

ORIGINAL ARTICLE

# Examination of disease-based selection, demographic history and population structure in European Y-chromosome haplogroup I

Efe Sezgin<sup>1</sup>, Alyssa Drosdak<sup>1</sup>, Carl McIntosh<sup>1</sup>, Bailey Kessing<sup>2</sup>, James A Lautenberger<sup>1</sup>, James J Goedert<sup>3</sup>, John P Phair<sup>4</sup>, Jennifer L Troyer<sup>2</sup>, Michael W Smith<sup>5</sup> and Stephen J O'Brien<sup>1</sup>

We attempted to refine the understanding of an association of Y-chromosomal haplogroup I (hg-I) with enhanced AIDS progression that had been previously reported. First, we compared the progression phenotype between hg-I and its phylogenetically closest haplogroup J. Then, we took a candidate gene approach resequencing *DDX3Y*, a crucial autoimmunity gene, in hg-I and other common European Y-chromosome haplogroups looking for functional variants. We extended the genetic analyses to *CD24L4* and compared and contrasted the roles of disease-based selection, demographic history and population structure shaping the contemporary genetic landscape of hg-I chromosomes. Our results confirmed and refined the AIDS progression signal to hg-I, though no gene variant was identified that can explain the disease association. Molecular evolutionary and genetic analyses of the examined loci suggested a unique evolutionary history in hg-I, probably shaped by complex interactions of selection, demographic history and high geographical differentiation leading to the formation of distinct hg-I subhaplogroups that today are associated with HIV/AIDS onset. Clearly, further studies on Y-chromosome candidate loci sequencing to discover functional variants and discern the roles of evolutionary factors are warranted.

*Journal of Human Genetics* (2010) 55, 613–620; doi:10.1038/jhg.2010.77; published online 24 June 2010

**Keywords:** AIDS progression; *CD24L4/DDX3Y*; population growth; population structure; selection; Y chromosome

## INTRODUCTION

The importance of Y chromosome in human health is becoming more evident with the accumulation of data from disease association studies. The nonrecombining region of the human Y chromosome (NRY) makes up 95% of the total Y chromosome and contains over 78 genes involved in a range of developmental and physiological processes such as sex determination, spermatogenesis, stature formation, enamel formation, immunological responses and behavioral traits.<sup>1,2</sup> Several slowly mutating bi-allelic polymorphisms on NRY can be used to group human Y chromosomes into monophyletic lineages called Y-chromosome haplogroups that have been essential markers for investigating human population histories.<sup>1</sup> Mutations on the Y chromosome ranging from single-nucleotide polymorphisms (SNPs) to large deletions and inversions have been linked to several diseases such as infertility and cancers.<sup>3</sup> Recently, our group showed faster AIDS progression associated with a European-specific Y-chromosome haplogroup, the haplogroup I (hg-I), in European American patients.<sup>4</sup> However, the same features of the Y chromosome that make it an excellent marker for population structure/history

make it rather difficult to identify causative variants in the Y chromosome associated with disease.

The lack of recombination in the NRY makes mapping of a locus/loci responsible for disease, such as the AIDS progression signal, nearly impossible with traditional approaches, because all genes can be positional candidate genes. In addition, the Y chromosome contains large sections of low complexity, repetitive sequences, making resequencing efforts to detect causative alleles challenging. Therefore, to further investigate the human Y chromosome in association studies, alternative evolutionary and population genetic analyses have been proposed to determine potential disease associations of specific Y-chromosome haplogroups rather than specific Y-chromosome loci.<sup>5</sup>

As for any fitness characteristics, Y-chromosome-specific phenotypic variations can be subject to selection. For example, infertility will be subjected to strong purifying selection, whereas weaker selection may be acting on more subtle phenotypic differences, such as the immune response, over a period of time under changing environmental conditions. These selective pressures can leave DNA polymorphism signatures detectable by population genetic analyses.

<sup>1</sup>Laboratory of Genomic Diversity, National Cancer Institute at Frederick, Frederick, MD, USA; <sup>2</sup>Laboratory of Genomic Diversity, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD, USA; <sup>3</sup>Viral Epidemiology Branch, National Cancer Institute, Rockville, MD, USA; <sup>4</sup>Northwestern University Medical School, Comprehensive AIDS Center, Chicago, IL, USA and <sup>5</sup>Advanced Technology Program, SAIC-Frederick, Frederick, MD, USA  
Correspondence: Dr E Sezgin, Laboratory of Genomic Diversity, National Cancer Institute at Frederick, Building 560, Room 21-42, NCI-Frederick, Frederick, MD 21702-1201, USA. E-mail: sezginef@mail.nih.gov

Received 24 January 2010; revised 28 May 2010; accepted 31 May 2010; published online 24 June 2010

However, demographic history and population structure will also effect the observed DNA sequence variation and must be considered, so that they do not confound potential disease-based selection signatures.

In this study, we aimed to refine our understanding of AIDS susceptibility associated with hg-I. We first analyzed AIDS progression differences between hg-I subhaplogroups and the phylogenetically closest haplogroup, haplogroup J (hg-J). Previously, we reported faster progression to AIDS outcomes in hg-I compared to the most common European haplogroup R, but did not analyze hg-J. As hg-I and hg-J are sister clades with similar time depths, and likely to have relatively closely related sequence organizations, it is interesting to see if the disease susceptibility is unique to hg-I or common among the hg-IJ clades. Once we clearly identified AIDS susceptibility to be hg-I specific, we adopted a candidate gene approach and resequenced a single-copy, X-degenerate, nonrecombining Y-chromosome locus, *DDX3Y*, a crucial gene in human autoimmunity, to identify potential functional variants that may influence immune response. The X chromosome paralog of *DDX3Y*, *DDX3X*, has been shown to be involved in HIV-1 replication and antiviral activity.<sup>6,7</sup> We also focused on detecting signals of potential selective pressure on the hg-I Y chromosomes that may indicate that the natural history of this haplogroup was influenced by infectious disease susceptibility/resistance. In addition, we examined the potential effects of historical population size changes and contemporary population structure on the observed distribution of DNA sequence polymorphism on the Y chromosome. For population genetics and molecular evolutionary analyses, besides *DDX3Y*, we chose another single-copy, nonrecombining Y-chromosome locus with a distinct evolutionary history, *CD24L4*, an intronless pseudogene of a functional autosomal *CD24* gene (on chromosome 6) that retrotransposed after the human chimpanzee lineage split.<sup>8</sup> We expect fewer evolutionary constraints

on pseudogenes. Therefore, the selective forces on *CD24L4* and *DDX3Y* are expected to be different, acting as an internal control to allow for differentiation of selective and demographic inferences in our samples.

