

Rapid Evolution, Genetic Variations, and Functional Association of the Human Spermatogenesis-Related Gene *NYD-SP12*

Qü Zhang,^{1,2} Feng Zhang,³ Xiao-hua Chen,^{1,2} Yin-qiu Wang,^{1,2} Wei-qi Wang,⁴ Alice A. Lin,⁵ Luca L. Cavalli-Sforza,⁵ Li Jin,³ Ran Huo,⁶ Jia-hao Sha,⁶ Zheng Li,⁷ Bing Su^{1,2}

¹ Key Laboratory of Cellular and Molecular Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China

² Kunming Primate Research Center, Chinese Academy of Sciences, Kunming, Yunnan, China

³ State Key Laboratory of Genetic Engineering and Center for Anthropological Studies, School of Life Sciences, Fudan University, Shanghai, China

⁴ Department of Pathology, Third Affiliated Hospital of Kunming Medical College, Kunming, Yunnan, China

⁵ Department of Genetics, Stanford University, Stanford, California, USA

⁶ Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing, Jiangsu, China

⁷ Renji Hospital, Shanghai Second Medical University, Shanghai, China

Received: 3 August 2006 / Accepted: 22 January 2007 [Reviewing Editor: Dr. Manyuan Long]

Abstract. *NYD-SP12* is a recently identified spermatogenesis-related gene with a pivotal role in human testis development. In this study, we analyzed between-species divergence and within-species variation of *NYD-SP12* in seven representative primate species, four worldwide human populations, and 124 human clinical subjects. Our results indicate that *NYD-SP12* evolves rapidly in both the human and the chimpanzee lineages, which is likely caused by Darwinian positive selection and/or sexual selection. We observed significant interpopulation divergence among human populations, which might be due to the varied demographic histories. In the association analysis, we demonstrated significant frequency discrepancy of a synonymous sequence polymorphism among the clinical groups with different sperm traits.

Key words: Genetic variations — Darwinian positive selection — Spermatogenesis — Sexual selection — Human evolution

Introduction

Reproductive proteins, which mediate gamete generation, storage, usage, signal transduction, and fertilization, diverge rapidly, often as a result of adaptive evolution (Civetta and Singh 1995; Swanson and Vacquier 2002). This may contribute to the establishment of reproductive barriers leading to speciation, as well as other important biological function. Accordingly, there has been an intense interest in identifying reproductive-related genes with accelerated amino acid changes (Dorus et al. 2004; Swanson and Vacquier 2002; Wang and Zhang 2004; Wyckoff et al. 2000).

NYD-SP12 is a newly reported testis-specific gene (Xu et al. 2003), involved in the formation of acrosome during spermatogenesis (Lu et al. 2006). It is located on human chromosome 3q26.32, encoding a protein of 569 amino acids, and highly expressed in testis during human sexual maturation (Xu et al. 2003). Furthermore, patients with spermatogenic arrest and Sertoli cell-only syndrome have reduced expression of *NYD-SP12* transcripts in the testis, indicating a possible role of *NYD-SP12* in spermatogenesis (Xu et al. 2003).

In this study, we conducted an evolutionary survey of sequence substitution patterns of *NYD-SP12* in diverse primate species to detect potential positive selection. We also compared the sequence variants of *NYD-SP12* in worldwide human populations, as well as three clinical groups with different sperm traits to detect regional selection and functional association. Our data demonstrated that there were adaptive amino acid changes in the primate lineages leading to human and chimpanzee, which were likely caused by Darwinian positive selection. Also, there is significant interpopulation divergence, which might be due to the varied demographic histories. Among the clinical subjects, we observed a significant difference in the frequencies of one-nucleotide polymorphisms.

Materials and Methods

DNA Samples

We sequenced 72 human individuals from the major continental populations, including 19 Africans, 13 Europeans, 30 East Asians (20 Chinese and 10 Cambodians), and 10 Melanesians. We also sequenced 124 Chinese clinical samples, whose sperm number, motility, and morphology were analyzed according to the World Health Organization (1999) criteria. The clinical samples were categorized into three groups: (1) the asthenozoospermia group (AST), whose sperm counts were >20 million/ml but with low sperm motility ($A + B < 50\%$; 43 samples); (2) the normal group (NOR), with normal sperm counts and motility (20 million–200 million/ml; $A + B > 50\%$; 48 samples), and (3) the donor group (DON), with high sperm counts (>200 million/ml; $A + B > 50\%$; 33 samples). In addition, six nonhuman primate species were also sequenced, reflecting a 45 million-year history of primate evolution (Goodman et al. 1998). The nonhuman primate panel included three great ape species (one chimpanzee, *Pan troglodytes*; one gorilla, *Gorilla gorilla*; and one orangutan, *Pongo pygmaeus*), one lesser ape species (white-browed gibbon, *Hylobates hoolock*), one Old World monkey species (rhesus monkey, *Macaca mulatta*), and one New World monkey species (black-handed spider monkey, *Ateles geoffroyi*). All the DNA samples were from collections at the Kunming Cell Bank of CAS, Kunming Blood Center, Shanghai Renji Hospital, and Stanford University. Informed consent was obtained from the clinical subjects.

