Human populations show reduced DNA sequence variation at the Factor IX locus

Eugene E. Harris* and Jody Hey†

Levels and patterns of human DNA sequence variation vary widely among loci. However, some of this variation may be due to the different populations used in different studies. So far, few studies of diverse human populations have compared different genetic loci for the same samples of populations and individuals. Here, we present new polymorphism data from intron 4 of the Factor IX gene (*FIX*) sequenced in diverse Old World populations. An explicit comparison is made with another X-linked gene, *PDHA1*, for which the sampling of individuals was very similar. Despite having a similar amount of divergence from chimpanzees, as do other nuclear genes, *FIX* has comparatively much less DNA sequence variation among humans. Nucleotide diversity at *FIX* is the lowest among the existing non-Y chromosome nuclear gene datasets and is less than 10% of the diversity found at *PDHA1*. Estimates of effective population size based on *FIX* are 8,558, about half of the value obtained for *PDHA1*, and the time to the most recent common ancestry among human *FIX* gene copies (282,000 years) is one of the most recent estimates reported for human genes. Analyses presented here suggest a history for the *FIX* region that includes recent positive directional selection, or background, selection. The general conclusion emerging is that very large variations can exist between the histories of similar genomic regions, even when sampling differences are minimized.

Addresses: *Universidade de São Paulo, Cidade Universitaria, Departamento de Biologia, Instituto de Biociências, Rua do Matão-Travessa 14 number 321, CEP 05508-900, São Paulo, Brasil. †Department of Genetics, Rutgers University, 604 Allison Road, Piscataway, New Jersey, 08854, USA.

Correspondence: Eugene E. Harris and Jody Hey
E-mail: eeharris@ib.usp.br (E.E.H.); jhey@mbcl.rutgers.edu (J.H.)

Received: **9 February 2001**
Revised: **27 March 2001**
Accepted: **27 March 2001**

Published: **15 May 2001**

Current Biology 2001, 11:774–778

0960-9822/01/$ – see front matter

Results and discussion

DNA sequences were collected from 36 human males from diverse Old World populations including Asians, Europeans, and Africans. The variation found among 36 copies of a portion of the Factor IX (*FIX*) gene is given in Figure 1. Only six polymorphic sites (plus 1 indel) were found in over 3700 base pairs, and only one of these polymorphisms was found among non-African sequences. Values derived for both of the commonly used estimates of the population mutation parameter θ (equal to $4Nu$ or $3Nu$ for X-linked genes like *FIX*, where N is the effective population size and u is the neutral mutation rate) are considerably lower than values typically reported for human genes. The average pairwise difference, or π, is 0.00014 per base pair, and Watterson's estimator, θ_W, is 0.00039 per base pair. The fact that π is lower than θ_W reflects the low frequency of the polymorphic bases listed in Figure 1. For five of the six human polymorphisms, the derived base (by comparison with the chimpanzee sequence) appears only once or twice (as a singleton or a doubleton) among the human sequences, and the most common polymorphism, at position 550, appears in only 3 sequences. The disparity in estimates of $3Nu$ is also reflected in a negative value of Tajima's test statistic ($D_T = -1.7132$, $0.10 > p > 0.05$) and Fu and Li's test statistic ($D_{FL} = -1.3272$) [1, 2].

The *FIX* locus data revealed no evidence of recombination; although, with only two informative base positions, there was very little power to detect recombination [3]. The dearth of informative polymorphisms is also reflected in the gene tree estimate (Figure 2), which has very little structure. We used GENETREE for maximum likelihood coalescent analyses and to estimate the time to the most recent common ancestral sequence (TMRCA) and the population mutation parameter. The maximum likelihood estimate of θ for the entire region was 1.86. Based on this value and the estimated mutation rate determined using the net average sequence divergence from chimpanzees and assuming a 5MY date of divergence between chimpanzees and humans, N_e was estimated at 8,378. The TMRCA of the full *FIX* gene tree was estimated at 281,501 (Figure 2). Examination of the parsimony tree (data not shown) rooted with an orangutan sequence shows that there is little relative rate difference along human and chimpanzee lineages. Of 37 fixed substitution differences between these groups, almost equal numbers were assigned to the human (20) and chimpanzee (17) branches.

