Differing Evolutionary Histories of WFDC8 (Short-Term Balancing) in Europeans and SPINT4 (Incomplete Selective Sweep) in Africans

Zélia Ferreira,1,2,3 Belen Hurle,3 Jorge Rocha,1,2 and Susana Seixas*1
1Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal
2Department of Zoology and Anthropology, Faculty of Sciences, University of Porto, Porto, Portugal
3Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD
*Corresponding author: E-mail: sseixas@ipatimup.pt.
Associate editor: Willie Swanson

Abstract

The whey acidic protein four-disulfide core (WFDC) gene cluster on human chromosome 20q13, harbors 15 small serine protease inhibitor genes with roles in innate immunity, reproduction, and regulation of endogenous proteases kallikreins. The WFDC cluster has emerged as a prime example of rapid diversification and adaptive evolution in primates. This study sought a better understanding of the evolutionary history of WFDC genes in humans and focused on exploring the adaptive selection signatures found in populations of European (Utah residents with ancestry from northern and western Europe [CEU]) and African (Yoruba from Ibadan, in Nigeria [YRI]) ancestry in a genome-wide scan for putative targets of recent adaptive selection. Our approach included resequencing coding and noncoding regions of WFDC6, EPPIN, and WFDC8 in 20 CEU and of SPINT4 in 20 YRI individuals. We generated 302 kb and 60 kb of high-quality sequence data from CEU and of YRI populations, respectively, enabling the identification of 72 single nucleotide polymorphisms. Using classic neutrality tests, empirical and haplotype-based analysis, we pinpointed WFDC8 and SPINT4 as the likely targets of short-term balancing selection in the CEU population, and recent positive selection (incomplete selective sweep) in the YRI population. Putative candidate variants targeted by selection include 44A (rs7273669A) for WFDC8, which may downregulate gene expression by abolishing the binding site of two transcription factors; and a haplotype configuration [Ser73+98A] (rs6017667A–rs6032474A) for SPINT4, which may simultaneously affect protein function and gene regulation. We propose that the evolution of WFDC8 and SPINT4 has been shaped by complex selective scenarios due to the interdependence of variant fitness and ecological variables.

Key words: WFDC, natural selection, innate immunity, serine protease inhibitor, reproduction.

Introduction

The availability of dense catalogues of human genomic variation has led to identification of intriguing “outlier” loci that present unusual patterns of genomic variation compared with the rest of the genome. These catalogues help unravel the confounding effects of natural selection (affecting genomic variation at specific loci) and population demographic history (shaping patterns of variation at all loci in a genome) (Biswas and Akey 2006; Biswas et al. 2009).

A growing number of genome-wide scans (GWS) for positive selection in humans suggest that certain types of genes are overrepresented among those that have been targets of positive selection: Among those are proteolysis genes and genes related to reproduction and immune functions (Wang et al. 2006; Akey 2009). Consistent with this, a recent GWS using the integrated haplotype score (IHS) and HapMap Phase II data (Voight et al. 2006) identified candidate genes bearing putative signals of recent positive selection within the whey acidic protein (WAP) four-disulfide core domain (WFDC) gene cluster on human chromosome 20q13 (fig. 1). Specifically, two signals of positive selection were found spanning the intervals from WFDC6 to EPPIN (in populations of European descent or CEU) and from WFDC10A to SPINT4 (in populations of African descent or YRI), respectively.

The WFDC cluster encompasses 15 genes encoding small serine protease inhibitors with characteristic WAP and/or for Kunitz domains that confer serine protease inhibitor and antibacterial activities (Clauss et al. 2005; Macedo-Ribeiro et al. 2008; McCrudden et al. 2008). Thus, a number of WFDC genes encode proteins with confirmed roles in innate immunity (SLPI, Zhu et al. 2002; PI3, Sallenave 2010), reproduction (EPPIN; O’Rand M et al. 2004), and regulation of the endogenous protease kallikreins (KLK, Lundwall et al. 2006). Although the functions of most other WFDC genes are poorly characterized, their similar structural domains suggest related functions. The neighboring seminal genes Semenogelin 1 and 2 (SEMG1 and SEMG2) also play a central role in fertility and immunity (Lundwall 2007; Edstrom et al. 2008).

With their concomitant key roles in reproduction and innate immunity, WFDC and SEMG genes appear to occupy crossroads of multiple, interconnected biological processes...
including host-pathogen interactions, the effects of different mating systems on postcopulatory sperm competition, and male attempts to counteract the female immune response (Dorus et al. 2004; Hurle et al. 2007; Ramm et al. 2008). The relative contributions of these (or other) selective pressures to the overall evolution of these genes are currently unknown.

Complementary reports indicate that the number of genes under selection at the WFDC cluster may be quite large. For instance, initial comparison of the human and chimpanzee genome sequences identified the WFDC cluster as one of 16 genomic regions with an unusually high density of rapidly evolving genes (Chimpanzee Sequencing and Analysis Consortium 2005). Likewise, an independent study involving resequencing of the WFDC centromeric subcluster in 12 primates showed strong patterns of positive selection in a striking number of contiguous genes (WFDC12, P3, SEMG1, SEMG2, and SLPI) (Hurle et al. 2007). Others have confirmed strong positive selection acting on SEMG genes (Kingan et al. 2003; Dorus et al. 2004; Ramm et al. 2008).

The present study sought a better understanding of the selection pressures acting on WFDC genes and focused on the positive selection signatures found in CEU and YRI populations by Voight et al. 2006. Our approach included reanalyzing HapMap Phase II haplotype data and resequencing the coding and noncoding regions of WFDC6, EPPIN, and WFDC8 in 20 CEU samples and of SPINT4 in 20 YRI samples. Individual [H] single nucleotide polymorphism (SNP) scores and classic neutrality tests enabled us to identify haplotypes linked to advantageous alleles that are consistent with positive selection centered in WFDC8 and in SPINT4. We propose that WFDC8 and SPINT4 have been shaped by short-term balancing selection (in CEU) and by an incomplete selective sweep (in YRI), a likely result of the interdependence of variant fitness and ecological variables.