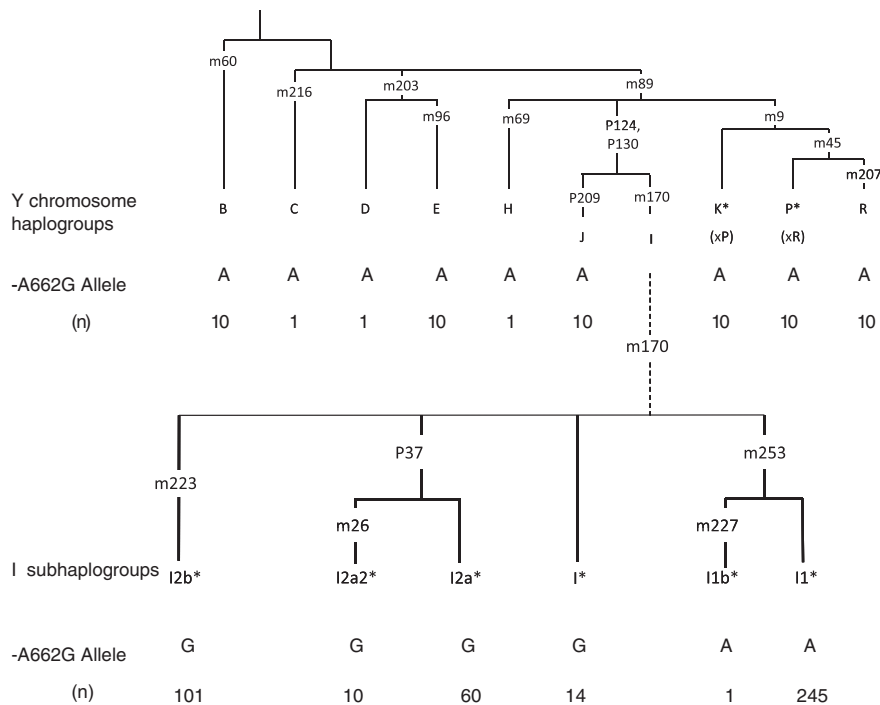
Although we did not identify a causal variant that can explain AIDS association in hg-I, based on interpreting AIDS survival associations in the context of population genetic analyses, we detected evidence for historical purifying selection, population expansion and substructure that have shaped the observed genetic diversity in hg-I Y chromosomes.

## MATERIALS AND METHODS

### Study subjects and assignment of Y-chromosomal haplogroups

The origin of individual samples, DNA extraction and assignment to specific Y-chromosome haplogroups were described by Sezgin *et al.*<sup>4</sup> The Y-chromosome haplotype of each male was determined by typing 10 SNPs (M60, M216, M203, M96, M89, M69, M170, M9, M45 and M207) located on the NRY using the 5' nuclease assay (TaqMan Assay-by-Design SNP genotyping products; Applied Biosystems, Foster City, CA, USA). The samples that were combined to form the juxtaposed F\* (xI, H, K) group in the previous study were further genotyped in this study for P130 (rs17250887), P124(rs17315772) and P209 (rs17315835) and assigned to respective hg-J and hg-I following the Y chromosome consortium (YCC 2008) nomenclature.<sup>9</sup> The hg-I samples were genotyped for five additional markers, M253, M227, P37, M26, and M223, to define the I1\*, I2a\*, I2a2\* and I2b\* subhaplogroups (the corresponding YCC 2002 names are I1a, I1a4, I1b, I1b2 and I1c, respectively). The remaining hg-I samples that lacked the derived alleles at these five markers were assigned into the I\* paragroup (Figure 1).

The *CD24L4* and *DDX3Y* sequencing panels consisted of 90 European Americans: 10 samples from each of the haplogroups E, J, K\*, R and subhaplogroups I1\*, I2a\*, I2a2\*, I2b\* and I\*. Further genotyping of the *CD24L4* -A662G polymorphism included additional 484 samples: 431 hg-I; 10 samples from each of the haplogroups R, P(xR), J, E and B; and 1 sample



**Figure 1** Presented is a Y-chromosome haplogroup tree with the indicated haplogroup defining markers following the Y-chromosome consortium 2008 nomenclature. Indicated below is number of individuals (*n*) genotyped for the -A662G allele that was discovered upon sequencing of *CD24L4* region. The AIDS susceptible haplogroup I samples were further divided into the respective subhaplogroups.

from each of the C, D and H haplogroups (Figure 1). This study was approved by the Protocol Review Office of the institutional review board of the National Cancer Institute. Informed consent was obtained from all individuals.

### AIDS progression analyses

Progression to four AIDS-related outcomes (time to CD4+ T lymphocyte count of fewer than 200 per mm<sup>3</sup> (CD4 < 200), AIDS-1993 definition, AIDS-1987 definition and death from AIDS-defining illnesses before HAART use was compared in J and hg-I subhaplogroup seroconverter (date of HIV-1 infection known) subjects using Cox proportional hazard model and Kaplan–Meier analyses as described in Sezgin *et al.*<sup>4</sup> The Cox models were adjusted for the known autosomal AIDS restriction/susceptibility genes (*CCR5-Δ32*, *CCR2-64I*, *CCR5-P1/P1*, *HLA-B\*27*, *HLA-B\*57*, *HLA-B\*35*, *HLA-B\*35P<sub>x</sub>*, *KIR3DS1*, *TSG101*, *SDF*, *RANTES* and HLA class I heterozygosity) using a genetic propensity index.<sup>10</sup> Statistical analyses were performed with SAS 9.1 (SAS Institute, Cary, NC, USA). *P*-values reported were nominal and uncorrected for multiple comparisons.

### Autosomal marker-based population stratification/substructure discovery

700k autosomal SNP data from an Affymetrix (Affymetrix, Santa Clara, CA, USA) 6.0 Genechip platform with call rates ≥ 95% was available (Troyer *et al.*, unpublished data) to perform a principal components analysis (PCA) by EIGENSOFT package.<sup>11</sup> The distribution of samples belonging to hg-I and hg-J was compared along the top two eigenvectors by an analysis of variance *F* statistic. Moreover, the eigenvectors were used in AIDS association Cox models as covariates to adjust for potential autosomal stratification.