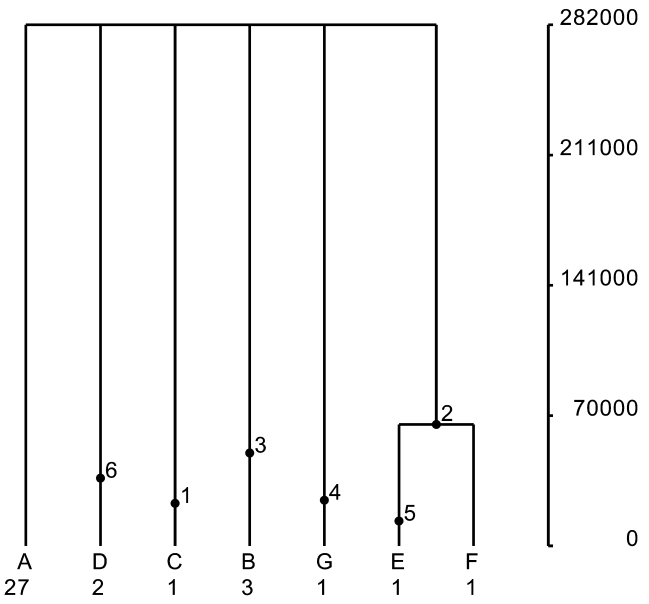
Previously, we reported on DNA sequence variation in

Figure 1

Haplotypes	Base Position	Populations								
	333									
	125345									
	675191	African				Non-African				
	150417	S	P	B	K	M	Sy	F	V	C
Chimpanzee	TAATAT									
A	------	4	3	2	2	2	1	3	5	5
B	--G---	-	-	1	-	-	-	2	-	-
C	C-----	-	1	-	-	-	-	-	-	-
D	-----A	1	1	-	-	-	-	-	-	-
E	-G--G-	-	-	1	-	-	-	-	-	-
F	-G----	-	-	-	1	-	-	-	-	-
G	---C--	-	-	-	1	-	-	-	-	-

Unique haplotypes and polymorphic sites found in humans are shown with respect to the chimpanzee sequence. Each position is numbered according to its position in the alignment. A dash for a given haplotype at a given site indicates that the nucleotide at that site is identical to the nucleotide in the chimpanzee sequence. On the right are the given counts of the haplotypes within each sample: S = Senegalese; P = Pygmy (including Mbuti from northeastern Zaire and Biaka from the Central African Republic); B = Southern African Bantu speakers; K = Khoisan, from the Angola/Namibia border; M = Mongolian; Sy = a male from the Druze Islamic sect, Syria; F = French; V = Vietnamese; and C = Chinese. The indel is found at site number 2838 and occurs in a single Mbuti Pygmy sequence and a single Senegalese sequence.

Figure 2



The gene tree estimate for *FIX*, with estimated ages of polymorphic mutations. Each mutation is identified by its position in the sequence and by its branch location (as determined by maximum parsimony and maximum likelihood). The single mutation that appears at a branch point indicates the emergence of the haplotype G.

the *PDHA1* gene in largely the same subpopulations and in many of the same individuals as for *FIX* [4]. *PDHA1* is a subunit of the pyruvate dehydrogenase complex located on the X chromosome (Xp22.2–22.1). The *PDHA1* sequence consists of 4220 base pairs of contiguous sequence, most of which is from noncoding intron regions. Because of the similarities between these genomic regions (i.e., both are X-linked, similar in length, and largely or entirely noncoding) and the very different patterns of variation they display, we present an explicit comparison between them.

PDHA1 contains over ten times the average pairwise diversity that *FIX* contains (0.179 versus 0.014); the effective population size of *PDHA1* is twice that of *FIX* (18,184 versus 8,558), the age of coalescence for *PDHA1* is about six times that of *FIX* (281,501 versus 1,727,480), and while *PDHA1* displays an excess of intermediate polymorphism, *FIX* shows an excess of rare variants (Tajima's D statistic, 0.7843 versus −1.7132). This difference in Tajima's D values was tested using a coalescent simulation method [5] and was found to be statistically significant ($p = 0.034$).

An explicit contrast between the gene trees for *FIX* and *PDHA1*, based on the 24 identical individuals sequenced for both loci, underscores the large differences in nucleotide diversity and histories between these two regions (Figure 3). The gene tree for *PDHA1* contains long and deep branches in the African sample and indicates strong geographic structure due to the fixed difference occurring between Africans and non-Africans. The tree is also dominated by mutations falling on internal branches of the tree. In contrast, the *FIX* tree reveals very little population structure (Figure 2), stemming from its low level of polymorphism and shallow history. Most mutations found at *FIX* occur on the tree's terminal branches.

