human genome have found an excess of selection acting on genes that mediate response to microbial attack and that play a role in reproduction. By studying the detailed patterns by which natural selection generates deviations from neutral evolution within the *WFDC* region, we can gain insights into the biological roles of specific *WFDC* and *SEMG* genes in host defense and reproduction. This effort to pinpoint signals of selective pressures shows a remarkable degree of interpopulation heterogeneity, identifying different genes under positive selection in three human populations. We hypothesize that this interpopulation heterogeneity is driven by the lack of homogeneity of pathogenic agents across the globe—however, further work will be required to identify agents specifically accounting for the specific patterns seen with the *WFDC* and *SEMG* genes.

The discovery of selective signals highlighted in this study takes into account that demographic history and genetic drift can affect both population differentiation. Although demographic processes affect the whole genome, natural selection acts on specific loci. Hence, the effect of demography must be controlled by comparing the genes of interest with an empirical distribution built from neutrally evolving regions of the genome. In this study, such control regions were represented by 47 unlinked, neutrally evolving pseudogenes (Andrés et al. 2010). We used a strict set of criteria to control the number of false positives and to maximize the detection of specific footprints of natural selection. In addition, because we performed various summary statistic-based tests to describe genetic variation across the *WFDC* locus, we corrected the *P* values for multiple testing by calculating *q* values from the obtained *P* values, estimating the proportion of false positives among the tests found to be significant (Benjamini and Hochberg 1995; Storey 2002; Storey and Tibshirani 2003; Storey et al. 2004). Notwithstanding, most studies presenting a candidate region approach usually present nominal *P* values when referring to a comparison with a particular demographic scenario (Barreiro et al. 2008, 2009; Fornarino et al. 2011; Hancock et al. 2011). Although the tests based on summary statistics failed to survive multiple test correction, they prompted us to pursue an analysis that combined the results of the different tests to probe distinct aspects of the data, including SNP allele frequencies, EHH, and population differentiation (F_{ST}).

Summary statistics, represented by Tajima's *D* and Fay and Wu's *H*′, suggest that *PI3*, *SEMG2*, and *SLPI* show a skew toward low-frequency variants in the Asian population, signals of a population expansion or positive selection. To discriminate between these possibilities, we performed coalescent simulations under a neutral demographic model (Gutenkunst et al. 2009) and compared Tajima's *D* statistic calculated for the sequenced neutrally evolving regions. The test results led us to conclude that *PI3*, *SEMG2*, and *SLPI* are not evolving under neutrality. In addition, the sliding window of Tajima's *D* performed in this region shows extremely negative values in *PI3* and *SEMG*s. Consistently, the summary statistics of these genes in the 1000 Genomes data set present low nucleotide diversity and strongly negative Tajima's *D* values, especially *SEMG1*, which presents the lowest values of the entire *WFDC* locus. These results point toward positive selection acting in this region.

Subsequent analysis of population differentiation of these loci found that some of the SNPs have elevated values of F_{ST}, suggesting the possibility that they might be under region-specific selective pressures. The single NS SNP among those with high F_{ST} values is rs2301366, a variant located on the second exon of *SEMG1* and responsible for the Thr56Ser replacement (ACC→TCC). The derived state of this variant is present in 88% of the Asian samples and defines a haplotype that spans 160 kb. The haplotype-based tests (Hudson haplotype test, DIND, and EHH/REHH) indicate that Ser56 haplotype has unusually low levels of intrahaplotypic diversity and long-range extension given its frequency, which significantly deviates from neutrality under the calibrated model of Asian demography.