Our study is the first to identify WFDC8 and SPINT4 as bearers of unique, complex selective signatures that signal their prominent biological role among other WFDC genes, and the need for their further functional characterization. The results contribute significantly to mounting evidence that WFDCs are genes under strong adaptive pressures in primate evolution, diversifying in ways that are perceptible even within the short timescale of modern humans.

**Methods**

**DNA Samples**

All human samples used in the current study belong to the collection of the International HapMap Project Phase I/II. For the sequencing variation study, we surveyed a subset of 20 Europeans (CEU: Utah residents with ancestry from northern and western Europe) for WFDC6, EPPIN, WFDC8 genes and another subset of 20 African individuals (YRI: Yoruba from Ibadan, in Nigeria) for SPINT4 (supplementary table S1, Supplementary Material online). For typing the rs7273669 SNP, all available HapMap PhaseI/Phase II samples (232) were analyzed including those from Asians (Japanese from Tokyo, Japan [JPT]; and Han Chinese from Beijing, China [CHB]) (supplementary table S1, Supplementary Material online).

**Polymerase Chain Reaction and Sequencing**

Primers for amplification and sequencing were designed based on the GenBank (www.ncbi.nlm.nih.gov) sequence entries for chromosome 20 (NC_000020.9) between bases 43596249 and 43787010. Sequencing was performed using the ABI BigDye Terminator version 3 cycle sequencing chemistry (Applied Biosystems, Foster City, CA), and electrophoretic analysis was done on an ABI 3130 automated sequencer. All human sequences were assembled and analyzed using the Phred-Phrap-Consed package (Nickerson et al. 1997). All putative polymorphisms and software-derived genotype calls were manually inspected and individually confirmed using
Consect. Further details of polymerase chain reaction (PCR) and sequencing conditions are available from the authors upon request.

**Genotyping of the WFDC8 Candidate Variant**

The putative candidate variant found in the CEU samples (rs7273669A or -44A for brevity) was genotyped by PCR-restriction fragment length polymorphism analysis with the restriction enzyme BseRI (New England BioLabs) in full HapMap Phase I/II sample collection (CEU, YRI, JPT, and CHB). The primers used for amplification of a fragment encompassing rs7273669 SNP were 5’-acatagaagtt-gagttgct-3’ and 5’-acgcctacctgtcataagtg-3’.

**Statistical Analysis**

Phased haplotypes from the International HapMap Project Phase II (release 21), for a 200 kb region centered on *EPPIN* in the CEU sample and on *SPINT4* in the YRI sample, were downloaded from the HapMap website (http://hapmap.ncbi.nlm.nih.gov/). Haplotype data were then annotated with additional SNP information regarding ancestral allele state and potential selected sites. Ancestral allele state was retrieved from dbSNP (http://www.ncbi.nlm.nih.gov/), and potential selected sites were identified using the Haplotter application (http://hg wen.uchicago.edu/selection/haplotter.htm), which displays the results of a GWS for positive selection based on the iHS statistic and relying on HapMap Phase I/II data sets (Voight et al. 2006). A |iHS| > 2 threshold, corresponding to the top 5% iHS values of the entire genome (Voight et al. 2006), was used to identify the potential selected sites.

Phased HapMap data for chromosome 20 were uploaded to Sweep 1.81 software (http://www.broad.mit.edu/mpg/sweep) and used to compute extended haplotype homozygosity (EHH) and relative extended haplotype homozygosity (REHH) (Sabeti et al. 2002). We defined cores as the longest nonoverlapping core haplotypes with at least three SNPs and no more than 20 SNPs, encompassing at least one SNP with significant iHS values, and containing the largest number of chromosomes carrying the candidate selected allele. Long-range haplotype tests were conducted at the largest physical distances, with REHH maximized and EHH close to 0.05. Significance of REHH, given the frequency of core haplotype, was calculated in Sweep 1.81 assuming 5% frequency bins.

For the WFDC resequencing survey, summary statistics of population genetic variation were calculated using SLIDER (http://genapps.uchicago.edu/slider/index.html); MAXDIP (http://genapps.uchicago.edu/labweb/index.html); and H test (http://www.genetics.wustl.edu/jflab/htest.html). Linkage disequilibrium (LD) and nucleotide diversity of the haplotypes were determined using the software DnaSP version 4.9 (Rozas 2009).

The haplotypes of the *WFDC6-EPPIN-WFDC8* array and *SPINT4* were inferred by using the program PHASE 2.02 (Stephens et al. 2001; Stephens and Donnelly 2003), where SNPs previously inferred by the International HapMap Project Phase II were input as known phase to benefit from the trios genotyped in both CEU and YRI samples. To provide a temporal dimension to the phylogenetic relationships among haplotypes and to estimate the coalescent times and ages of relevant mutations, we used GENETREE version 9.0 (Griffiths and Tavare 1994). Five rare recombinant haplotypes carrying homoplastic mutations in the CEU population were removed from the WFDC8 analysis. The mutation rate per gene, per generation, was deduced from the average number of nucleotide substitutions per site between human and chimpanzee reference sequences, calculated with DnaSP v.4.9 (Rozas 2009). Time estimates in generations were converted into years using a 25-year generation time. Human/chimpanzee divergence was assumed to have occurred 5.4 Ma (Patterson et al. 2006).

To assess the statistical significance of summary statistics, we ran 100,000 coalescent simulations (Hudson 2002) using estimates of the population recombination and mutation rate parameters calculated from our own data using MAXDIP and SLIDER. Simulations were produced using the “ms” program, assuming distinct demographic models including constant population size, expansion, bottleneck, structured population, and African and European best-fit models (Sabeti et al. 2002; Schaffner et al. 2005; Voight et al. 2005; Wang et al. 2006). For each model, we obtained a null distribution of summary statistics values and calculated the 2.5th and 97.5th percentiles.