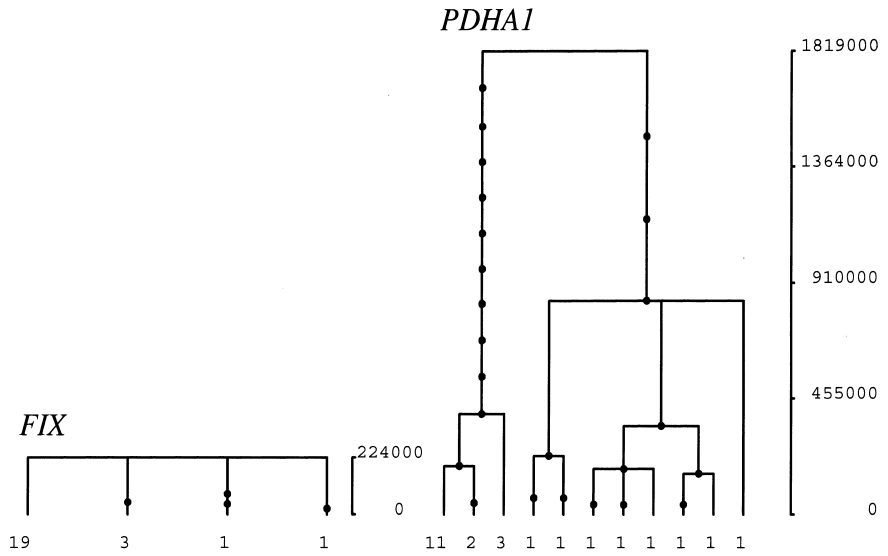
Reduced diversity at *FIX* does not appear to be due to a reduced mutation rate at this locus, since it has an estimated mutation rate that is closely similar to that of *PDHA1* (Table 1). Both have very similar levels of sequence divergence from chimpanzees, with 37 fixed mutations between chimpanzees and humans at *FIX* and 35 fixed mutations between this pair at *PDHA1*.

Crossover rates per generation for *FIX* and *PDHA1* in humans have been estimated by Nachman (see http://eebweb.arizona.edu/nachman/publications/data/microsats.html) to be 3.30 cM/Mb and 4.42 cM/Mb, respectively. Since these values are very similar, it appears doubtful that a difference in recombination rates accounts for the different levels of nucleotide diversity between *FIX* and *PDHA1*.

Natural selection can play a large role in shaping patterns of variation at specific loci. The HKA test [6] provides a means of detecting significant departures from neutral expectations in the amount of divergence and levels of polymorphism in a single locus by comparing these mea-

Figure 3

Comparison of the gene trees estimated for the *FIX* and *PDHA1* genes. Mutations that generated new haplotypes are indicated at branch points. Multiple mutations along the same branch are ordered arbitrarily with respect to time. For the *PDHA1* tree, lines are drawn under the non-African haplotypes (to the left) and the African *PDHA1* haplotypes (to the right).



sures across different loci. We performed three tests: *FIX* versus *PDHA1*, *FIX* versus β-*globin*, and *PDHA1* versus β-*globin*. Only the comparison between *PDHA1* and *FIX* closely approached significance ($p = 0.0692$). This may result for three reasons: reduced levels of observed polymorphism in *FIX*, increased levels of observed polymorphism in *PDHA1*, or their deviation from each other in different directions. Since diversity at β-*globin* is intermediate between that at *PDHA1* and *FIX* (comparing African diversity), the latter explanation seems likely.

It is possible that the reduced levels of variation at *FIX*, and its excess of low frequency variants, reflects a history shaped by natural selection. Two forms of selection can produce such a pattern. Positive directional selection for a favorable allele can produce a selective sweep in which neutral mutations linked to this allele rise rapidly in frequency, causing a loss of polymorphism in the region [7]. Alternatively, under a background selection model, the recurrent removal of deleterious mutations and their linked neutral mutations [8] causes reduced polymorphism. Tests of these different selection regimes have been shown to have low resolving power when there is low variation [9], as seems generally the case for humans. Since the *FIX* region that was sequenced is entirely noncoding, it is possible that the region sequenced is linked

Table 1

Estimates of population genetic parameters compared for the *FIX* and *PDHA1* gene regions.