Network analysis of the composite haplotypes for all populations suggests the Thr56Ser as the most plausible target of selection. Although the haplotype cladogram shows a star-like structure that can be characteristic of a population expansion, the previous tests performed suggest positive selection in the *PI3-SEMG1-SEMG2* region in Asians, centered on Thr56Ser. Ancestral Thr56 in SEMG1 is highly conserved among primates, dating back to Old World Monkeys, and conserved at position 56 of paralogous gene SEMG2, 79% similar in sequence to SEMG1 (Hurle et al. 2007). The derived allele, Ser56, is also present in African (38% frequency) and European (80% frequency) populations and consistently has a 0.287 Ma age estimate before the "Out-of-Africa" migrations. For the expectations of a classical selective sweep, Ser56 may have weak footprints as indicated also by the shorter haplotype and borderline summary statistics. Conversely, Ser56 may provide a good fit for a model of selection on standing variation in which an allele already segregating in a population is favored by a sudden change in selective pressures (Przeworski et al. 2005; Pritchard et al. 2010; Hernandez et al. 2011). Although the variant Ser56 does not seem to affect SEMG1 protein structure or stability, SEMG1 has a well-established role in forming semen coagulum, crosslinking with SEMG2 to entrap spermatozoa, priming them for optimal fertilization potential (sperm capacitation). Later, this coagulum is degraded by the action of PSA, which cleaves the crosslinking matrix, and releases spermatozoa along with SEMG1- and SEMG2-derived peptides. The N-terminal peptides from SEMG1, with a Serine in position 56, have been described

to have antimicrobial and antiviral activity both in male and female reproductive tracts, whereas the peptides originated from SEMG2, with a Threonine in position 56, do not present antimicrobial activity (Robert and Gagnon 1999; Bourgeon et al. 2004; Edstrom et al. 2008; Zhao et al. 2008; Martellini et al. 2009). The location of Ser56, six amino acids upstream of a mapped PSA cleavage site (Tyr63), may alter the efficacy of the cleavage at this site compared with other primates (Robert et al. 1997; Bourgeon et al. 2004). Here, we hypothesize that the change from Thr to Ser may change the proteolytic cleavage of SEMG1 by PSA, leading to a modified peptide profile and antimicrobial activities.

A signal of short-term balancing selection in Europeans centered in *WFDC8* and the incomplete selective sweep in *SPINT4* were previously described at the *WFDC* region in CEU and YRI, respectively (Ferreira et al. 2011). When the selective signal in *WFDC8* was re-examined in an independent sample, the remarkable differentiation between the Haplotypes A and B in CEU, along with the intermediate frequency at which they are found, solidifies the evidence that *WFDC8* is under a balancing selection in this population. Variant -44(A/G) remains the best candidate SNP under selection, potentially regulating the expression of *WFDC8*. Specifically, the intermediate frequency of these alleles may regulate the levels of *WFDC8* expression, maximizing its role in proteolysis cascades linked to sperm maturation, as well as its antimicrobial functions. In fact, Wfdc8 has been shown to have antibacterial activity in rat male reproductive tract (Rajesh et al. 2011), and its ortholog in humans, *WFDC8*, has been shown to be associated to impaired fertility (Thimon et al. 2008).

Similarly, *SPINT4* is expressed only in testis and epididymis, and it has been associated with sperm maturation in mice (Penttinen et al. 2003; Clauss et al. 2005). A genome-wide association study identified the Gly73Ser (rs6017667) allele of *SPINT4* as being associated with the multifactorial autoimmune disease Type I diabetes (Todd et al. 2007), which has been previously associated with impairments of male reproductive function in humans (Agbaje et al. 2007; Navarro-Casado et al. 2010).

Considering the distinct selective signatures of *SEMG1*, *WFDC8*, and *SPINT4*, we propose that the selection acting on these genes may be related to innate immune functions in the reproductive tract, with possible consequences for fertility. This hypothesis is easier to reconcile with the geographic restriction of selective signatures and the contribution of different alleles from paralog genes to the overall fitness that could be correlated with host–pathogen interaction and with the pathogen load, which largely differs in type and number across geographic regions (Prugnolle et al. 2005; Barreiro et al. 2008; Coop et al. 2009; Fumagalli et al. 2009, 2011; Pritchard et al. 2010; Seixas et al. 2011). However, because of lack of biological knowledge for some of these genes, the precise form of natural selection driving the departures from neutrality remains unclear.