Sequencing Analysis

All human and nonhuman primate samples were sequenced for the full-length coding region of *NYD-SP12* (1707bp). Primers for all the primates were designed by aligning the published sequences of human and mouse (*Ensembl* genome browser; <http://www.ensembl.org>). The primer sequences are listed below.

- Exon2_f: 5' CCACTGGAATTCAAAATATTGG 3'
- Exon2_r: 5' TTCATAGTTTCTGTCAACCAATG 3'
- Exon3_f: 5' TAAGCAAGCATAAGCAAAACATT 3'
- Exon3_r: 5' AAGCTCAGATAGTACCCTCCACA 3'
- Exon4_f: 5' ATTCATTCCATTGCACTTGTGC 3'
- Exon4_r: 5' AATTGCTGGTTTGACGTTGATAGT 3'
- Exon5_f: 5' AAAAATCCCTACTTCAACACTCTT 3'
- Exon5_r: 5' CCAGAGCTACCTTTTCTTTCAGTG 3'
- Exon6_f: 5' ATATGTTATCTTTGCAGTACAGTAA 3'
- Exon6_r: 5' CAGAAACAATAAACTAATAAACCAC 3'

- Exon7_f: 5' GATATGTTATCTATTTGACCTTTTC 3'
- Exon7_r: 5' CTAACCCACAAAGAAGATTAGAT 3'
- Exon8_f: 5' CATCTCCTCTGCCTTTCTCCG 3'
- Exon8_r: 5' CGTTTCCAAAATCCCCTAACAG 3'
- Exon9_f: 5' ATCCGTCTCAACAGTTCTCCAAT 3'
- Exon9_r: 5' TGTTGGCTCTGGGCCTATGT 3'
- Exon10_f: 5' CAATAATATTTCGGCAATGACAAGG 3'
- Exon10_r: 5' TATTGATTTTGGTCACACTTCCAC 3'
- Exon11_f: 5' CCTGCACTTCTGCTTTGTGTTAT 3'
- Exon11_r: 5' GTGGATGCCCTTGCCTCTTG 3'

PCRs were performed with rTaq under conditions recommended by the manufacturer (Takara Company).

Sequencing was performed in both directions with forward and reverse primers using the BigDye terminator sequencing kit on an ABI 3130 automated sequencer. There are 11 exons in the *NYD-SP12* gene, and the first exon is nontranslational and, therefore, was not sequenced in this study.

Data Analysis

DNA sequences were aligned using DNASTAR (DNASTAR, Inc.) and the MEGA3.0 program (Kumar et al. 2004) and checked manually. The Z-test was performed by MEGA3 (Kumar et al. 2004). The PAML (Yang 1997) package was used to identify positive selection and individual sites under selection. To detect selection, we first inferred the ancestral sequences of the internal nodes from the phylogenetic tree by Baseml and calculated lineage-specific dN/dS (ω) by yn00. We also used the branch models and branch-site models in Codeml to identify branches and sites under selection without inferring the ancestral sequences. The sliding window calculation was performed by K-Estimator (Comeron, 1995, 1999).

McDonald-Kreitman's (1991) neutrality test was used to test selection and performed by DnaSP (Rozas and Rozas 1999). DnaSP was also used to perform three tests: Tajima's (1989) D statistic, which considers the difference between θ_W and π ; Fay and Wu's (2000) H test, which considers the difference between θ_H and π ; and Fu and Li's (1993) D and F tests (Fu and Li 1993), which compares the observed number of singleton polymorphisms with those expected under a neutral model. Statistic significance was determined by 5000 simulations. F_{ST} was estimated using the program Arlequin, which can be found at <http://www.lgb.unige.ch/arlequin>. Phase I of the HapMap project (Altshuler et al. 2005) provides SNP information from 90 individuals of 30 trios of Yoruba in Ibadan, Nigeria (YRI); 90 individuals of 30 trios of Utah residents with ancestry from northern and western Europe (CEU); 45 unrelated Han Chinese individuals from Beijing, China (CHB); and 44 unrelated Japanese individuals from Tokyo, Japan (JPT). The web tool, Haplotter, at <http://www.pritch.bsd.uchicago.edu/data.html>, which pooled CHB and JPT samples to form a single sample pool, was used to query the HapMap data for evidence of selection by calculating the empirical distributions of the iHS , Tajima's D , Fay and Wu's F , and the F_{ST} statistics (Voight et al. 2006). We used Haploview at <http://www.broad.mit.edu/mpg/haploview/index.php> to calculate linkage disequilibrium, and the SNP data with location information were downloaded from the HapMap Project.