**Prediction of Transcript Factor Binding**

Binding of transcript factors was predicted for both 5’ and 3’ untranslated regions (UTRs) of all the sequenced genes by the online available tool Mapper (Marinescu et al. 2005b). This program searches for putative transcription factor binding sites based on hidden Markov models built on alignments with experimentally known sites provided by the TRANSFAC (http://www.biobaseinternational.com/pages/index.php?id=transfac) and JASPAR (http://jaspar.cgb.ki.se/) databases.

**Results**

**Selection Signatures at the WFDC Cluster According to HapMap Phase II Data**

According to the GWS for recent positive selection based on the iHS statistic described by Voight et al. 2006 and HapMap Phase II data, two signals of positive selection reside in the *WFDC* cluster. One spans genes *WFDC6* and *EPPIN* in CEU populations (significant empirical *P* values of *P* = 0.0448); another lies in the genomic segment containing *WFDC10A, WFDC11, WFDC10B, WFDC13,* and *SPINT4* in YRI populations (significant *P* values ranging from 0.0406 to 0.0441). For more detailed evaluation of the putative signals, we defined nonsliding windows of 200 kb centered either on *EPPIN* (CEU) or *SPINT4* (YRI) and mined phased HapMap Phase II data to identify all SNPs within the region with significant |iHS| values (supplementary files 1 and 2, Supplementary Material online). Consequently, ancestral allele state information was combined with the positive or negative significant |iHS| score of each SNP to determine...
configurations of tightly linked alleles defining the longest haplotypes in the region. Hereafter, we will refer to extended haplotypes associated with a potential target of selection as "A" haplotypes and those unassociated as "B" haplotypes.

The extended "A" haplotype centered on EPPIN in the CEU population had an approximate frequency of 41% with very low genetic diversity (supplementary file 1 and supplementary fig. 1A and B, Supplementary Material online). Surprisingly, all 16 SNPs with significant \( i_{HS} \) positive values within the 200 kb-long window were clustered with the EPPIN adjacent gene, WFDC8 (2.08 < \( i_{HS} \) < 2.10); in other words, none was linked to WFDC6 or EPPIN as previous findings would have predicted (Voight et al. 2006). Thus, in subsequent studies, WFDC8 was included as a putative target for recent positive selection in the CEU population.

The extended "A" haplotype centered on SPINT4 in the YRI population was a short, ~30 kb fragment spanning SPINT4 and nearby pseudogene HNRPA1P3 (G: 51511747) and included a large cluster of 30 SNPs with significant \( i_{HS} \) scores (2.0 < \( i_{HS} \) < 3.6) (supplementary file 2, Supplementary Material online). Given its frequency of approximately 80%, this "A" haplotype appears to present unexpectedly low variation (supplementary fig. 1C and D, Supplementary Material online). Thus, we hypothesize that SPINT4 may be the single target of recent positive selection in the YRI population.

Resequencing for SNP Discovery at the WFDC Cluster

We next resequenced the WFDC6-EPPIN-WFDC8 array and the SPINT4 gene in a subset of 20 CEU and 20 YRI HapMap individuals, respectively. This step was necessary to avoid distortion of the allele frequency spectrum due to the ascertainment nature of HapMap Phase II data and to uncover functional variation in the candidate genes. Figure 1 summarizes the resequencing strategy, which targeted coding exons and splicing junctions of the candidate genes but also included a number of small introns. Three tagger SNPs were also included in the resequencing survey to act as anchors for the "A" and "B" haplotype configurations as defined by HapMap Phase II data and to tag the two clusters of SNPs with significant \( i_{HS} \) scores centered in either WFDC8 (represented by tagger SNP rs6104221 in CEU) or SPINT4 (represented by tagger SNPs rs1386504 and rs6032474 in YRI), respectively.

Overall, we generated 302 kb of high-quality sequence data from the CEU sample and 60 kb from the YRI sample, enabling identification of 72 SNPs (fig. 2). Importantly, 24 (33%) of those 72 sites lacked any previously associated reference SNP identification code in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/); and 19 (26%) of them, although tagged with a dbSNP number, had never been typed by the HapMap initiative. Upon integrating the new SNPs into the HapMap Phase II shell, we reexamined the haplotypes and nonscertained patterns of diversity across the WFDC6-EPPIN-WFDC8 array and the SPINT4 gene. Specific findings for WFDC8 (in the CEU population) and SPINT4 (in the YRI population) are discussed below.

WFDC8 in the CEU Population

Our sequencing and genotyping initiatives more than doubled the SNPs typed across the WFDC6-EPPIN-WFDC8 gene array, exposing a total of 18 SNPs that differentiate the "A" and the "B" haplotypes. Initially, two positions in complete LD with the "A" haplotype stood out for their potential functional consequences: nonsynonymous amino acid replacement Asn137Ser in WFDC8; and SNP rs7273669G/A located 44 bp upstream from the transcription-start site of WFDC8 (hereafter −44A for brevity).

However, upon closer examination, Asn137Ser was considered an unlikely target of selection for several reasons: The allele associated with the "A" haplotype is the ancestral one; codon 137 is not conserved among WFDC8 orthologues; and the amino acid replacement is not predicted to affect protein stability or structure, as determined by Polyphen analysis (Ramensky et al. 2002).