In summary, we propose that the *WFDC* and *SEMG* loci are under adaptive pressures within the short timescale of modern human evolution. *SEMG1*, *WFDC8*, and *SPINT4* are highlighted as the most likely primary targets of selection in this genomic region. Although the signals found in this locus lead us to hypothesize that immune response to pathogens and fertility drive the selective signatures observed, other unknown biological function(s) of the *WFDC* genes cannot be discarded. Additional studies are needed to address how the molecular evolution of *SEMG1* may alter its biochemical properties and how *WFDC8* and *SPINT4* variants can influence the proteolytic and antimicrobial activity in the human reproductive system.

Materials and Methods

DNA Samples

To study genetic variation in the *WFDC* locus, we resequenced the coding regions of 18 *WFDC* and *SEMG* genes (66 exons total) and a number of intervening noncoding regions (spaced every ~10 kb). In parallel, 47 unrelated, neutrally evolving autosomal regions were polymerase chain reaction (PCR) amplified and sequenced as controls. These regions consist of unlinked, ancient processed pseudogenes expected to evolve neutrally in humans and other primates and were previously used as a proxy for neutral sites (Andrés et al. 2010). See supplementary table S1, Supplementary Material online, for the complete list of loci.

All human samples come from the collection of the International HapMap Project Phase I/II. These included a subset of 21 European (CEU: Utah residents with ancestry from northern and western Europe), 25 African (YRI: Yoruba from Ibadan in Nigeria), and 25 Asian (20 CHB: Han Chinese from Beijing in China and 5 JPT: Japanese from Tokyo in Japan) individuals. See supplementary table S2, Supplementary Material online, for sample identification.

Sequence Generation

Primers for amplification and sequencing of the regions of interest were designed based on the Human Genome Reference Sequence from the March 2006 assembly (v36.1), available at the Genome Browser (http://genome.ucsc.edu/, last accessed January 14, 2013). All samples were PCR amplified and analyzed by bidirectional Sanger sequencing. Further details about PCR and DNA sequencing are available from the authors upon request.

Polymorphic sites were detected with the Phred-Phrap-Consed package (Nickerson et al. 1997). Sites found to have a quality score under 99 were manually curated to minimize sequencing errors. The sequencing data were aligned to the Human RefSeq (hg18), and the ancestral state of each SNP was inferred by comparison with the chimpanzee, orangutan, and macaque genome sequences (Chimpanzee Sequencing and Analysis Consortium 2005; Gibbs et al. 2007; Andrés et al. 2010; http://genome.ucsc.edu/, last accessed January 14, 2013).

Statistical Analysis

The DNA sequence data were analyzed using the classical neutrality tests Tajima's *D*, Fay and Wu's *H*', Hudson, Kreitman, and Aguade (HKA), and MWU_{high} (Hudson et al. 1987; Tajima 1989; Fay and Wu 2000; Zeng et al. 2005; Nielsen

et al. 2009). Although none of these tests constitutes a formal test of natural selection, they do provide useful metrics for detecting patterns of departure from neutral variation. Tajima's *D* statistic (Tajima 1989) summarizes the polymorphic DNA frequency spectrum; when significantly negative, it is indicative of excess rare variants, consistent with positive selection, purifying selection, or population expansion. When significantly positive, Tajima's *D* identifies a pattern of variation that is consistent with balancing selection or population subdivision, detecting an increased level of common polymorphisms. The Fay and Wu's *H* statistic (Fay and Wu 2000; Zeng et al. 2006) detects an elevated level of high-frequency derived alleles. When significantly negative, Fay and Wu's *H* indicates a signature of a nearly completed selective sweep. MWU_{high} compares the SFS of a region of interest with the SFS from a neutrally evolving region using the MWU statistical test (Nielsen et al. 2009). MWU_{high} is significant only when there is an excess of intermediate-frequency alleles in the locus of interest. Another approach to detect older positive selection is the HKA test (Hudson et al. 1987), which is based on a contrast between polymorphic and fixed differences levels.

MWU was calculated using an in-house C program, whereas Tajima's *D*, Fay and Wu's *H*, and HKA were calculated using the package `libsequence` (Thornton 2003). To control for demographic effects, we assessed the significance of the obtained summary statistics by comparing them to the distributions of statistics from 10,000 neutral demography-corrected coalescent simulations (ms, Hudson 2002), with population recombination estimates predicted from hg18 (http://genome.ucsc.edu/, last accessed January 14, 2013).