### DNA sequencing and polymorphism screening

A DNA fragment of 6795 base pairs (bp) covering 3500 bp upstream of putative *CD24L4* mRNA, 2195 bp putative *CD24L4* mRNA and 1100 bp downstream regions was selected according to NCBI Reference Sequence (RefSeq) NC\_000024.8 and accession number AC009235. This region was divided into six overlapping amplicons (Supplementary Figure 1) and sequenced with primer pairs listed in Supplementary Table 1. Newly discovered *CD24L4* – A662G polymorphism was screened in a larger set of samples using the 5' nuclease assay (TaqMan Assay-by-Design SNP genotyping products; Applied Biosystems; Supplementary Table 2). For *DDX3Y* (formerly known as *DBY*), we selected a DNA fragment of 22613 bp covering 3570 bp upstream of first exon, 15613 bp mRNA (longest transcript covering all 17 exons) and 3430 bp downstream regions according to NCBI RefSeq NC\_000024.8 and RefSeq mRNA NM\_001122665. The *DDX3Y* region was divided into 11 overlapping amplicons (Supplementary Figure 1) and sequenced with primer pairs listed in Supplementary Table 3. A 1900-bp low-complexity region between the third and fourth amplicons was not sequenced. The Repeat Masker web server (<http://www.repeatmasker.org>) was used to identify the repetitive elements around and within the sequenced regions. As both *CD24L4* and *DDX3Y* have highly similar homologs in the genome (*CD24* on chromosome 6 and *DDX3X* on the X chromosome), female samples were used as a control that the primer pairs only amplified the Y-chromosome copies of these genes. All DNA

fragments were Sanger sequenced and analyzed by ABI (Foster City, CA, USA) 3730xl DNA Analyzer. Sequence fragments were assembled and aligned with computer program Sequencher, version 4.6 (GeneCodes, Ann Arbor, MI, USA) and MEGA version 4.0.<sup>12</sup> The nucleotide sequence data reported are available in the GenBank databases.

The NCBI, Seattle SNPs and ENCODE web sites were searched for previously reported *CD24L4* and *DDX3Y* mutations. Also, the 236 resequenced (accession codes BV678971–BV679207) PCR fragments by Repping *et al.*<sup>13</sup> were aligned and examined for *DDX3Y* region nucleotide changes.

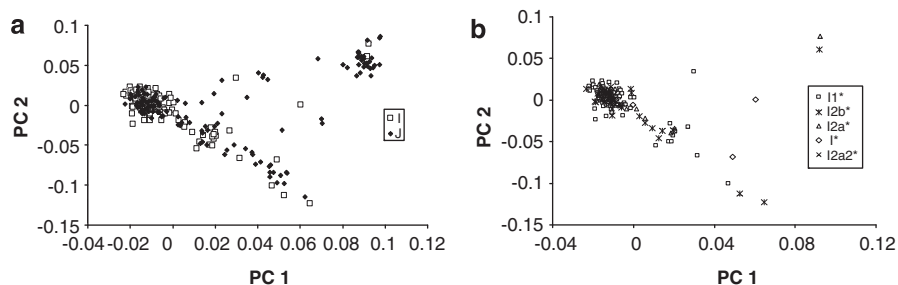
### Summary statistics and population genetic analyses

Two estimators, one based on the number of segregating sites in a sample,  $\theta_w$ ,<sup>14</sup> and another based on average number of pairwise nucleotide differences per site between two sequences,  $\pi$ ,<sup>15</sup> were used to summarize the nucleotide diversity statistics. To test the neutrality of the polymorphism frequency distribution and demographic equilibrium, we applied Tajima's *D*<sup>16</sup> and Fu and Li's *D*\*<sup>17</sup> tests. Further inferences of population size changes were tested using Fu's *F*<sub>s</sub>,<sup>18</sup> Ramos-Onsins and Rozas' *R*<sub>2</sub><sup>19</sup> and mismatch distributions.<sup>20</sup> The significance of *D*, *D*\*, *F*<sub>s</sub> and *R*<sub>2</sub> was inferred from the null distributions generated by 10 000 replicates of coalescent simulations without recombination and should be interpreted as the probability of obtaining values equal or lower than the observed values. Parameter estimates and statistical tests were calculated by DNAsp version 4.50.3.<sup>21</sup> The sum of square deviations between the observed and expected mismatch distributions,<sup>22</sup> and Chakraborty's test of population amalgamation<sup>23</sup> were calculated using ARLEQUIN version 3.11.<sup>24</sup> A potential role for sampling bias influencing the *D*\* and *D* estimates was investigated by sampling hg-I subhaplogroups and constructing 5, 10, 8, 5 and 1 sample populations made up of 1 (individual I subhaplogroups), 2, 3, 4 and 5 (all hg-I subhaplogroups) subhaplogroups, respectively. The *D*\* and *D* for each constructed sample population were estimated as mentioned above.

## RESULTS

### Autosomal and Y haplogroup population substructure/stratification in the study population

As human Y-chromosome variation shows high levels of geographic differentiation, disease association studies involving the Y chromosome are particularly prone to population stratification/substructure. Moreover, in European populations the gene variants associated with susceptibility to/protection against AIDS may also be susceptible to population structure. For example, the AIDS protective *CCR5-Δ32* variant shows a north–south gradient in European populations,<sup>25</sup> a trend similar to that of hg-I distribution in Europe.<sup>26</sup> Moreover, the geographical distributions of hg-J and hg-I are also different in Europe, where the hg-J is mostly found in southern European populations.<sup>27</sup> To prevent autosomal population stratification from confounding Y-chromosome-based AIDS association signal, we applied a PCA approach to 700k autosomal SNP data and observed that the distribution of hg-I and hg-J samples were significantly



**Figure 2** Distribution of individuals belonging to (a) haplogroup I, haplogroup J and (b) haplogroup I subhaplogroups among the top principal components (PCs) based on 700k autosomal SNP data.