Gene and sample	n[1]	S[2]	π[3]	θ_{ML}[4]	D[5]	N_e[6]	*TMRCA*[7]	u[8]
FIX								
Old World	36	6	0.00014	1.86	−1.71	8,378	281,501	9.950×10^{-10}
African	18	6	0.00021	2.4	−1.66	10,811	–	–
Non-African	18	1	0.00006	0.30	−0.53	1,351	–	–
PDHA1								
Old World	35	25	0.00179	4.41	0.78	18,184	1,727,480	9.76×10^{-10}
African	16	23	0.00197	5.78	0.66	23,833	–	–
Non-African	19	2	0.00011	0.62	−0.48	2,561	–	–

[1] n is the number of chromosomes sequenced.
[2] S is the number of polymorphic sites observed in the sample.
[3] π is the average pairwise difference expressed (Nei, 1987).
[4] θ_{ML} is the maximum likelihood estimate of 3Nu obtained using the GENETREE program.
[5] D is Tajima's statistic (1989), measuring the departure of polymorphic site frequencies from those expected under neutrality and constant population size.
[6] N_e is the effective number of gene copies estimated from θ_{ML} and u.
[7] *TMRCA* is the time to the most recent common ancestor estimated using the GENETREE program.
[8] u is the mutation rate per base pair per year obtained by solving $D = 2ut$, where D is the average number of fixed differences between humans and chimpanzees, and t is the estimated time since they last shared a common ancestor (5MY). Because of indels in the data, the average lengths of the sequence we were able to compare between the human dataset and the chimpanzees were 3,717 and 4,142 for *FIX* and *PDHA1*, respectively.

Table 2

Comparison of population parameters for nuclear loci.

Gene (Chromosome)	L[1]	N[2]	S[3]	π[4]	D[5]	*TMRCA*[6]	N_e[7]	Reference
F9 (X)	3,731	36	6	0.00014	−1.713	281,501	8,378	this study
PDHA1 (X)	4,220	35	25	0.00179	0.784	1,727,480	18,184	[4]
DYS (X)	8,000	250	35	0.00101	0.962	∼600,000	11,200	[19]
ZFX (X)	1,089	336	10	0.00082	−0.95	1,100,000	21,700	[20]
Xq13.3 (X)	10,200	69	33	0.00036	−1.61	535,000	16,000	[15]
DMD intron 44 (X)	3,000	41	19	0.00141	−0.16	1,560,000	26,000	[16]
DMD intron 7 (X)	2,389	41	9	0.00034	−1.79	210,000	3,500	[16]
β-*globin* (11)	2,670	349	28	0.00180	1.158	800,000	11,661	[21]
22q11.2 (22)	9,834	128	71	0.00088	−1.03	1,288,000	12,200	[22]
16p13.3 (16)	1,700	100	42	0.00147	−2.14	1,040,000	16,800	[23]
1q24	9.7	122	48	0.00058	−1.20	1,559,000	12,600	[24]
ACE (17)	24,070	22	78	0.00093	−0.32	1,113,000	10,200	[25]
LPL (8)	9,734	142	79	0.0020	0.909	1,200,000	16,900	[26]
SMCY (Y)	39,931	53	47	0.000072	−2.31	48,000[8]	6,100	[27,28]

[1] L is the length of the sequences in units of base pairs.
[2] N is the number of chromosomes sampled.
[3] S is the number of polymorphic sites observed in the sample.
[4] π is the average pairwise difference for the specified sample over the total length in base pairs of the sequence.
[5] D is Tajima's statistic (1989), measuring the departure of polymorphic site frequencies from those expected under neutrality and constant population size.
[6] *TMRCA* is the time to the most recent common ancestral DNA sequence, as reported in the references.
[7] N_e is the effective number of gene copies, as given in the reference.
[8] This value is based on a model of exponential population growth (see reference).

to a genomic region that is under selection (i.e., exons of *FIX* or its regulatory regions).