Results

Positive Selection in Human and Chimpanzee

We sequenced the complete coding region of *NYD-SP12* in seven primate species including human. The

Human	MDAGSSRSLE	NAVNR1YHDQ	LVPKINTSKK	MSTL ^A HPPNI	LEMSQEIKKN	CGGKQVEITL
ChimpanzeeV.....P
GorillaP.....
Orangutan
Gibbon	S.....	R.....
rhesus_monkeyK.....P.....
Spider_monkeyK.S.PP.M.....D.....P
Human	ERTKMTKGIK	EKQSNL ^E KA	A ^E KRKAEGEE	KPTRK ^A AKI	TELDNQLITM	PLPHIPLKNI
Chimpanzee	..I.....Q.....	..L.....E.....
Gorilla	..GL.....R.....E.....
Orangutan	..I.....S.....E.....I.....
Gibbon	..I.....S.....E.....I.....R.....
rhesus_monkey	..I.....S.....E.....I.....
Spider_monkeyS.....V.....AG.....E.....I.P.....R.....
Human	MDVEMKLVYI	DEMGVRYEFV	ESFMSTGSQP	TCQAAEIVDP	LSVHNFSFLP	QIDKWLQVAL
ChimpanzeeV.....R.....
GorillaV.....I.....G.....
OrangutanV.....R.A.....
GibbonV.H.....I.R.....S.....
rhesus_monkeyV.....RV.....P.....
Spider_monkeyE.S.....P.....P.....
Human	KDASSCYRQK	KYALAAGQFR	TALELCSKGA	VLGEPFDAPA	EDIASVASFI	ETKLVTCYLR
ChimpanzeeA.....
Gorilla
Orangutan
Gibbon
rhesus_monkeyA.....
Spider_monkeyM.....E.....G.....
Human	MRKPDALANH	AHRSIVLNPA	YFRNHLRQAT	VFRCLERYSE	AARSAMIADY	MFWLGGGR ^E
ChimpanzeeQ
GorillaQ
OrangutanQ
GibbonQ
rhesus_monkeyQ
Spider_monkeyA.....A.....S.Q
Human	SISKLIKLYW	QAMIEEAITR	AESFSVMYTP	FATK ^R RADKI	EKVKDAFTKT	HPAYA ^E MYT
ChimpanzeeM.....Q.....
Gorilla
OrangutanV.....I.....
Gibbon	D.....A.....	R.....I.....
rhesus_monkeyV.....I.....
Spider_monkeyG.....V.....I.P
Human	DLQALHMLPQ	TVDWSSFPPQ	QYLLTLGFKN	K ^D DGKFLEKI	SSRKLPIFTE	HKTPFGLTRE
ChimpanzeeE.....
GorillaE.....
OrangutanE.....
GibbonE.....S.....
rhesus_monkeyT.....E.....
Spider_monkeyL.....E.....T.....V.....Q.....
Human	DTVRQMETMG	KRILPILDFI	RSTQLNGSFP	ASSGVMEK ^L Q	YASLLSQLQR	VKEQSQVINQ
ChimpanzeeF.....
Gorilla
OrangutanG.....
GibbonV.....D.....
rhesus_monkeyG.....
Spider_monkeyR.....C.....V.....E.....
Human	AMAELATIPY	LQDISQGEAE	LLQSLMADAM	DTLEGRNNN	ERVWNMIQKV	GQIEDFLYQL
ChimpanzeeR.....D.....T.....
GorillaR.....S.....T.....
OrangutanW.....D.....T.....
GibbonT.....D.....T.....I.....
rhesus_monkeyR.....D.....T.....
Spider_monkeyL.....DK.....A.K.....R.....
Human	EDSFLK ^E KKL	RTARRQKTKM	KRLQTV ^E QR
ChimpanzeeA.....H.S
GorillaS
OrangutanS
GibbonS
rhesus_monkeyI.S
Spider_monkeyN.....RS

Fig. 1. Protein sequence alignment of *NYD-SP12* in human and nonhuman primates. The amino acid sites under positive selection by the branch-site model are shaded.

protein sequence alignment is shown in Fig. 1. To investigate the general pattern of nucleotide substitutions, the well-established phylogenetic tree of the major primate lineages was used (Goodman et al. 1998; Page and Goodman 2001) (Fig. 2). Following the strategy in Messier and Stewart's (1997) work on primate lysozyme, we reconstructed the ancestral sequences. The nonsynonymous substitution number