On the other hand, −44A was predicted by Mapper (Marinescu et al. 2005a) to abolish, in the selected "A" haplotype, the binding sites of two ubiquitously expressed transcript factors—the myeloid zinc finger (MZF) and the peroxisome proliferator activated receptor (PPAR). Interestingly, G to A substitutions in transcription factor binding sites strongly contribute to loss of function and are more prone than other substitutions to alter gene expression (Lapidot et al. 2008). No functional variants in LD were detected with the "A" haplotype in association with WFDC6 and EPPIN. Thus, we regard −44A as the best candidate variant among those that have been identified for being targeted by selection at the WFDC6-EPPIN-WFDC8 gene array. The status of −44A as a putative candidate variant warranted extended genotyping of this SNP in 232 HapMap Phase II individuals (see supplementary table S1, Supplementary Material online). When all available −44G/A data were inserted into the HapMap Phase II haplotypes, the −44A allele was not found in the CHB and JPT populations but was present in the YRI and the CEU populations (15% and 41%, respectively).

We questioned whether the long-range haplotypes would be maintained when haplotype analysis was centered on −44A. Thus, the haplotype structure across the WFDC8 locus was reanalyzed with the EHH and REHH statistical tests, which use multiple SNPs in a core to increase statistical power (Sabeti et al. 2002; Sabeti et al. 2005) (fig. 3). In the CEU population, three major core haplotypes were identified, spanning positions from −4672 to +8722 (SNPs rs6032336 to rs6104239) relative to the ATG start codon of WFDC8. The observed frequencies reached 39% for the haplotype associated with the derived allele −44A (core GTTAGTGAAGCAAAGTTC), and frequencies of 29% and 14% for the two haplotypes associated with the ancestral allele −44G (cores GCTAATCAAGGCCGATATCC and GCTAAACAGCGCGGATATCC, respectively; fig. 3A). Moreover, the high REHH values of core GTTAGTGAAGCAAAGTTC extend an additional 410 kb (or 0.15 cM) on the telomeric side of WFDC8, with a value of 5.096—significantly elevated when compared with the distribution of REHH scores for HapMap Phase II data from
chromosome 20 and for the haplotype cores found at equivalent frequencies (35–40%; \( p = 0.034 \); fig. 38). A similar analysis in the YRI population, centered on position –44, found narrow cores, short haplotypes, and nonsignificant REHH \( P \) values (result not shown).

Using classic neutrality tests and empirical distribution analyses, \( WFDC8 \) significantly departed from neutral expectations in the CEU population. For instance, \( WFDC8 \) showed unusually high nucleotide diversity (\( \pi = 21.0 \)) and a strong positive value (2.38) of Tajima’s \( D \) (Tajima 1989). Meanwhile, flanking genes \( WFDC6 \) and \( EPPIN \) yielded unremarkable values of \( \pi \) and Tajima’s \( D \) (table 1). Also for \( WFDC8 \), both statistics significantly deviated from neutrality in seven out of eight theoretical null distributions generated by coalescent simulations that assumed distinct demographic models (table 2). The departure from neutrality was particularly significant when the demographic model by (Schaffner et al. 2005) was tested with parameters calibrated to better fit the CEU data from HapMap Phase I. Furthermore, an empirical comparison with the subset of 316 genes from the SeattleSNPs project (http://pga.gs.washington.edu/; Crawford et al. 2005) that have been sequenced in CEU samples confirmed the outlier nature of \( WFDC8 \), with \( \pi \) and Tajima’s \( D \) values falling within the 97.5th percentiles of the two empirical distributions in SeattleSNPs (supplementary fig. 2A and B, Supplementary Material online). Thus, we concluded that \( WFDC8 \) significantly deviates from global trends of diversity in Europeans, with \( \pi \) and Tajima’s \( D \) values indicating high sequence divergence and excess intermediate frequency alleles in the region.

We reconstructed the genealogy of \( WFDC8 \) and estimated the time to most recent common ancestor (\( T_{\text{MRCA}} \)) using a maximum likelihood coalescent analysis (Griffiths and Tavare 1994). Interestingly, \( WFDC8 \) showed an atypical structure dominated by two deep-root branches with intermediate frequencies. Each branch precisely segregated with either the potential selected allele –44A or the opposite allele –44G (fig. 4). Significantly, the tree topologies of adjacent genes in the same haplotype block \( WFDC6 \) and \( EPPIN \) (supplementary fig. 3, Supplementary Material online) were in conformity with collected statistics for genes evolving under neutral evolution (Excoffer 2002; Tishkoff and Verrelli 2003; Satta and Takahata 2004; Garrigan and Hammer 2006). Given that previous findings (significant values of \( IHS \), REHH, Tajima’s \( D \), and \( \pi \)), support a nonneutral interpretation of the evolution of \( WFDC8 \), the uncommon tree structure can be attributed to balancing selection acting exclusively on \( WFDC8 \). The recent \( T_{\text{MRCA}} \) estimate of 1.40 ± 0.27 Ma, coupled with the atypical deep-root tree topology, does not suggest the effect of ancestral population subdivision, as genes with deep-root branches affected by population structure in humans have longer \( T_{\text{MRCA}} \) estimates of 2–3 Ma (Hey and Harris 1999; Garrigan and Hammer 2006; Hayakawa et al. 2006; Patin et al. 2006; Kim et al. 2010).