**Table 1** Comparison of progression to AIDS outcomes among hg-J and hg-I subhaplogroup seroconverters

(Sub)haplogroup <sup>a</sup>	CD4 < 200			AIDS 1993			AIDS 1987		
	n/events	RH <sup>a</sup> (95% CI)	P	n/events	RH <sup>a</sup> (95% CI)	P	n/events	RH <sup>a</sup> (95% CI)	P
J	97/48	1.00	—	97/49	1.00	—	98/47	1.00	—
I1*	55/30	1.95 (1.00–3.81)	0.05	55/30	1.83 (0.96–3.48)	0.07	55/29	1.62 (0.84–3.13)	0.15
I2a*	12/6	1.97 (0.71–5.48)	0.19	12/6	1.74 (0.64–4.78)	0.28	12/5	1.97 (0.65–5.96)	0.22
I2b*	23/11	2.85 (1.21–6.70)	0.02	23/11	2.56 (1.11–5.93)	0.03	24/10	2.13 (0.93–4.89)	0.07
I*	4/1	1.70 (0.22–13.16)	0.61	4/2	3.21 (0.71–14.50)	0.36	4/3	4.96 (1.30–18.91)	0.02
I (all)	97/48	2.07 (1.12–3.81)	0.02	97/49	2.02 (1.12–3.64)	0.02	98/47	2.02 (1.11–3.68)	0.02
<i>Within hg-I subhaplogroups</i>									
I1*	55/30	0.74 (0.37–1.48)	0.40	55/30	0.77 (0.39–1.49)	0.43	55/29	0.71 (0.36–1.39)	0.32
I2a*	12/6	0.98 (0.38–2.54)	0.97	12/6	0.88 (0.34–2.27)	0.79	12/5	1.06 (0.37–3.02)	0.91
I2b*	23/11	1.60 (0.74–3.46)	0.23	23/11	1.44 (0.67–3.10)	0.35	24/10	1.28 (0.60–2.73)	0.53
I*	4/1	0.85 (0.09–8.27)	0.89	4/2	2.30 (0.39–13.62)	0.36	4/3	3.69 (0.52–26.21)	0.19

<sup>a</sup>Haplogroup J is the reference group. For within hg-I subhaplogroup comparisons, the reference group is all other subhaplogroups combined against the subhaplogroup of interest. Due to small sample size ( $n=3$ ), there were no events in the I2a2\* group and similar analyses were not conducted.

different along the first axis ( $P=0.001$ ; Figure 2). Moreover, when only hg-I subhaplogroups were compared against each other, I\* differed significantly compared to other subhaplogroups along the first axis ( $P=0.02$ ; Figure 2). Although the recovered eigen axes were informative about the genomic ancestry differences between the (sub)haplogroups, they were not significantly correlated with the AIDS progression outcome phenotypes (that is, for AIDS-1987, eigen 1:  $r^2=0.02$ ,  $P=0.09$ ; eigen 2:  $r^2=0.01$ ,  $P=0.15$ ) upon a regression analysis. For completeness, we still corrected the Cox models of AIDS progression analyses for these top two most informative eigen vectors.

#### AIDS progression analyses

There were significant differences between the hg-I subhaplogroups and hg-J samples for AIDS progression, where hg-I subhaplogroups depleted CD4+ T cells and progressed to AIDS 1993 and AIDS 1987 faster (Table 1). However, the faster AIDS progression signal (compared to the hg-J samples) was not significant for every individual hg-I subhaplogroup, probably due to small sample sizes resulting in lack of statistical power. When all hg-I subhaplogroups were combined (representing hg-I) and compared against hg-J, the significant AIDS progression differences were more evident ( $P \leq 0.02$  for each AIDS end point; Table 1).

In contrast, when the hg-I subhaplogroup samples were compared against each other, the Cox analyses did not show a significant AIDS progression differences between them (Table 1).

#### Sequencing to identify functional variants and population genetic analyses of *DDX3Y*

No mutation in *DDX3Y* corresponded with AIDS outcome. However, the number of polymorphic sites in and around this gene allowed for population genetic analyses within and between Y haplogroups. There were 14 polymorphic sites defining 13 haplotypes in the full set of 90 European American samples selected for DNA sequencing (Figure 3). No polymorphism was observed in the coding regions. Only two of the polymorphisms, –2303 deletion (–T2303D) and –A1850G were in the low complexity/repeat regions and all segregating sites mutated only once, thus the data fit the infinite sites model quite well. The deletion site was excluded from further analyses. The summary statistics of polymorphisms in the 15 003 bp of noncoding region are presented in Table 2. The estimates of nucleotide diversity, negative

$D$  result and the lack of polymorphism in the coding region are in agreement with previous reports.<sup>28–30</sup> Five of the polymorphisms T2766G, G4675A, A9577C, G10682T and C13905T correspond to the binary NRY markers M191, P170, M173, M201 and M198, respectively. The M201 marker was used to assign three of the samples to the haplogroup G.

#### Excess of singleton mutations in the hg-I samples: selection or population expansion

Analyses of all samples ( $n=90$ ) showed that six of the polymorphisms were singletons driving negative but not significant  $D$  and  $D^*$  results. However, the more powerful test for detecting an excess of rare alleles,  $F_s$ , showed that the number of singletons observed was significantly higher than neutral expectations, suggesting either positive selection or population growth. The samples were then assigned to their corresponding haplogroups and hg-I samples ( $n=50$ ) were compared against non-hg-I ( $n=40$ ) samples. Although both groups of samples have the same number of segregating sites ( $S=7$ ) generating seven haplotypes, in the hg-I samples most of the polymorphisms were singletons (5/7) scaling up to five times more singletons compared to the non-hg-I samples (1/7). The excesses of low-frequency polymorphisms in the hg-I samples were also detected by the rather negative and significant  $D$ ,  $D^*$  and  $F_s$  values (Table 2), suggesting either positive selection or a population history out of demographic equilibrium. The conservative mismatch distribution based tests also did not reject the null hypothesis of population growth for any of the haplogroups including hg-I (Table 2). Interestingly, the unimodality of mismatch distributions was much more evident in the non-hg-I samples (Supplementary Figure 2), suggesting a larger population expansion compared to hg-I samples. The small and significant  $R_2$  test value in the hg-I samples was also consistent with a recent population growth. Further examination of each I subhaplogroup showed that, except the I\*, all others had negative but not significant  $D$ ,  $D^*$  and  $F_s$  values. The  $R_2$  is a superior test for detecting population growth when the sample sizes and the number of segregating sites are small. However, again, none of the individual hg-I indicated population growth upon the  $R_2$  test. The  $D$ ,  $D^*$  and  $F_s$  estimates for E, and R samples were also not significantly different from zero and population demographic equilibrium cannot be rejected by the  $R_2$  test. Similar analyses were not conducted in the G, J and K samples as there were no segregating sites. The I\* samples, with positive  $D$  value, showed the