(N) and synonymous substitution number (S), as well as the d_N/d_S ratio (ω), for each branch of the tree were calculated (Fig. 2). Our results identified episodes of adaptive evolution in both the human lineage ($\omega = 4.17$, $P < 0.05$, Z-test) and the chimpanzee lineage ($\omega = \infty$, $P = 0.01$). Since the reconstructed ancestral sequences involve random errors and systematic biases (Yang 1998), we used the

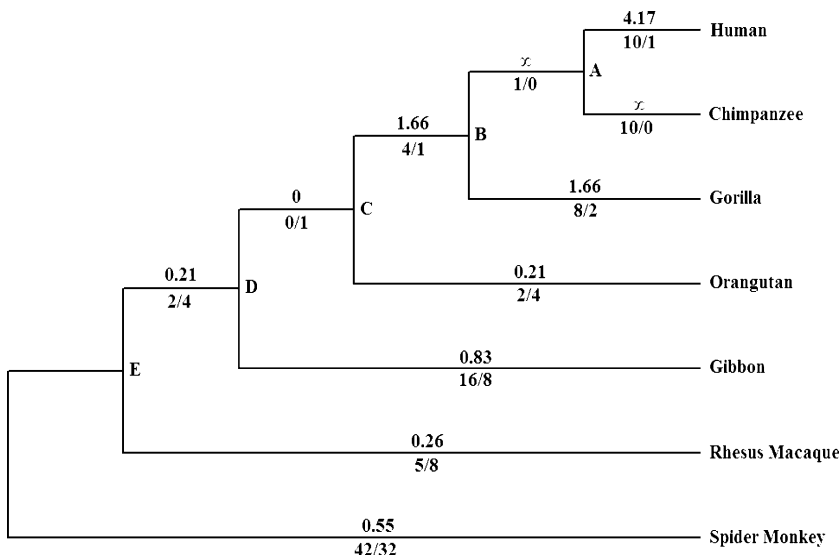


Fig. 2. The ω values and numbers of nonsynonymous (N) and synonymous (S) substitutions of the primate species tested. The ω value is shown above each branch, and the N/S ratio below. A–E refer to internal nodes.

Table 1. Parameter estimates and log-likelihood values under the branch-specific model and the branch-site model

Model	N_p	LnL	Estimates of parameters
A: one ratio	14	-3222.61	$\omega_0 = \omega_C = \omega_H = 0.620$
Branch-specific			
B: Two ratios: $\omega_0 = \omega_C = \omega_G, \omega_H$	15	-3218.84	$\omega_0 = \omega_C = \omega_G = 0.559, \omega_H = \infty$
C: Two ratios: $\omega_0 = \omega_H = \omega_G, \omega_C$	15	-3217.68	$\omega_0 = \omega_H = \omega_G = 0.549, \omega_C = \infty$
D: Two ratios: $\omega_0 = \omega_H = \omega_C, \omega_G$	15	-3221.62	$\omega_0 = \omega_H = \omega_C = 0.586, \omega_G = 2.061$
E: Two ratios: $\omega_C = \omega_H = \omega_G, \omega_0$	15	-3214.2	$\omega_0 = 0.459, \omega_H = \omega_C = \omega_G = 4.809$
F: Two ratios: $\omega_C = \omega_H, \omega_G = \omega_0$	15	-3215.34	$\omega_0 = \omega_G = 0.499, \omega_C = \omega_H = 7.639$
G: Two ratios: $\omega_C = \omega_G, \omega_H = \omega_0$	15	-3218.09	$\omega_0 = \omega_H = 0.521, \omega_C = \omega_G = 3.438$
H: Two ratios: $\omega_G = \omega_H, \omega_C = \omega_0$	15	-3217.78	$\omega_0 = \omega_C = 0.518, \omega_G = \omega_H = 6.293$
I: Two ratios: $\omega_0 = \omega_C = \omega_G, \omega_H = 1$	14	-3221.19	$\omega_0 = \omega_C = \omega_G = 0.562, \omega_H = 1$
J: Two ratios: $\omega_0 = \omega_H = \omega_G, \omega_C = 1$	14	-3220.83	$\omega_0 = \omega_H = \omega_G = 0.551, \omega_C = 1$
K: Two ratios: $\omega_0 = \omega_C = \omega_H, \omega_G = 1$	14	-3221.93	$\omega_0 = \omega_H = \omega_C = 0.587, \omega_G = 1$
L: Two ratios: $\omega_0, \omega_C = \omega_H = \omega_G = 1$	14	-3217.86	$\omega_0 = 0.458, \omega_H = \omega_C = \omega_G = 1$
M: Two ratios: $\omega_G = \omega_0, \omega_C = \omega_H = 1$	14	-3219.05	$\omega_0 = \omega_G = 0.499, \omega_C = \omega_H = 1$
N: Two ratios: $\omega_H = \omega_0, \omega_C = \omega_G = 1$	14	-3219.96	$\omega_0 = \omega_H = 0.518, \omega_C = \omega_G = 1$
O: Two ratios: $\omega_C = \omega_0, \omega_G = \omega_H = 1$	14	-3220.24	$\omega_0 = \omega_C = 0.524, \omega_G = \omega_H = 1$
Site-specific			
M1: neutral ($K=2$)	2	-3216.84	$p_0 = 0.466, (p_1 = 0.534), \omega_0 = 0$
M2: selection ($K=3$)	4	-3210.70	$p_0 = 0.973, p_1 = 0, (p_2 = 0.027), \omega_0 = 0.465, \omega_2 = 8.892$
M7: β	2	-3216.93	$p = 0.018, q = 0.016$
M8: β & $\omega > 1$	4	-3210.70	$p_0 = 0.973, p = 86.079, q = 99, (p_1 = 0.027) \omega = 8.907$
Branch-site			
Model A with $\omega_2 = 1$	3	-3216.10	$p_0 = 0.219, p_1 = 0.215, (p_{2a} + p_{2b} = 0.566), \omega_0 = 0$
Model A	4	-3212.48	$p_0 = 0.258, p_1 = 0.221, (p_{2a} + p_{2b} = 0.521), \omega_0 = 0.045, \omega_2 = 10.947$