Forty-seven neutrally evolving regions (pseudogenes) were sequenced in the CEU, YRI, and Asian (CHB + JPT) populations and analyzed using the same methodology as in Andrés et al. (2010). This was done to further control for the demographic history effects in the studied populations. Among seven demographic models tested (data not shown), the model proposed by Gutenkunst et al. (2009) provided the best fit (goodness of fit) to our control data set. Thus, those were the population demographic parameters that were subsequently used in the neutral coalescent simulations to provide critical values of test statistics.

In addition to the above model-based approach and taking advantage of having sequenced the *WFDC*s and the control regions in the same individuals, we assessed significance of departure from neutrality by contrasting the distribution of test statistics (e.g., Tajima's *D*) generated from the control regions to the observed statistic from each *WFDC* gene. Specifically, we generated an empirical null distribution by calculating these statistics for each of the control regions in each population. We estimated the upper and lower 2.5 percentiles of each distribution and used these thresholds to assess significance of the statistics of each gene.

The levels of population differentiation at the SNP level were calculated with the classical F_{ST} statistic, which describes the proportion of genetic variance attributable to between-population effects (Excoffier 2002). To identify SNPs presenting extreme levels of F_{ST}, the observed F_{ST} at each SNP within the *WFDC* region was compared with the control regions through a locus-by-locus Analysis of Molecular Variance (AMOVA) approach using 20,000 simulations (Arlequin software package; Excoffier et al. 2007).

The potential functional impact of NS SNPs and fixed differences at the protein level was estimated with the PolyPhen-2v HumDiv (Adzhubei et al. 2010) and SIFT (Kumar et al. 2009) algorithms. Although computational predictions are no substitute for molecular studies that identify measureable functional consequences of protein variants, the consistency of SIFT and PolyPhen results combined with the population genetic inferences can be informative.

Haplotype phasing for all samples was inferred separately for the *WFDC*-CEN and *WFDC*-TEL subloci using PHASE2.1 (Stephens et al. 2001; Stephens and Donnelly 2003). Haploview 4.2 (Barrett 2009) used these phased genotype data to calculate LD statistics (r^2 and D') and to identify clusters of high-LD variants (haplotype blocks) (Gabriel et al. 2002). Cladistic (network) relationships among the haplotypes (Bandelt et al. 1999) were infered with Network 4.5.01 software package.

The recent occurrence of an incomplete (or partial) selective sweep is expected to produce a derived haplotype of unusually elevated frequency, and several tests have been devised to detect such events. One of the first of such tests was developed by Hudson et al. (1994), which estimates the probability of finding a subset of haplotypes with a high frequency and low variation, given the total number of segregating sites in the sample. The haplotype test was performed by simulating 10,000 replicates under neutrality with restricted number of segregating sites, incorporating the recombination rate and demographic model previously described (Gutenkunst et al. 2009). To determine statistical significance, the values estimated for the *PI3-SEMG1-SEMG2-SLPI* haplotype were compared against the obtained background neutral distributions. To evaluate the levels of diversity along the haplotype, we calculated values for Tajima's *D*, π, and haplotype diversity using a sliding window (1,000-bp window size and 500-bp increments) and SLIDER online tool (http://genapps.uchicago.edu/slider/index.html, last accessed January 14, 2013).

We also used the DIND test (Barreiro et al. 2009), which considers the ratio of ancestral to derived intrahaplotypic nucleotide diversity ($i\pi_A/i\pi_D$) plotted against the frequency of the derived allele. Specifically, the DIND test was applied to the sequencing data gathered from the *WFDC*-CEN sublocus for each population. A high-frequency-derived allele associated with an elevated $i\pi_A/i\pi_D$ is indicative of an incomplete selective sweep targeting the derived allelic state.