	---	11	1111		---	11	1111
	21	22349903	5666		21	22349903	5666
	38327660569	3178			38327660569	3178	
Y haplogroup samples	05346079780	6876		Y haplogroup samples	05346079780	6876	
	30156458725	8055			30156458725	8055	
Chimp	TACTTTGAAAC	CAAC			TACTTTGAAAC	CAAC	
E_1	...C..A..G.	....		I2a*_7	.....C...G.	....	
E_2	...G.A..G.	....		I2a*_8	.....G.	....	
E_3	...C..A..G.	....		I2a*_9	.....G.	....	
E_4	...C..A..G.	....		I2a*_10	.....G.	....	
E_5	...C..A..G.	....		I2a2*_1	.....G.	....	
E_6	...C..A..G.	.G..		I2a2*_2	.....G.	....	
E_7	...G.A..G.	....		I2a2*_3	.....G.	....	
E_8	...C..A..G.	....		I2a2*_4	.....G.	....	
E_9	...C..A..G.	....		I2a2*_5	.....G.	....	
E_10	...C..A..G.	....		I2a2*_6	.....G.	....	
G_1	.....T.	....		I2a2*_7	.....G.	....	
G_2	.....T.	....		I2a2*_8	.....G.	....	
G_3	.....T.	....		I2a2*_9	.....G.	....	
J_1	.....G.	....		I2a2*_10	.....G.	....	
J_2	.....G.	....		I2b*_1	.....G.	....	
J_3	.....G.	....		I2b*_2	.....G.	....	
J_4	.....G.	....		I2b*_3	.....G.	....	
J_5	.....G.	....		I2b*_4	.....G.	....	
J_6	.....G.	....		I2b*_5	.....G.	....	
J_7	.....G.	....		I2b*_6	.....G.	....	
I*_1	.....G.	T.G.		I2b*_7	.....G.	....	
I*_2	.....G.	T.G.		I2b*_8	.....G.	....	
I*_3	.....G.	....		I2b*_9	.....G.	....	
I*_4	.....G.	....		I2b*_10	.....T.G.	....	
I*_5	.....G.	T.G.		R_1	.....CGT	....	
I*_6	.....G.	...T		R_2	.....CG.	....	
I*_7	.....G.	T.G.		R_3	.....CG.	....	
I*_8	.....G.	T.G.		R_4	.....CG.	....	
I*_9	.....G.	....		R_5	.....CG.	....	
I*_10	..A.....G.	T.G.		R_6	.....CGT	....	
I1*_1	.....G.	....		R_7	.....CG.	....	
I1*_2	.....G.	....		R_8	.....CG.	....	
I1*_3	.....G.	....		R_9	.....CG.	....	
I1*_4	.G.....G.	....		R_10	.....CG.	....	
I1*_5	.....G.	....		K (xP)_1	.....G.	....	
I1*_6	.....G.	....		K* (xP)_2	D.....G.	....	
I1*_7	.....G.	....		K* (xP)_3	.....G.	....	
I1*_8	.....G.	....		K* (xP)_4	.....G.	....	
I1*_9	.....G.	....		K* (xP)_5	.....G.	....	
I1*_10	.....G.	....		K* (xP)_6	.....G.	....	
I2a*_1	.....G.	....		K* (xP)_7	.....G.	....	
I2a*_2	.....G.	....		K* (xP)_8	.....G.	....	
I2a*_3	.....G.	....		K* (xP)_9	.....G.	....	
I2a*_4	.....G.	....		K (xP)_10	.....G.	....	
I2a*_5	.....G.	....					
I2a*_6	.....G.	....					

**Figure 3** Distribution of *DDX3Y* polymorphisms among 90 individuals belonging to 6 Y-chromosome haplogroups and 5 hg-I subhaplogroups (see Figure 1). The numbers indicate the position of changes. Negative base numbers are with respect to the start position of the first exon. None of the changes are in the exons. Chimpanzee (Chimp) sequence is used to show the ancestral states of the observed polymorphisms in human *DDX3Y* sequences.

highest haplotype diversity in the hg-I samples, where two distinct haplotypes characterized by C15368T and A16775G were evident (Table 2 and Figure 3).

We also calculated *D* and *D\** values for *DDX3X* (X chromosome homolog of *DDX3Y*) and *CD24* (chromosome 6 homolog of *CD24L4*) using the HapMap data from 60 CEU samples. The original *DDX3X* variation data based on 18 segregating sites generated significantly positive *D* (2.36,  $P < 0.05$ ) and *D\** (1.68,  $P < 0.05$ ) values. However, it is known that the HapMap genotype data is biased toward

high-frequency polymorphisms and excludes singleton changes (<http://www.hapmap.org/>). After applying a correction for the ascertainment bias in the HapMap data generation,<sup>31</sup> the *D* and *D\** values were not significantly different from neutral expectation of zero.

Next, we examined if sampling from nonhomogeneous populations, in addition to population growth, can bias the *D* and *D\** results. First, we tested if our samples originated from a single homogeneous population following approach by Chakraborty<sup>23</sup> and compared the observed number of alleles against their expected values given



**Table 2** Summary statistics for the 15 003 bp noncoding region of *DDX3Y*

Samples	N <sup>a</sup>	S <sup>b</sup>	$\pi$ (%)	$\theta_w$ (%)	$h_d^c$	$ns^d$	Tajima's D	Fu-Li's D*	Fu's Fs	R <sub>2</sub>	SSD P <sup>e</sup>
All	90	14	0.010	0.018	0.60	6	-1.260	-1.781	-4.915 <sup>#</sup>	0.052	0.33
Iall <sup>f</sup>	50	7	0.004	0.010	0.36	5	-1.577 <sup>#</sup>	-2.651 <sup>#</sup>	-3.710 <sup>#</sup>	0.053 <sup>#</sup>	0.20
Non-hg-I <sup>f</sup>	40	7	0.012	0.011	0.77	1	0.212	0.512	-0.388	0.125	0.15
E	10	3	0.006	0.007	0.51	1	-0.507	0.174	0.300	0.155	0.37
G	3	0	0.000	0.000	0.00	0					
J	7	0	0.000	0.000	0.29	0					
I1*	10	1	0.001	0.002	0.20	1	-1.111	-1.243	-0.339	0.300	0.09
I2a*	10	1	0.001	0.002	0.20	1	-1.111	-1.243	-0.339	0.300	0.12
I2a2*	10	0	0.000	0.000	0.00	0					
I2b*	10	1	0.001	0.002	0.20	1	-1.111	-1.243	-0.339	0.300	0.10
I*	10	4	0.010	0.009	0.71	2	0.143	-0.338	-0.077	0.167	0.46
R	10	1	0.002	0.002	0.36	0	0.015	0.804	0.417	0.178	0.46
K*(xP)	10	1 <sup>g</sup>	0.000	0.000	0.00	0 <sup>g</sup>					