Note. N_p , number of parameters in the ω distribution; ω_C , ω in the chimpanzee lineage; ω_G , ω in the gorilla lineage; ω_H , ω in the human lineage; ω_0 , ω in all other lineages; p_0 , frequency of site classes with $0 < \omega < 1$; p_1 , frequency of site classes with $\omega = 1$; p_2 , frequency of site classes with $\omega > 1$.

likelihood approach embodied in the PAML package (Yang 1997) to confirm the results, which averages over all possible ancestral sequences at each interior node in the tree and weights appropriately according to their relative likelihoods of occurrence. The log likelihood values and maximum-likelihood estimates of parameters under the branch-specific models, the site-specific models, and the branch-site models

(Yang 1998; Yang and Nielsen 2002; Zhang et al. 2005) are reported in Table 1. In the branch-specific models, the two-ratio model assigns one ω ratio for the human, chimpanzee, and gorilla branches ($\omega_H = \omega_C = \omega_G$), which were shown to have ω values > 1 , and ω_0 for all other branches. This model is significantly better than the one-ratio model ($2\Delta l = 16.82$; $P = 0.00004$, $df = 1$). When fixing

the foreground ω_C , ω_G , and ω_H to 1 in the two-ratio model, we found that this model was significantly worse than the two-ratio model without constraining $\omega_C = \omega_G = \omega_H = 1$ ($2\Delta l = 7.33$; $P = 0.007$, $df = 1$), indicating that the great ape clade evolved more rapidly than other primate lineages due to higher level of selection. We also conducted a series of tests (Table 1) to identify which branch or branches are significantly > 1 , and the data indicate that both the human and the chimpanzee lineages are likely under positive selection ($P = 0.03$ and $P = 0.01$, respectively). To rule out the possibility that the high ω ratio might result from a decrease in K_s , we compared the average K_s in *NYD-SP12* to the average value of more than 10,000 genes in the genome (Nielsen et al. 2005) and no significant decrease in K_s was observed for *NYD-SP12* (data not shown). Therefore, the high ω values of human and chimpanzee are most likely caused by Darwinian positive selection.

With the use of the site-specific models, we can predict whether there are certain sites under selection. We first compared the M2 model (selection) with the M1 model (neutral), and the former is significantly better than the latter ($2\Delta l = 12.28$; $P = 0.002$, $df = 2$). The M8 model (β and $\omega > 1$) is also much better than the M7 model (β) ($2\Delta l = 12.46$; $P = 0.002$, $df = 2$). Both comparisons imply that there is selection in *NYD-SP12* among the primate lineages.