**SPINT4** in the YRI Population

The density of SNPs typed across the \( SPINT4 \) locus increased from eight SNPs typed in HapMap Phase II to a total of 27 SNPs. The divergence of the “A” and “B”

---

**FIG. 2.** “A” and “B” haplotypes for \( WFDC6 \), \( EPPIN \), \( WFDC8 \), and \( SPINT4 \) as inferred by PHASE2.02 software. The ancestral state at each site was inferred based on orthologous nonhuman primate sequences. SNPs with significant \( IHS \) statistics are identified. Coding variants are marked by an asterisk. These included, in the CEU population, one synonymous amino acid replacement in \( EPPIN \) (Thr10Thr) and two nonsynonymous amino acid replacements in \( WFDC8 \) (Asn137Ser and Thr96Met); and in the YRI population three nonsynonymous replacements in \( SPINT4 \) (Ala30Glu, Gly73Ser, and Ser73Arg). SNP identifiers and their chromosomal positions based on NC000020 reference sequence are indicated in columns. SNPs typed in HapMap Phase II are in a white background. SNPs not typed by HapMap with a dbSNP reference number are in light gray background.
haplotypes was determined by a total of 11 SNPs, two of which had potential functional consequences for SPINT4: a nonconservative change at position Gly73Ser and non-coding SNP rs6032474A/G in the 3' UTR. Conveniently, both candidate SNPs had been surveyed by the HapMap Phase II project in all HapMap individuals. The derived allele Ser73 introduces a nonconservative change at a highly conserved codon within the Kunitz domain, which in turn may significantly affect SPINT4 structure and stability. Likewise, derived allele +98G is predicted to abolish a binding site for the Forkhead Box J2 (FOXJ2) transactivator factor in the 3' UTR of SPINT4. Interestingly, in the YRI population, the derived allele Ser73 is in complete LD with the ancestral allele +98A (\(\text{\small jD^2} = 1\), \(\text{\small r}^2 = 1\)), making it infeasible to dissociate the selective advantages of each individual allele. Thus, the best candidate for selection in the SPINT4 locus may be a two-SNP haplotype configuration rather than a single allele. For brevity, we will refer to the proposed advantageous variant as [Ser73 +98A] and to the opposite variant as [Gly73 +98G]. The [Ser73 +98A] signature is prevalent in the YRI population, at 80% frequency, significantly higher than its modest frequency in CEU (36%) and JPT+CHB (33%) populations.

FIG. 3. Haplotype-based tests of selection using HapMap Phase II data. Plots of EHH and REHH over physical distance for the largest nonoverlapping cores encompassing the -44G/A (rs7274789) WFDC8 variants (core haplotypes from rs6032336 to rs6104239 SNPs) (A) and for Gly73Ser+98A/G SPINT4 variants (rs11908541–rs6130908) (C). Plots of REHH versus frequency for chromosome 20 using the CEU sample and a 410 kb distance (B) and using the YRI sample and a 150 kb distance (D). Core haplotype sequences are indicated below REHH plots and candidate variants are underlined. The stars represent the results for -44A bearing chromosomes (B) and for [Ser73+98G] bearing chromosomes (D).

Table 1. Summary Statistics of Population Variation.

<table>
<thead>
<tr>
<th>Population</th>
<th>Gene</th>
<th>N(^a)</th>
<th>(L)(^b)</th>
<th>(S)(^c)</th>
<th>(\theta_w)(^d)</th>
<th>(\pi)(^e)</th>
<th>(D)(^f)</th>
<th>(H)(^g)</th>
<th>(D)(^h)</th>
<th>(r)(^i)</th>
<th>(p)(^j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europeans (CEU: Utah residents with ancestry from northern and western Europe)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WFDC6</td>
<td>40</td>
<td>4243</td>
<td>9</td>
<td>4.98</td>
<td>6.08</td>
<td>0.65</td>
<td>0.41</td>
<td>1.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPPIN</td>
<td>40</td>
<td>5288</td>
<td>6</td>
<td>2.67</td>
<td>2.43</td>
<td>-0.24</td>
<td>0.69</td>
<td>6.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WFDC8</td>
<td>40</td>
<td>5675</td>
<td>30</td>
<td>12.4</td>
<td>21.0</td>
<td>2.38*</td>
<td>0.35</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Africans (YRI: Yoruba from Ibadan in Nigeria)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPINT4</td>
<td>40</td>
<td>2877</td>
<td>27</td>
<td>22.06</td>
<td>21.19</td>
<td>-0.134</td>
<td>-1.99*</td>
<td>1.81</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* N = number of chromosomes.
\(L\) = total number of sites surveyed.
\(S\) = number of segregating sites.
\(\theta_w\) = Watterson’s estimator of \(\theta\) (\(4Ne\)) (Watterson 1975) per base pair (\(\times 10^6\)).
\(\pi\) = Nucleotide diversity per base pair (\(\times 10^{-8}\)).
\(D\) = Tajima’s D statistic (Tajima 1989).
\(H\) = Fay and Wu H statistic (Fay and Wu 2000; Zeng et al. 2006).
\(D\) = Hudson’s estimator of \(q\) (\(4Ne\)) per base pair (\(\times 10^{-8}\)) based on a conversion-to-crossover ratio of 2 and a mean conversion tract length of 500 bp (Frisse et al. 2001; Hudson 2001).
\(p\) = Statistically significant P < 0.05.

Phase II project in all HapMap individuals. The derived allele Ser73 introduces a nonconservative change at a highly conserved codon within the Kunitz domain, which in turn may significantly affect SPINT4 structure and stability. Likewise, derived allele +98G is predicted to abolish a binding site for the Forkhead Box J2 (FOXJ2) transactivator factor in the 3' UTR of SPINT4. Interestingly, in the YRI population, the derived allele Ser73 is in complete LD with the ancestral allele +98A (\(\text{\small jD} = 1\), \(\text{\small r}^2 = 1\)), making it infeasible to dissociate the selective advantages of each individual allele. Thus, the best candidate for selection in the SPINT4 locus may be a two-SNP haplotype configuration rather than a single allele. For brevity, we will refer to the proposed advantageous variant as [Ser73+98A] and to the opposite variant as [Gly73+98G]. The [Ser73+98A] signature is prevalent in the YRI population, at 80% frequency, significantly higher than its modest frequency in CEU (36%) and JPT+CHB (33%) populations.

EHH/REHH statistics identified two major haplotype cores in the YRI population. These spanned positions +109 to +4776 (or SNPs rs11908541–rs6130908) with respect to the start codon of SPINT4, and included 13 SNPs (three with significant iHS values). The observed frequencies reached...
### Table 2. Percentile 97.5 of the Null Distributions Generated by Coalescent Simulations.