EHH and REHH (Sabeti et al. 2002) were calculated with SWEEP (http://www.broadinstitute.org/mpg/sweep/, last accessed January 14, 2013). The LRH test (Sabeti et al. 2002), performed to assess statistical significance of REHH, included 50 chromosomes simulations under the Gutenkunst demographic model (Gutenkunst et al. 2009; ms, Hudson 2002), for five 500 kb sequence assuming the same population mutation parameter and recombination rate as estimated for the entire *WFDC-CEN* region. Core

haplotypes were set using SWEEP as the longest nonoverlapping cores with no more than three SNPs, with EHH/REHH statistics calculated for 80 kb distance from cores. The significance of EHH/REHH statistics was estimated by comparing values with the null distribution of core haplotypes within the same 5% frequency bin of Ser56.

The T_{MRCA} and neutral parameter θ_{ML} for all the populations were estimated using a maximum likelihood coalescent method implemented in GENETREE version 9 (Griffiths and Tavare 1994). Rare recombinant haplotypes carrying homoplasic mutations were removed from the analysis. We took into account the possibility of selective forces acting on one mutation by estimating the β parameter (Coop and Griffiths 2004). Strictly, the model of Coop and Griffiths constructs a likelihood ratio test on the selection parameter β, contrasting the likelihoods under a null (neutral) model ($\beta = 0$) to that with selection ($\beta \neq 0$), where the neutral model is for a population of constant size. Time, scaled in $2N_e$ generations, was derived from $\theta_{ML} = 4N_e\mu$. The estimate of the mutation rate per gene per generation (μ) was obtained from the average number of nucleotide substitutions per site (D_{xy}) between human and chimpanzee reference sequences, as calculated in DnaSP v.5.1 (Rozas et al. 2003). Time estimates in generations were converted into years using a 25-year generation time. Human/chimpanzee divergence was assumed to have occurred approximately 5.4 Ma (Patterson, Richter, et al. 2006). The likelihood ratio test of Coop and Griffiths was found by simulation to be robust to the impact of demographic change in the parameter range of human population growth. In addition, when we tested for selection in the control genome regions by this test, the neutral null hypothesis was not rejected.

To increase sample size and further test the robustness of our results, we downloaded the corresponding sequenced regions from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). We performed a principal component analysis using EIGENSOFT to study the population structure in the *WFDC* locus, using the information of all the SNPs regardless of the LD between them (Patterson, Price, et al. 2006; Price et al. 2006) and calculated summary statistics for every gene in each population using SLIDER. The correlations and independence between the summary statistics of both data sets were determined by Kendall's rank and Spearman's ρ correlations, and χ^2 test.

Supplementary Material

Supplementary tables S1–S7 and figures S1–S7 are available online at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge Anh-Dao Nguyen for the help inferring the ancestral allele state of each SNP and Guillaume Laval for making available the scripts for the DIND test. This work was supported in part by the Intramural Research Program of the National Human Genome Research Institute, by the SFRH/BD/45907/2008 fellowship from the Portuguese Foundation for Science and Technology (FCT) to Z.F., by the POPH-QREN-Promotion of scientific employment to Z.F. and S.S., supported by the European Social Fund and national funds of the Portuguese Ministry of Education and Science, and by the Wellcome Trust Centre for Human Genetics (WT097307) to W.W.K. IPATIMUP is an Associated Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods.* 7:248–249.

Agbaje IM, Rogers DA, McVicar CM, McClure N, Atkinson AB, Mallidis C, Lewis SE. 2007. Insulin dependant diabetes mellitus: implications for male reproductive function. *Hum Reprod.* 22:1871–1877.

Andrés AM, Dennis MY, Kretzschmar WW, et al. (13 co-authors). 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* 6:e1001157.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16:37–48.

Barreiro LB, Ben-Ali M, Quach H, et al. (18 co-authors). 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5:e1000562.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet.* 40:340–345.

Barrett JC. 2009. Haploview: visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc.* 2009:pdb.ip71.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 57:289–300.

Bingle CD, Vyakarnam A. 2008. Novel innate immune functions of the whey acidic protein family. *Trends Immunol.* 29:444–453.