<sup>a</sup>Number of individuals.<sup>b</sup>Number of polymorphic sites.<sup>c</sup>Haplotype diversity.<sup>d</sup>Number of singletons.<sup>e</sup>Probability of sum of square deviations (mismatch distribution).<sup>f</sup>Iall: all haplogroup I samples analyzed together; non-hg-I: E, G, J, R and K\* haplogroup samples analyzed together.<sup>g</sup>The -T2303deletion site was excluded in the following statistical analyses.<sup>#</sup> $P < 0.05$ .

the observed homozygosity in our samples. For all samples ( $n=90$ ), the observed number of alleles ( $k=14$ ) was significantly higher than the expected number ( $k=6.8$ ,  $P=0.0003$ ) indicating nonhomogeneous population sampling. When hg-I ( $n=50$ ) and other (non-hg-I,  $n=40$ ) haplogroup samples are examined separately, the observed number of alleles ( $k=7$ ) was significantly higher than the expected number ( $k=3.4$ ,  $P=0.004$ ) for the hg-I, but not significantly different for the non-hg-I samples (8 versus 6.2,  $P=0.63$ ). The results were not significant for individual I subhaplogroups and haplogroups E, G, J, R and K\*. Then, the number of subhaplogroups sampled versus  $D$  and  $D^*$  estimates were analyzed. A significant negative correlation between the number of subhaplogroups sampled and the estimated  $D$  and  $D^*$  values were observed ( $r^2=0.95$ ,  $P < 0.001$  and  $r^2=0.84$ ,  $P < 0.01$ , respectively. Supplementary Figure 3). A similar analysis using only the non-hg-I samples did not indicate a significant correlation between the number of populations sampled and the estimated  $D$  and  $D^*$  values (for  $D$ :  $r^2=0.01$ ,  $P=0.74$ ; for  $D^*$ :  $r^2=0.001$ ,  $P=0.97$ ).

#### Population genetic analyses of CD24L4

An initial DNA sequence analysis of 50 samples (10 from each of the haplogroups E, I, J, K\* and R) found an A to G transition at 662 bp upstream of *CD24L4* mRNA starting site (-A662G, Supplementary Figure 1) segregating only in the hg-I individuals. The -A662G change was in a repeat region characteristic of LINES. This high LINE resembling stretch of sequence is most probably a remnant of the retrotransposition event that generated the *CD24L4* on the Y chromosome. Further sequencing of 10 samples from each of the I1\*, I2a\*, I2a2\*, I2b\* and I\* subhaplogroups and additional genotyping of 484 samples from 10 haplogroups showed that the derived G allele was found only in the I2a\*, I2a2\*, I2b\* and I\* samples, whereas all I1\* samples had the ancestral A allele (Figure 1). This observation is in agreement with the recently published YCC 2008 Y-chromosomal haplogroup tree<sup>9</sup> in which the P215 marker shows that I2a\* and I2b\* lineages share a more recent common ancestor with each other than with the I1\* lineages. When I2a\* and I2b\* lineages were combined and compared against I1\* lineage in terms of progression

to AIDS outcomes, no significant progression difference was observed ( $P > 0.3$ ).

#### DISCUSSION

In a previous study, we discovered faster AIDS progression associated with European-specific hg-I individuals when compared against the individuals belonging to the most common European haplogroup R.<sup>4</sup> A standing question was whether AIDS susceptibility was specific to hg-I or it was also shared by its sister clade hg-J. In this study, using new genotyping data, we first identified individuals belonging to hg-J, the phylogenetically closest lineage to hg-I, and we confirmed the fast AIDS progression signal is specific to hg-I individuals using the hg-J as a reference group. The robustness of hg-I disease susceptibility association to correction for known effects of several autosomal loci and genome-wide allelic variants suggests that the disease signal seen here reflects Y-chromosome-specific loci. Moreover, we observed that none of the hg-I subhaplogroups can account for the faster progression signal alone and the AIDS signal can be best explained due to genetic variants present in all hg-I Y chromosomes. However, we note that generalization of our conclusions to individual hg-I subhaplogroups should be interpreted cautiously given the small sample sizes in some of these groups. Finally, a candidate gene approach of resequencing *DDX3Y*, a crucial autoimmunity gene, did not identify a causal variant that can explain the disease association.

Although resequencing of *DDX3Y* did not identify a causal variant, the sequence data indicated a significant deviation from neutrality in the frequency distribution of *DDX3Y* polymorphisms in the hg-I samples. Therefore, we conducted comprehensive comparative evolutionary and population genetic analyses on Y chromosomes belonging to hg-I and other common European haplogroups to examine potential roles of disease-based selection, demographic history and geographic differentiation affecting the evolutionary history of hg-I.

The significantly excess number of singletons observed in the *DDX3Y* among the hg-I samples can be caused by positive/directional selection or presence of weakly deleterious mutations. However, none

of the observed mutations are in the coding regions. It is highly unlikely that these mutations will be advantageous (or weakly deleterious) and give selective advantage (or disadvantage) to the hg-I Y chromosome. It is more likely that purifying selection is acting on *DDX3Y*. A recent survey of X-degenerate, single-copy protein-coding genes from Y chromosomes representing worldwide diversity also found very little variation in the exons of these genes (including *DDX3Y*), suggesting purifying selection as the primary factor in the evolutionary history of these genes.<sup>32</sup> Therefore, an explanation is needed for the excess of presumably neutral intronic polymorphisms observed in the hg-I samples.