<table>
<thead>
<tr>
<th>Model</th>
<th>WDFC8 (CEU)</th>
<th>SPINT4 (YRI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td>n</td>
</tr>
<tr>
<td>Constant</td>
<td>1.78*</td>
<td>10.70*</td>
</tr>
<tr>
<td>Recent expansion</td>
<td>0.91*</td>
<td>8.93*</td>
</tr>
<tr>
<td>2-fold growth</td>
<td>1.74*</td>
<td>10.43*</td>
</tr>
<tr>
<td>Short and severe bottleneck</td>
<td>2.25*</td>
<td>11.67*</td>
</tr>
<tr>
<td>Long and mild bottleneck</td>
<td>1.11*</td>
<td>9.35*</td>
</tr>
<tr>
<td>Structure</td>
<td>2.08*</td>
<td>11.31*</td>
</tr>
<tr>
<td>Structure</td>
<td>2.15*</td>
<td>11.45*</td>
</tr>
<tr>
<td>Best fit</td>
<td>2.43</td>
<td>12.02*</td>
</tr>
<tr>
<td></td>
<td>1.93*</td>
<td>11.03*</td>
</tr>
</tbody>
</table>

**Notes:**
- N—effective size; t—time in generations; b—bottleneck intensity; npop—number of populations; m—migration rate per generation.
- Tajima's D statistic (Tajima 1989).
- Nucleotide diversity per base pair ($\times 10^4$).
- Model from Wang et al. 2006.
- Model from Sabeti et al. 2002.
- Model from Schaffner et al. 2005.
- *Statistically significant—the observed statistic is higher than the 97.5th values.

77% for core TACCACACTGTGC, including the proposed advantageous variant [Ser73+98A], and 15% for core TCACGCGGCCCATC, including the opposite variant (Gly73+98G). Furthermore, core TACCACACTGTGC extended an additional 150 kb (0.02 cM) on the 3' side of SPINT4 with stronger LD and higher EHH/REHH statistics ($REHH = 2.24$) than core TCACGCGGCCCATC (Fig. 3C). This REHH value was also significantly higher than the REHH values for other haplotype cores found at equivalent frequencies on chromosome 20 (75–80%; $P = 0.015$; Fig. 3D). By contrast, in non-African populations, core TACCACACTGTGC did not yield significant REHH P values.

SPINT4 deviated from neutral expectations in all eight theoretical null distributions of π generated by coalescent simulations assuming distinct demographic models (table 2). The H test (Fay and Wu 2000; Zeng et al. 2006) was also significant for SPINT4 ($-1.99; P = 0.032$). Unexpectedly, when we estimated the Tajima's D distribution for the YRI population, SPINT4 showed unremarkable, slightly negative Tajima's D values that generally followed the trend in the YRI population (supplementary fig. 2C, Supplementary Material online); however, when individuals were sorted with respect to intrahaplotype diversity, the haplotype linked to [Ser73+98A] was associated with significantly negative Tajima's D and D* statistics of Fu and Li (Fu and Li 1993), whereas the haplotype linked to [Gly73+98G] was not (table 3). Importantly, the haplotype linked to [Ser73+98A] was more skewed toward rare variants (six singletons and two doubletons) and lower nucleotide diversity than was the opposite haplotype (supplementary fig. 4, Supplementary Material online). Empirical comparisons of SPINT4 with the subset of 316 SeattleSNPs genes studied in the YRI population (103 genes) or African-American population (213 genes) showed that SPINT4 had an unusually high π value achieved by few other SeattleSNPs genes (supplementary fig. 2, Supplementary Material online). In conclusion, SPINT4 presents a significant H test, a high π value deviating from the global trends of diversity in Africans, and a haplotype linked to [Ser73+98A] that shows strongly negative Tajima's D and D* values of Fu and Li, hinting at a nonneutral haplotype configuration.

As in the case of WDFC8, the tree topology of SPINT4 was dominated by an atypical pattern with two deep-root branches (fig. 4). However, in the case of SPINT4, the higher $T_{MRCA}$ estimate of 3.07 ± 0.55 Ma could theoretically be attributed to either selection or ancestral population substructure. Likewise, the branch associated with the candidate signature [Ser73+98A] was linked to a “star”-shape genealogy, which might also be connected with either a positive selective event (occurring about 0.93 ± 0.34 Ma for node A and 0.58 ± 0.18 Ma for node B) or with population expansion. When a selection parameter $\beta$ (Coop and Griffiths 2004) is taken into account in reconstructing the coalescent process, positive selection is more likely than no selection. In addition, when selection is assumed, minimal age estimates of the candidate signature [Ser73+98A] are significantly reduced, bringing the selective event to a more recent time (−0.26 ± 0.05 Ma). Also taken into consideration was that Africa’s demographic history does not suggest that population substructure could significantly account for two highly divergent haplotypes segregating at intermediate frequencies in the YRI population. Taken together, the tree topology and unusual statistics suggest that SPINT4 is undergoing an incomplete selective sweep.

**Discussion**

GWS have made a critical contribution to understanding the genetic bases of natural selection (Biswas and Akey 2006; Scheinfeldt et al. 2009). Nevertheless, it is essential to validate the putative selective signals and follow up with detailed case-by-case gene analysis (Biswas and Akey 2006; Teshima et al. 2006). The WDFC cluster, which harbors genes involved in reproduction and immunity, was identified as containing putative targets of recent positive selection in a GWS based on iHS statistics. In this study, we pinpoint WDFC8 and SPINT4 as the likely targets of balancing selection in the CEU population, and recent positive selection in the YRI population, respectively. In addition, we identify putative allele variants with potential functional consequences that might confer selective advantage.