The modified branch-site models (Zhang et al. 2005) were applied to test whether there are amino acid sites under positive selection along the branches of interest. Only the result from the Bayes empirical Bayes (BEB) method was considered because it performs better than the naïve empirical Bayes (NEB) method by taking into account the sampling errors (Yang et al. 2005). The human and chimpanzee branches were used as the foreground branches. The modified model A (Zhang et al. 2005) is significantly better than that with ω_2 fixed to 1 ($2\Delta l = 7.24$; $P = 0.007$, $df = 1$) (Zhang et al. 2005). The parameter estimates under model A suggested that 52% of the sites are under strong positive selection along the foreground branches with $\omega = 10.95$. At the $P > 95\%$ level, we identified the following sites under positive selection: 35A, 78E, 82F, 97Q, 300E, 335I, 356E, 392D, 459L, 547T, 567Q, and 569R (shaded in Fig. 1).

When conducting McDonald-Kreitman's test, we detected 12 polymorphic sites (8 nonsynonymous and 4 synonymous) in the coding region of *NYD-SP12* in human populations. The interspecies nonsynonymous/synonymous substitution ratio (20/1) is higher than the nonsynonymous/synonymous polymorphism in human populations (8/4) ($p = 0.047$, Fisher's two-tailed exact test), again confirming the presence of Darwinian positive selection between the human and the chimpanzee lineages.

We next sought to identify the specific domains in *NYD-SP12* that might be subject to positive selection because this might reveal the driving force and facilitate further functional analysis. For instance, studies on major histocompatibility complex proteins suggested that positive selection is confined to the antigen-recognition sites (Hughes and Nei 1988; Yang and Swanson 2002), which participate in protein-protein interaction. The *NYD-SP12* contains a known functional domain called TPR (tetratricopeptide repeat; amino acids 169–280), which is involved in chaperone, cell cycle, transcription, and protein transport complexes and may function in protein-protein interaction. Using sliding window analysis, we compared the ω ratios between ancestral node A and human or chimpanzee, respectively. We did not observe signature of positive selection on the TPR domain. Instead, the other segments of the gene, which could not be classified into any known domain, have undergone positive selection (data not shown). This calls for further studies to understand the biological meanings of the observed positive selection on these segments.

NYD-SP12 Sequence Polymorphisms in Human Populations

In many cases, selection increases the degree of differentiation among populations. A selective sweep can have dramatic impacts on the level of population subdivision, especially when the sweep has not spread to all populations (Majewski and Cohan 1999; Nielsen 2005; Slatkin and Wiehe 1998; Yang and Swanson 2002). The allele frequency differences in subpopulations can be evaluated by F_{ST} (Wright 1950). Under evolutionary neutrality, all loci in the genome would have the same expected degree of differentiation, which can be used to detect the action of selection (Cavalli-Sforza 1966; Lewontin and Krakauer 1973). If allele-frequency data are available for a set of putatively neutral loci, then an empirical distribution of F_{ST} values can be constructed and used to identify loci with unusual patterns of differentiation due to positive selection (Cavalli-Sforza et al. 1994; Fullerton et al. 2002). In our population samples, we measured the interpopulation frequency divergences (Table 2), and we conducted the F_{ST} analysis to assess population differentiation. As shown in Table 3, four of the six pairwise comparisons have significant interpopulation differences. Both regional selection and demographic factors, such as population growth and random genetic drift, could account for the observed differentiation.

We conducted several neutrality tests, including Tajima's D test, Fu and Li's D and F tests, and Fay and Wu's H test (data not shown). None of the tests

Table 2. The *NYD-SP12* polymorphic sites in human populations

Location	Property			Derived allele frequency			
	Ancestral	Derived	AA change	African (<i>n</i> = 38)	Caucasian (<i>n</i> = 26)	Asian (<i>n</i> = 60)	Melanesian (<i>n</i> = 20)
149	A	G	N/S	0.158	0.077	0.000	0.000
232	G	A	E/K	0.211	0.038	0.300	0.450
397	G	A	M/V	0.368	0.731	0.700	0.500
440	G	A	G/E	0.026	0.038	0.300	0.100
675	C	T	S	0.289	0.115	0.133	0.150
720	G	A	R	0.000	0.038	0.000	0.000
1029	A	T	V	0.026	0.000	0.000	0.000
1249	C	T	L	0.000	0.000	0.017	0.000
1267	G	A	V/M	0.000	0.000	0.000	0.350
1526	C	T	A/V	0.000	0.077	0.000	0.000
1577	C	T	T/M	0.947	0.923	10.000	10.000
1691	A	G	Q/R	0.000	0.000	0.017	0.000

Note. *n* is the number of chromosomes.