Next, we examined if a population growth model can account for the observed patterns of variation in *DDX3Y* and applied three more population genetic tests in addition to the *D* and *D\** estimates. All of the tests indicated population growth for hg-I, although population equilibrium could not be rejected for the non-hg-I samples. However, individual hg-I subhaplogroups did not significantly deviate from population equilibrium. We acknowledge a sampling caveat that may be inflating the population growth estimators. As we were interested in a thorough genetic investigation of hg-I, we sampled representatives of all common hg-I subhaplogroups for sequencing. Due to the unique topology of Y haplogroup phylogenetic tree, this sampling scheme means that we are more prone to sampling from the tips of the Y haplogroup phylogenetic tree for the hg-I samples, capturing more singleton changes and inflating the *D* and *D\** estimates (Supplementary Figure 3).

The hg-I makes up only around 18% of the European Y-chromosome haplogroups,<sup>26</sup> therefore it is unlikely that hg-I have experienced a more drastic exponential growth compared to the other haplogroups. Moreover, when non-hg-I samples are pooled together, they did not show any evidence of population growth. It is expected that exponential population growth should affect polymorphism patterns throughout the genome. Some studies suggested that the DNA sequence data from autosomes and X chromosome did not show a rapid expansion in Europeans,<sup>33–35</sup> whereas others argued that the data were consistent with population growth after a past bottleneck.<sup>36,37</sup> To see what kind of population demography can be inferred from the X chromosome and autosomal homologs of *DDX3Y* and *CD24LA*, we calculated *D* and *D\** values for *DDX3X* and *CD24* using the HapMap data from 60 CEU samples. After considering for the ascertainment bias inherent in the HapMap genotype data, we cannot reject demographic equilibrium hypothesis for these loci. However, we note that Northern European descendant HapMap (CEU) *DDX3X* and *CD24* results may not be representative of genotype variation in all European populations. Adequate sampling of geographically diverse European populations and unbiased resequencing are crucial for understanding the demographic history of European populations.<sup>38</sup>

Due to biological forces (such as drift) and cultural practices (such as patrilocality), Y chromosome shows the most geographically structured variation pattern in the human genome. Recently, high levels of population substructure within European populations have been shown using high-density marker studies.<sup>39,40</sup> How much of this European population structure can be traced by Y-chromosome haplogroups is currently unknown. Hg-I shows strong geographic differentiation in Europe. Subhaplogroups I1\* and I2b\* are mostly found in north and west Europe, respectively. I2a\* is most frequent in Eastern Europe and Balkans, whereas I2a2\* is the predominant subhaplogroup in Sardinia.<sup>26</sup> Our high-density autosomal marker analyses with European descendant Americans suggest that there are genome-wide allelic differences between hg-I and other haplogroups

and within hg-I subhaplogroups. These differences may be tracing the different geographic origin of European American individuals with different Y (sub)haplogroups.

A different and more complex population and demographic history within the hg-I subhaplogroups was also evident. For the *DDX3Y* region, the I\* samples had at least twice the number of segregating sites and more than three times the haplotype diversity compared to the other I subhaplogroups, where two distinct haplotypes were evident. Also, the positive *D* estimate in I\* samples was in the opposite direction compared to other subhaplogroups. These observations suggest that I\* samples represent an older and/or more mixed population compared to other hg-I subhaplogroups that is in demographic equilibrium. Moreover, the presence of *CD24LA* –A662G change in I2a\*, I2b\* and I\* samples suggests that either this mutation originated independently on the I2\* and I\* Y chromosomes or that, at least some, I\* chromosomes share a more recent common ancestor with the I2\* clade. The first scenario of recurrent mutations on different Y chromosomes is unlikely given the low mutation rates observed on the Y chromosome and lack of observation of any other change on other examined (older) haplogroups (as they have existed for longer periods of time they would have had more chances to accumulate mutations). The paragroup I\* is a collection of hg-I Y chromosomes defined by the lack of derived I subhaplogroup markers. Although the paragroup represents the oldest hg-I lineages, it makes up only about 5% of the hg-I in European populations<sup>26</sup> and has not been thoroughly studied. Highest frequency I\* samples are found in southwestern Europe (2.9–5.8%) and Balkans (3.6%), the most probable last glacial period refuge areas of I1\*, I2b\* and I2a\*, respectively.<sup>26</sup> However, whether these geographically separate I\* populations represent a panmictic population or a long-term separated, and structured population is currently unknown. Considering that the high-density SNP data also suggest genome wide allelic differences between I\* and the rest of the subhaplogroups, testing a scenario of historic geographical population division for the two major haplotypes observed in I\* samples should be interesting using larger samples from Iberian peninsula, western and southeastern Europe.

The participants of this study were self-reported European descendants from the US population. We assumed that we will be able to detect the pre- and post-Neolithic evolutionary factors that had been acting on the European Y-chromosome lineages, without any significant influence of the post-Columbian expansion in the New World. This should be a reasonable assumption. First, the mutation rates on the human Y chromosome are rather slow and it is unlikely that there had been enough genetic diversity accumulation on the paternal lineages since post-Columbian expansion in the New World. Second, previously we observed that haplogroups I, J and R made up 17, 10 and 63% of the 2000 Y chromosomes studied from the US population, respectively.<sup>4</sup> These haplogroup frequencies sampled from the US population are rather similar to the frequency distribution of these haplogroups in Europe, not indicating any bias for or against hg-I expansion in the New World.

In conclusion, we present a unique evolutionary history of hg-I, the first human Y-chromosome haplogroup showing infectious disease association. Our analyses suggest that genetic changes occurred on the ancestor of hg-I Y chromosomes, after its split from its sister clade hg-J, may be responsible for AIDS susceptibility. However, we were not able to identify a causative locus. We also observed that even within a continent, Y chromosomes belonging to the same haplogroup may have very different life histories representing their diverse genomic backgrounds. This is an important observation that needs to be

considered in designing of future Y-chromosome-related disease association and replication studies. Determining the roles of disease-based selection, demographic history and population structure during the evolutionary history of a haplogroup is obviously complex. Although AIDS is a recent epidemic and might not have had enough time to affect Y chromosome sequence diversity, overall infectious disease susceptibility might be acting on certain Y chromosomes for longer periods of time. Our analyses could not show a clear selection against (or for) the hg-I. However, this study should be considered as early stages of a relatively new approach and our conclusions should serve as a stimulant for further research.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ACKNOWLEDGEMENTS

We thank the individuals and staff of all the participating cohorts in the study. San Francisco City Cohort samples were provided by Susan Buchbinder. We also thank Michael Malasky and Mary Thompson for their assistance. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government. The project described in this paper has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract Nos. HHSN261200800001E and N01-CO-12400.