Bouchard D, Morisset D, Bourbonnais Y, Tremblay GM. 2006. Proteins with whey-acidic-protein motifs and cancer. *Lancet Oncol.* 7: 167–174.

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

Carnahan SJ, Jensen-Seaman MI. 2008. Hominoid seminal protein evolution and ancestral mating behavior. *Am J Primatol.* 70:939–948.

Chowdhury MA, Kuivaniemi H, Romero R, Edwin S, Chaiworapongsa T, Tromp G. 2006. Identification of novel functional sequence variants in the gene for peptidase inhibitor 3. *BMC Med Genet.* 7:49.

Clauss A, Lilja H, Lundwall A. 2005. The evolution of a genetic locus encoding small serine proteinase inhibitors. *Biochem Biophys Res Commun.* 333:383–389.

Clauss A, Persson M, Lilja H, Lundwall.Å. 2011. Three genes expressing Kunitz domains in the epididymis are related to genes of WFDC-type protease inhibitors and semen coagulum proteins in spite of lacking similarity between their protein products. *BMC Biochem.* 12:55.

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

Coop G, Griffiths RC. 2004. Ancestral inference on gene trees under selection. *Theor Popul Biol.* 66:219–232.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation. *PLoS Genet.* 5:e1000500.

de Lamirande E. 2007. Semenogelin, the main protein of the human semen coagulum, regulates sperm function. *Sem Thromb Hemost.* 33:60–68.

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.

Drmanac R, Sparks AB, Callow MJ, et al. (65 co-authors). 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81.

Edstrom AM, Malm J, Frohm B, Martellini JA, Giwercman A, Morgelin M, Cole AM, Sorensen OE. 2008. The major bactericidal activity of human seminal plasma is zinc-dependent and derived from fragmentation of the semenogelins. *J Immunol.* 181: 3413–3421.

Excoffier L. 2002. Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev.* 12:675–682.

Excoffier L, Laval G, Schneider S. 2007. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online.* 1:47–50.

Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

Ferreira Z, Hurle B, Rocha J, Seixas S. 2011. Differing evolutionary histories of WFDC8 (short-term balancing) in Europeans and SPINT4 (incomplete selective sweep) in Africans. *Mol Biol Evol.* 28: 2811–2822.

Fornarino S, Laval G, Barreiro LB, Manry J, Vasseur E, Quintana-Murci L. 2011. Evolution of the TIR domain-containing adaptors in humans: swinging between constraint and adaptation. *Mol Biol Evol.* 28: 3087–3097.

Fu YX. 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics* 143:557–570.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.* 19:199–212.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355.

Gabriel SB, Schaffner SF, Nguyen H, et al. (18 co-authors). 2002. The structure of haplotype blocks in the human genome. *Science* 296: 2225–2229.

Gibbs RA, Rogers J, Katze MG, et al. (177 co-authors). 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.

Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344: 403–410.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5: e1000695.

Hancock AM, Clark VJ, Qian Y, Di Rienzo A. 2011. Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Mol Biol Evol.* 28:601–614.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920–924.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. 1994. Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* 136:1329–1340.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.

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

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

Kingan SB, Tatar M, Rand DM. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. *J Mol Evol.* 57:159–169.

Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 4:1073–1082.

Lundwall A. 2007. A locus on chromosome 20 encompassing genes that are highly expressed in the epididymis. *Asian J Androl.* 9: 540–544.

Lundwall A, Bjartell A, Olsson AY, Malm J. 2002. Semenogelin I and II, the predominant human seminal plasma proteins, are also expressed in non-genital tissues. *Mol Hum Reprod.* 8:805–810.

Lundwall A, Clauss A. 2011. Genes encoding WFDC- and Kunitz-type protease inhibitor domains: are they related? *Biochem Soc Trans.* 39: 1398–1402.

Martellini JA, Cole AL, Venkataraman N, Quinn GA, Svoboda P, Gangrade BK, Pohl J, Sorensen OE, Cole AM. 2009. Cationic polypeptides contribute to the anti-HIV-1 activity of human seminal plasma. *FASEB J.* 23:3609–3618.