Not surprisingly, iHS statistics performed on individual SNPs prevailed over gene-centered (window-based) iHS statistics to narrow down the selective signals from candidate regions to individual genes. This is most likely due to ascertainment bias and factors such as gene size and gene
density that influence or distort iHS statistics on window-based analyses (Seixas et al. 2007). For instance, gene-centered iHS analysis identified \(WFDC6\) and \(EPPIN\) as likely targets of selection in the CEU population; yet iHS statistics on individual SNPs indicated nearby \(WFDC8\) (with lower iHS values) as the likeliest target for selection. Similarly, it was not \(WFDC11\) or \(WFDC10B\) within the corresponding window, but rather \(SPINT4\) that had a more significant \(P\) value and the strongest evidence of selection in the YRI population.

Our results indicate that the \(-44A\) \(WFDC8\) allele (perhaps affecting the \(WFDC8\) regulatory control) has reached intermediate frequencies of \(\sim 40\%\) in Europe; and the haplotype configuration \([\text{Ser73} + 98A]\) (presumably modifying \(SPINT4\) function and regulation) is nearing fixation in Africa, with frequencies of \(\sim 80\%\). However, the sequence variation and haplotype patterns suggest complex selective histories for \(WFDC8\) and \(SPINT4\) that are difficult to resolve by modeling standard scenarios of recent positive selection or long-term balancing selection.

It is possible that the selection signals are difficult to tease apart because \(WFDC8\) and \(SPINT4\) are evolving under complex adaptive scenarios for which statistics that rely on the frequency spectrum of sequence variation have limited sensitivity. Several alternative selective scenarios could explain the large \(\pi\) values associated with \(WFDC8\) and \(SPINT4\). For example, in an ongoing selective sweep—with a new, rare allele being rapidly selected from introduction to fixation—\(\pi\) might increase its value as the allele reaches intermediate frequencies (Fay and Wu 2000; Przeworski et al. 2005; Zeng et al. 2006; Gordo et al. 2009). Alternatively, directional selection on standing variation—in which a neutral or mildly deleterious allele arises and drifts in the population, becomes beneficial at frequency \(f\), and eventually reaches fixation—can be associated with large \(\pi\) values (Przeworski et al. 2005; Chevin and Hospital 2008). Thus, depending on the value of \(f\), it can be challenging to discriminate between signals that indicate selective sweep acting on a newly arisen allele or selection on standing variation on a “rare” allele (Przeworski et al. 2005). It is also possible that incomplete sweeps and initial stages of balancing selection might produce quite

---

**Table 3. SPINT4 Intrahaplotype Diversity.**

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>(N)</th>
<th>(S)</th>
<th>(\theta_w)</th>
<th>(H^I)</th>
<th>(D^f)</th>
<th>(D^*)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPINT4 (YRI sample)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>30</td>
<td>10</td>
<td>8.77</td>
<td>3.16</td>
<td>-2.02*</td>
<td>-2.36*</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>7</td>
<td>8.60</td>
<td>7.72</td>
<td>-0.43</td>
<td>-0.13</td>
</tr>
</tbody>
</table>

\(N = \) number of chromosomes.

\(S = \) number of segregating sites.

\(\theta_w = \) Watterson’s estimator of \(\theta = (4N_e \mu)\) (Watterson 1975) per base pair (\(\times 10^4\)).

\(H^I = \) Nucleotide diversity per base pair (\(\times 10^4\)).

\(D^f = \) Tajima’s \(D\) statistic (Tajima 1989).

\(D^\* = \) Fu and Li (Fu and Li 1993).

* Statistically significant \(P < 0.05\).
similar overall frequency spectrum patterns (Fay and Wu 2000; Przeworski et al. 2005; Charlesworth 2006; Zeng et al. 2006; Gordo et al. 2009).

For WFDC8, the remarkable divergence between "A" and "B" haplotypes at first resembled a prototypical pattern of long-term balancing selection. However, the homogeneity and length of the "A" haplotype spanning WFDC8 was unusual: unlikely to result from local low recombination or even to derive from the ancestral structure of the CEU population (Schaffner et al. 2005; Garrigan and Hammer 2006; Gutenkunst et al. 2009). Instead, we believe that the extended haplotype observed for WFDC8 is likely to correlate with the effects of "short-term" balancing selection. Several arguments support our hypothesis: The haplotype extension is not spurious, but allele-specific and linked to −44A; the selective signal is clearly captured through iHS or REHH statistics for which variation in local recombination is controlled; −44A has few linked polymorphic sites suggesting a recent, nonneutral increase in frequency; and the recent \( T_{MRCA} \) estimate of 1.40 ± 0.27 Ma, coupled with the atypical deep-root tree topology, does not suggest the effect of ancestral population subdivision (>2 Ma) (Hey and Harris 1999; Garrigan and Hammer 2006; Hayakawa et al. 2006; Patin et al. 2006; Kim et al. 2010). Similarly, the summary statistics \( \pi \) and Tajima's \( D \) are not in full concordance with the expectations of recent positive selection favoring a newly arisen variant. Considering all these observations, we propose a short-term balancing selection event for WFDC8 with the following characteristics: an "A" haplotype swept to intermediate frequencies; linked variation including both rare and common variants (as if the selective sweep was recently stabilized); linked neutral common variation starting to accumulate in the newest allele; and differences between alleles not yet eroded by recombination.