- Jobling, M. A. & Tyler-Smith, C. The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* **4**, 598–612 (2003).
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
- Krausz, C., Quintana-Murci, L. & Forti, G. Y chromosome polymorphisms in medicine. *Ann. Med.* **36**, 573–583 (2004).
- Sezgin, E., Lind, J. M., Shrestha, S., Hendrickson, S., Goedert, J. J., Donfield, S. *et al.* Association of Y chromosome haplogroup I with HIV progression, and HAART outcome. *Hum. Genet.* **125**, 281–294 (2009).
- Tyler-Smith, C. An evolutionary perspective on Y-chromosomal variation and male infertility. *Int. J. Androl.* **31**, 376–382 (2008).
- Soulat, D., Burckstummer, T., Westermayer, S., Goncalves, A., Bauch, A., Stefanovic, A. *et al.* The DEAD-box helicase DDX3X is a critical component of the TANK-binding kinase 1-dependent innate immune response. *EMBO J.* **27**, 2135–2146 (2008).
- Yedavalli, V. S., Neuveut, C., Chi, Y. H., Kleiman, L. & Jeang, K. T. Requirement of DDX3 DEAD box RNA helicase for HIV-1 Rev-RRE export function. *Cell.* **119**, 381–392 (2004).
- Hughes, J. F., Skaletsky, H., Rozen, S., Wilson, R. K. & Page, D. C. Has the chimpanzee Y chromosome been sequenced? *Nat. Genet.* **38**, 853–854, author reply 854–855 (2006).
- Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L. & Hammer, M. F. New binary polymorphisms reshape and increase resolution of the human Y chromosome haplogroup tree. *Genome Res.* **18**, 830–838 (2008).
- O'Brien, S. J. & Nelson, G. W. Human genes that limit AIDS. *Nat. Genet.* **36**, 565–574 (2004).
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007).
- Repping, S., van Daalen, S. K., Brown, L. G., Korver, C. M., Lange, J., Marszalek, J. D. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38**, 463–467 (2006).
- Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
- Nei, M. *Molecular Evolutionary Genetics* (Columbia University Press, New York, 1987).
- Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
- Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
- Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925 (1997).
- Ramos-Onsins, S. E. & Rozas, J. Statistical properties of new neutrality tests against population growth. *Mol. Biol. Evol.* **19**, 2092–2100 (2002).
- Rogers, A. R. & Harpending, H. Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9**, 552–569 (1992).
- Rozas, J., Sanchez-DelBarrio, J. C., Messeguer, X. & Rozas, R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496–2497 (2003).
- Schneider, S. & Excoffier, L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**, 1079–1089 (1999).
- Chakraborty, R. Mitochondrial DNA polymorphism reveals hidden heterogeneity within some Asian populations. *Am. J. Hum. Genet.* **47**, 87–94 (1990).
- Schneider, S., Roessli, D. & Excoffier, L. *A Software for Population Genetics Data Analysis* (Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000).
- Libert, F., Cochaux, P., Beckman, G., Samson, M., Aksanova, M., Cao, A. *et al.* The deltaaccr5 mutation conferring protection against HIV-1 in Caucasian populations has a single and recent origin in Northeastern Europe. *Hum. Mol. Genet.* **7**, 399–406 (1998).
- Rootsi, S., Magri, C., Kivisild, T., Benuzzi, G., Help, H., Bermisheva, M. *et al.* Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am. J. Hum. Genet.* **75**, 128–137 (2004).
- Semino, O., Magri, C., Benuzzi, G., Lin, A. A., Al-Zahery, N., Battaglia, V. *et al.* Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am. J. Hum. Genet.* **74**, 1023–1034 (2004).
- Sun, C., Skaletsky, H., Birren, B., Devon, K., Tang, Z., Silber, S. *et al.* An azoospermic man with a *de novo* point mutation in the Y-chromosomal gene USP9Y. *Nat. Genet.* **23**, 429–432 (1999).
- Shen, P., Wang, F., Underhill, P. A., Franco, C., Yang, W. H., Roxas, A. *et al.* Population genetic implications from sequence variation in four Y chromosome genes. *Proc. Natl Acad. Sci. USA* **97**, 7354–7359 (2000).
- Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. & Feldman, M. W. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl Acad. Sci. USA* **97**, 7360–7365 (2000).
- Nielsen, R., Hubisz, M. J. & Clark, A. G. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics* **168**, 2373–2382 (2004).
- Rozen, S., Marszalek, J. D., Alagappan, R. K., Skaletsky, H. & Page, D. C. Remarkably little variation in proteins encoded by the Y chromosome's single-copy genes, implying effective purifying selection. *Am. J. Hum. Genet.* **85**, 923–928 (2009).
- Voight, B. F., Adams, A. M., Frisse, L. A., Qian, Y., Hudson, R. R. & Di Rienzo, A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl Acad. Sci. USA* **102**, 18508–18513 (2005).
- Pluzhnikov, A., Di Rienzo, A. & Hudson, R. R. Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics* **161**, 1209–1218 (2002).
- Wall, J. D. & Przeworski, M. When did the human population size start increasing? *Genetics* **155**, 1865–1874 (2000).
- Adams, A. M. & Hudson, R. R. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics* **168**, 1699–1712 (2004).
- Marth, G. T., Czabarka, E., Murvai, J. & Sherry, S. T. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* **166**, 351–372 (2004).
- Wall, J. D., Cox, M. P., Mendez, F. L., Woerner, A., Severson, T. & Hammer, M. F. A novel DNA sequence database for analyzing human demographic history. *Genome Res.* **18**, 1354–1361 (2008).
- Heath, S. C., Gut, I. G., Brennan, P., McKay, J. D., Bencko, V., Fabianova, E. *et al.* Investigation of the fine structure of European populations with applications to disease association studies. *Eur. J. Hum. Genet.* **16**, 1413–1429 (2008).
- Tian, C., Plenge, R. M., Ransom, M., Lee, A., Villoslada, P., Selmi, C. *et al.* Analysis and application of European genetic substructure using 300K SNP information. *PLoS Genet.* **4**, e4 (2008).

Supplemental Information accompanies the paper on Journal of Human Genetics website (<http://www.nature.com/jhg>)