Table 3. Pairwise F_{ST} values in human populations

	Asian	Melanesian	Caucasian	African
Asian	0.000			
Melanesian	0.154**	0.000		
Caucasian	0.0227	0.071**	0.000	
African	0.081*	0.040*	0.004	0.000

Note. * $P < 0.05$; ** $P < 0.01$.

rejected the neutral assumptions. Since our sample size is relatively small (72 individuals), we used the newly available single-nucleotide polymorphism (SNP) data from the International HapMap Project (Altshuler et al. 2005) to confirm our results by empirical studies in a large sample pool (209 individuals) in the *NYD-SP12* region. Beside the tests mentioned above, we also conducted the iHS (integrated haplotype score) test developed to test for recent selection using the HapMap data (Voight et al. 2006). The Haplotter program was used to calculate the iHS (Voight et al. 2006), as well as several other test statistics (Tajima's D , Fay and Wu's H , F_{ST}). Our result indicates that in the *NYD-SP12* region, none of the statistic values fell into the top 5% cutoff (which is defined as significant in empirical distributions; data not shown), implying no or very weak recent positive selection. However, it could also be due to a low density of SNPs in the Phase I HapMap data or to the selective sweep on amino acid changes occurring well before the split of modern human populations.

Functional Association of the *NYD-SP12* Polymorphisms in the Clinical Subjects

As *NYD-SP12* is related to spermatogenesis and shows a higher expression level in adults than in fetuses (Xu et al. 2003), we investigated the SNP

frequency spectrum in three clinical groups (Table 4). Although the majority of the SNPs showed no significant frequency discrepancy among the groups, the 675 C/T synonymous polymorphisms suggest a different pattern. The DON group, which has the highest sperm count and motility level, has the highest 675T frequency, and this is significantly higher than in the other two groups ($P = 0.0006$ for DON versus AST, $P = 0.016$ for DON versus NOR; Fisher's two-tailed exact test after Bonferroni correction). However, we did not observe the same association between AST and NOR.

Discussion

In this study, we have demonstrated that *NYD-SP12* evolves rapidly under positive selection in primates, and a synonymous polymorphism in human populations is associated with sperm traits in the clinical subjects. It was suggested that *NYD-SP12* is likely involved in the formation of acrosome during spermatogenesis (Lu et al. 2006), hence the observed adaptive evolution in human and chimpanzee could cause functional modification in sperm traits.

Genes expressed exclusively or preferentially in testis are likely involved in male reproduction and have been shown to evolve rapidly under positive selection in previous studies (Podlaha and Zhang 2003; Rooney and Zhang 1999; Wang and Zhang 2004). Our observation of rapid evolution in *NYD-SP12* is consistent with the previous findings. There are varied biological driving forces causing adaptive evolution; one possible force is sexual selection because sperm traits are critical for male reproduction. For example, it was shown that the sperm trait-related gene *SEMG2* was involved in sperm competition (Dorus et al. 2004). In the analysis of

Table 4. Allele frequency distribution of the 675T allele in the clinical samples

Group	Sample size	Genotype (CC/CT/TT)	T-allele frequency	Sperm count (millions/ml)	Sperm motility
Asthenozoospermia	86	40/3/0	.035	72.7	A + B = 16.9%
Normal	96	40/8/0	.083	80.4	A + B = 57.4%
Donor	66	19/14/0	.212	248.2	A + B = 74.3%

Note. Sample size refers to the number of chromosomes. Sperm motility was ranked according to the World Health Organization (1999) standard.

NYD-SP12, chimpanzee, the most promiscuous species, showed the strongest signal of positive selection ($\omega = \infty$), and human, whose mating system is considered less promiscuous, also had a strong signature of positive selection ($\omega = 4.20$) (Harcourt et al. 1995). However, the monandrous gorilla also has a relatively high ω ratio ($\omega = 1.67$), while the rhesus monkey has a low ratio ($\omega = 0.26$) though it is highly promiscuous (Anderson and Dixson 2002). Therefore, sperm competition may not be the major force leading to the rapid evolution of *NYD-SP12* in primates. An alternative hypothesis is that the adaptive evolution might lead to establishing reproductive barriers during speciation, which seems to fit the *NYD-SP12* data better (Yang and Swanson 2002), but more evidence is needed.

A previous study showed that the *NYD-SP12* protein was localized to the Golgi apparatus, and it was involved in the formation of the acrosome during spermatogenesis (Xu et al. 2003). This is consistent with the potential function of the TPR domain of *NYD-SP12*, and it is not surprising to see strong purifying selection on the TPR region. However, the biological significance of the observed positive selection is not clear. One possibility is that *NYD-SP12* gained new or modified function along the human and chimpanzee lineages. Further functional evidence is needed to address this issue.

The neutrality test for human populations does not show significant deviation from the neutral expectation, though the F_{ST} comparisons show significant interpopulation divergence. Considering the small sample size, we also analyzed the newly available HapMap data, in which none of the statistics was exceptional high. Therefore, the significant differences in F_{ST} may be caused by demographic factors though regional selection as well as the ascertainment bias in the HapMap data cannot be totally ruled out.