SPI2T4, on the other hand, presented unusually high values of \( \pi \), an overall slightly negative Tajima's \( D \) value, and a significant \( H \) test value, all of which may be found when an advantageous allele is increasing in frequency toward fixation (Fay and Wu 2000; Przeworski et al. 2005; Zeng et al. 2006; Gordo et al. 2009). Moreover, all tests centered on "A" haplotype aspects, including EHH/REHH, Tajima's \( D \), and \( D^* \) of Fu and Li, support SPI2T4 as a gene under an incomplete selective sweep—a new allele increasing its frequency rapidly toward fixation, thus becoming associated with an excess of rare variants. Interpretation of the selective scenario of SPI2T4 could be further complicated by the existence of two candidate variants in complete LD, or the haplotype configuration [Ser73+98A]. Still, these were traced to a single deep-rooted branch in the SPI2T4 tree, and to a "star"-shape genealogy typical of a positive selection event. Moreover, the literature describes other cases in which a two-SNP haplotype configuration has been swept to higher frequencies by natural selection. One example is a two-SNP haplotype configuration [Val72-Le93] in the N-acylsphingosine amidohydrolase (ASAH1) gene involved in hydrolysis of ceramides and regulation of neuronal development (Kim and Satta 2008).

Regardless of the complexity of the underlying selective scenario(s), a selective (nonneutral) interpretation of the highly divergent lineages observed at the WFDC8 and SPI2T4 loci is more plausible than alternative models based solely on neutrality or ancestral population substructure. The signals of selection on WFDC8 and SPI2T4 were initially identified through a GWS for positive selection in which possible effects of demography, recombination, and SNP density and the allele frequency spectrum were somewhat controlled, making a selective interpretation more plausible than demographic effects. In comparisons with null distributions generated by neutral coalescent simulations under different demographic scenarios, and in empirical comparisons with SeattleSNPs genes, both genes are almost invariably "outliers," which also supports a selective interpretation. Finally, the tree topologies of two genes within the same haplotype block as WFDC8 (WFDC6 and EPPIN) conformed with collected statistics for genes evolving under neutral evolution—making the very restricted, locus-specific signal on WFDC8 harder to explain under neutrality or demographic hypotheses alone.

Expression of WFDC8 and SPI2T4 is mainly restricted to testis and epididymis, with the gene activity potentially extending over a large area in the male reproductive tract. As with other serine protease inhibitors (i.e., SERPINAs, SERPINE2, EPPIN, SLPI, and PI3, Uhrin et al. 2000; Murer et al. 2001; O’Rand et al. 2007; Wang et al. 2007a, 2007b; McCrudden et al. 2008), WFDC8 and SPI2T4 are likely to play a role in regulating proteolysis cascades linked to the maturation and capacitation of sperm cells or they may be related to innate immune function in the male genital tract. For instance, WFDC8 expression in epididymis cauda is lost after vasectomy and may contribute to the impaired fertility after vas deferens reanastomosis (Thimon et al. 2008); and a GWS identified the Gly73Ser (rs6017667) allele of SPI2T4 as being associated with the multifactorial autoimmune disease, Type I diabetes (T1D) (Todd et al. 2007). Interestingly, T1D has also been associated with impairments of male reproductive function in humans and in animal models, which show structural modifications of testis and epididymis and damaged sperm cells (Agbaje et al. 2007; Navarro-Casado et al. 2010). Considering the complex selective signatures of WFDC8 and SPI2T4 (which might still represent recent adaptations), we propose that the selection acting on WFDC8 and SPI2T4 may be related to innate immune functions in the reproductive tract, with possible consequences for fertility levels. This hypothesis is easier to reconcile with the geographic restriction of selective signatures that could be correlated with host-pathogen interaction and with the pathogen load, which largely differs in genus and number across world geographic regions (Prugnolle et al. 2005; Fumagalli et al. 2009). We hypothesize that the WFDC8 candidate variant −44A abolishes the binding sites of two transcription factors, possibly downregulating gene expression. This could provide a selective advantage related to augmentation of proteolytic activity in male secretions, which ultimately may
facilitate sperm cell capacitation and gain of motility. Alternatively, this variant might contribute to lower anti-inflammatory and antibacterial properties of male secretions, increasing the risk of developing urogenital infections. Given the much higher pathogen burden in Africa compared with Europe, the lower frequencies of −44A variant in the YRI sample and the geographic specificity of the selective signature to the CEU sample might be explained by differential allele fitness, subject to environmental and ecological variables.

The same adaptive pressure, driven by host-pathogen interactions, may drive selection of the SPINT4 haplotype configuration [Ser73+98A] close to fixation in the YRI population. Here, joining the two alleles in a haplotype configuration—one leading to a modified Kunitz domain (of reduced inhibitory activity) and the other maintaining an active transactivator binding site—may increase the innate immunity function without compromising the proteolytic features (necessary for fertility) of male secretions. Exploring these hypotheses will clearly require functional characterization of the candidate variants; but the assumption of a fluid balance between the activities of WFDGs in regulating fertility and innate immunity is a plausible fit for the complex evolutionary scenarios evoked for WFDG8 and SPINT4.

In conclusion, the links between WFDG8 and SPINT4 and complex selective scenarios probably reflect their concomitant key roles in reproduction and innate immunity. Functional characterization of the proposed selected variants −44A (WFDG8) and [Ser73+98A] (SPINT4) will be fundamental to gaining greater understanding of the evolutionary forces driving their evolution as well as their contribution to male fertility. Additional primate studies on the intraspecific diversity of these and other WFDG genes will contribute importantly to our understanding of their critical roles in reproduction and immunity.

Supplementary Material
Supplementary table S1, files 1 and 2, and figures 1–5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
We would like to thank Aida Andres, Max Planck Institute, for critically reading the manuscript and for insightful discussions; Jacquelyn K. Beals for the editing; and the two anonymous reviewers for their helpful feedback. This research was supported by the Portuguese Foundation for Science and Technology (FCT), project grant to S. Seixas—PTDC/SAU-GMG/64043/2006. Z. Ferreira is supported by SFRH/BD/45907/2008 fellowship from FCT, supported by POPH-QREN—Promotion of scientific employment, supported by the European Social Fund and national funds of the Ministry of Science, Technology and Higher Education. S. Seixas is supported by POPH-QREN—Promotion of scientific employment, supported by the European Social Fund and national funds of the Ministry of Science, Technology and Higher Education. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by FCT.

References