McKiernan PJ, McElvaney NG, Greene CM. 2011. SLPI and inflammatory lung disease in females. *Biochem Soc Trans.* 39:1421–1426.

McNeely TB, Shugars DC, Rosendahl M, Tucker C, Eisenberg SP, Wahl SM. 1997. Inhibition of human immunodeficiency virus type 1 infectivity by secretory leukocyte protease inhibitor occurs prior to viral reverse transcription. *Blood* 90:1141–1149.

Navarro-Casado L, Juncos-Tobarra MA, Chafer-Rudilla M, Iniguez-de Onzono L, Blazquez-Cabrera JA, Miralles-Garcia JM. 2010. Effect of experimental diabetes and STZ on male fertility capacity: study in rats. *J Androl.* 31:584–592.

Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25:2745–2751.

Nielsen R, Hubisz MJ, Hellmann I, et al. (13 co-authors). 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* 19:838–849.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2:e190.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.

Penttinen J, Pujianto DA, Sipila P, Huhtaniemi I, Poutanen M. 2003. Discovery in silico and characterization in vitro of novel genes exclusively expressed in the mouse epididymis. *Mol Endocrinol.* 17: 2138–2151.

Peter A, Lilja H, Lundwall A, 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.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 38:904–909.

Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20:208–215.

Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 15:1022–1027.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59(11):2312–2323.

Rajesh A, Madhubabu G, Yenugu S. 2011. Identification and characterization of Wfdc gene expression in the male reproductive tract of the rat. *Mol Reprod Dev.* 78:633–641.

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.

Robert M, Gagnon C. 1999. Semenogelin I: a coagulum forming, multifunctional seminal vesicle protein. *Cell Mol Life Sci.* 55:944–960.

Robert M, Gibbs BF, Jacobson E, Gagnon C. 1997. Characterization of prostate-specific antigen proteolytic activity on its major physiological substrate, the sperm motility inhibitor precursor/semenogelin I. *Biochemistry* 36:3811–3819.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Seixas S, Ivanova N, Ferreira Z, Rocha J, Victor BL. 2011. Loss and Gain of Function in SERPINB11: an example of a gene under selection on standing variation, with implications for host-pathogen interactions. *PLoS One* 7(2):e32518.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68:978–989.

Storey JD. 2002. A direct approach to false discovery rates. *J R Stat Soc Ser B.* 64:479–498.

Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc Ser B.* 66: 187–205.

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide experiments. *Proc Natl Acad Sci U S A.* 100:9440–9445.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Thimon V, Calvo E, Koukoui O, Legare C, Sullivan R. 2008. Effects of vasectomy on gene expression profiling along the human epididymis. *Biol Reprod.* 79:262–273.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

Todd JA, Walker NM, Cooper JD, et al. (40 co-authors). 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet.* 39:857–864.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A.* 102:18508–18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

Wang Z, Widgren EE, Sivashanmugam P, O'Rand MG, Richardson RT. 2005. Association of eppin with semenogelin on human spermatozoa. *Biol Reprod.* 72:1064–1070.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–276.

Weldon S, McGarry N, Taggar CC, McElvaney NG. 2007. The role of secretory leucoprotease inhibitor in the resolution of inflammatory responses. *Biochem Soc Trans.* 35:273–276.

Weldon S, Taggart CC. 2007. Innate host defense functions of secretory leucoprotease inhibitor. *Exp Lung Res.* 33:485–491.

Whitlock MC. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol.* 18: 1368–1373.

Williams SE, Brown TI, Roghanian A, Sallenave JM. 2006. SLPI and elafin: one glove, many fingers. *Clin Sci.* 110:21–35.

Yenugu S, Richardson RT, Sivashanmugam P, Wang Z, O'Rand M G, French FS, Hall SH. 2004. Antimicrobial activity of human EPPIN, an androgen-regulated, sperm-bound protein with a whey acidic protein motif. *Biol Reprod.* 71:1484–1490.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174: 1431–1439.

Zhao H, Lee WH, Shen JH, Li H, Zhang Y. 2008. Identification of novel semenogelin I-derived antimicrobial peptide from liquefied human seminal plasma. *Peptides* 29:505–511.