The significant association of 675T allele frequency in the three clinical groups is intriguing. Since the C-to-T mutation at this site is a synonymous change, it seems unlikely that this polymorphism will encounter selection at the protein level. To test whether this site has any effect at gene expression pattern, we analyzed the expression of *NYD-SP12* in eight human testis samples with different genotypes, i.e.,

675 C/C and 675 C/T. No significant difference was observed (data not shown). Hence, the 675 C/T polymorphism may be a nonfunctional site which is linked to a nearby functional site due to genetic hitchhiking. To locate the potential functional site, we conducted linkage disequilibrium analysis in this gene by using SNP data downloading from the HapMap project. The 675 C/T polymorphism falls into a region showing high linkage disequilibrium, which spans about 30 kb covering exon 3 and intron 3, implying that there might be other sites with clinical significance. However, due to the low density of SNPs in the Phase 1 HapMap data, further re-sequencing and analysis of this region are needed.

Electronic Database Information

The URL for data in this article is <http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>. The GenBank accession numbers are DQ141860–DQ142639.

Acknowledgments. This study was supported by grants from the Chinese Academy of Sciences (KSCX1-YW-R-34), the National Natural Science Foundation of China (30370755, 30525028, 30630013), the Natural Science Foundation of Yunnan Province of China, and the National 973 Project of China (2006CB701506). We thank Hui Zhang and Yi-chuan Yu for technical help.

References

- Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. *Nature* 437:1299–1320
- Anderson MJ, Dixson AF (2002) Sperm competition: motility and the midpiece in primates. *Nature* 416:496
- Cavalli-Sforza LL (1966) Population structure and human evolution. *Proc R Soc Lond B Biol Sci* 164:362–379
- Cavalli-Sforza L, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ
- Civetta A, Singh RS (1995) High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J Mol Evol* 41:1085–1095
- Comeron JM (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J Mol Evol* 41:1152–1159

- Cameron JM (1999) K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* 15(9):763–764
- 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
- Fay JC, Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413
- Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133:693–709
- Fullerton SM, Bartoszewicz A, Ybazeta G, Horikawa Y, Bell GI, Kidd KK, Cox NJ, Hudson RR, Di Rienzo A (2002) Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. *Am J Hum Genet* 70:1096–1106
- Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, Gunnell G, Groves CP (1998) Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* 9:585–598
- Harcourt AH, Purvis A, Liles L (1995) Sperm competition: mating system, not breeding season, affects testes size of primates. *Funct Ecol* 9:468–476
- Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335:167–170
- Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5:150–163
- Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74:175–195
- Lu L, Lin M, Xu M, Zhou ZM, Sha JH (2006) Gene functional research using polyethylenimine-mediated in vivo gene transfection into mouse spermatogenic cells. *Asian J Androl* 8:53–59
- Majewski J, Cohan FM (1999) Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* 152:1459–1474
- McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654
- Messier W, Stewart CB (1997) Episodic adaptive evolution of primate lysozymes. *Nature* 385:151–154
- Nielsen R (2005) Molecular signatures of natural selection. *Annu Rev Genet* 39:197–218
- Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, Adams MD, Cargill M (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3:e170
- Page SL, Goodman M (2001) Catarrhine phylogeny: noncoding DNA evidence for a diphyletic origin of the mangabeys and for a human-chimpanzee clade. *Mol Phylogenet Evol* 18:14–25
- Podlaha O, Zhang J (2003) Positive selection on protein-length in the evolution of a primate sperm ion channel. *Proc Natl Acad Sci USA* 100:12241–12246
- Rooney AP, Zhang J (1999) Rapid evolution of a primate sperm protein: Relaxation of functional constraint or positive Darwinian selection? *Mol Biol Evol* 16:706–710
- Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 15:174–175
- Slatkin M, Wiehe T (1998) Genetic hitch-hiking in a subdivided population. *Genet Res* 71:155–160
- Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. *Nat Rev Genet* 3:137–144
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595
- Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4:e72
- Wang X, Zhang J (2004) Rapid evolution of mammalian X-linked testis-expressed homeobox genes. *Genetics* 167:879–888
- World Health Organization(1999) WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction. Cambridge University Press, Cambridge,
- Wright S (1950) Genetical structure of populations. *Nature* 166:247–249
- Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309
- Xu M, Xiao J, Chen J, Li J, Yin L, Zhu H, Zhou Z, Sha J (2003) Identification and characterization of a novel human testis-specific Golgi protein, NYD-SP12. *Mol Hum Reprod* 9:9–17
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556
- Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15:568–573
- Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917
- Yang Z, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* 19:49–57
- Yang Z, Wong WS, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118
- Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472–2479