Both the *FIX* and *PDHA1* regions lie within important protein coding loci [10–11], and such genes can be expected to be under purifying selection, which removes deleterious mutations [12]. If this is the only type of natural selection that has been occurring, then these and other such genes should show a good correspondence to the neutral model [13–14]. Indeed, our studies began without prior expectations that either locus might have been the target of adaptive, or balancing, selection, which would shift polymorphism levels, and so the neutral model is an appropriate statistical baseline. In a previous report, *PDHA1* was shown to have an unusual pattern of DNA sequence variation, with a very deep gene tree among samples from Africa, very little variation in non-Africans, and a fixed difference between the two geographic groups [4]. The *FIX* data reveal that this locus is also atypical, though in quite a different way. In this case, we found less variation than expected, both among Africans and non-Africans, and the gene tree estimated had a MRCA that was far more recent than that reported in all other studies of X-linked or autosomal loci (Table 2).

Table 2 summarizes the nucleotide diversity and population parameters estimated for nuclear datasets sequenced in large diverse human populations. Diversity at *FIX* is considerably reduced compared to all other X-linked genes, comprising only 15% of the average diversity at X-linked loci. Furthermore, it is less than half as diverse as *DMD* intron 7, otherwise the least diverse X-linked gene region known. At the other extreme, *PDHA1* is more than twice as diverse as the average X-linked gene and 30% more diverse than any other X-linked gene.

We note, however, that *FIX* does show some similarities with two other X-linked loci. Both *Xq13.3* [15] as well as *DMD* intron 7 [16] show comparatively reduced levels of nucleotide diversity (*Xq13.3*, $\pi = 0.00037$; *DMD* intron 7, $\pi = 0.00034$) (Table 2). Also, Tajima's D statistics for *Xq13.3* ($D = -1.61$) and DMD intron 7 ($D = -1.79$) indicate an abundance of rare polymorphisms. Furthermore, coalescence dates for these regions, 535,000 years estimated for *Xq13.3* and 210,000 years estimated for *DMD* intron 7, are recent when compared to other X-linked and autosomal genes (Table 2). Indeed, it has been suggested that positive natural selection has shaped variation at *DMD* intron 7 [16]. Regarding *Xq13.3*, we note that it has a low recombination rate (0.16 cM/Mb [15]) and that it is possible that natural selection may have caused the low variation at this putative intergenic region.

Few studies have compared genetic diversity and gene histories at different loci for closely similar samples of individuals. Although such comparisons are intended to remove the variance between locus-specific estimates that stems from different sampling strategies, we have shown that very large differences may still occur between similar noncoding genomic regions. Nachman et al. [17] and Nachman and Crowell [16] have also collected sequence datasets from different genomic regions in closely similar samples and have uncovered widely different levels of diversity and gene histories. Recently, Ingman et al. [18]

have sequenced the entire mitochondrial genome in a very similar sample of individuals, as sequenced for the *Xq13.3* intergenic region [15]. Interestingly, their estimates of TMRCA for mitochondria (171,500 years) and for *Xq13.3* (479,000 years) are in close agreement with the 3-fold difference in coalescence expected between these different genomic systems. However, we point out that the mitochondrial coalescence is inconsistent with the coalescent estimates for the majority of X-linked and other autosomal gene loci studied to date, which are disproportionately much older. It appears, therefore, that there is a large variance among human loci, in their apparent histories, and that this could greatly complicate our estimates of human populational level history.

Supplementary material

Supplementary material including full materials and methods is available at http://images.cellpress.com/supmat/supmatin.htm.

Acknowledgements

We thank Eldredge Bermingham, Laurent Excoffier, Trefor Jenkins, Lynne Jorde, Connie Kolman, and Jeff Rogers for providing DNA samples. Walter Neves and João Morgante gave assistance at Universidade de São Paulo, and Elbert Chen assisted in data analyses. This work was supported by the National Institutes of Health (grant R55GM54684) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, grant 99/12959-1).

References

1. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585-595.
2. Fu YX, Li WH: **Statistical tests of neutrality of mutations.** *Genetics* 1993, **133:**693-709.
3. Hudson R: **The sampling distribution of linkage disequilibrium under an infinite allele model without selection.** *Genetics* 1985, **109:**611-631.
4. Harris E, Hey J: **X chromosome evidence for ancient human histories**. *Proc Natl Acad Sci USA* 1999, **96:**3320-3324.
5. Hey J: **Mitochondrial and nuclear gene trees present conflicting portraits of human origins.** *Mol Biol Evol* 1997, **14:**166-172.
6. Hudson RR, Kreitman M, Aguade M: **A test of neutral molecular evolution based on nucleotide data.** *Genetics* 1987, **116:**153-159.
7. Maynard-Smith J, Haigh J: **The hitch-hiking effect of a favourable gene.** *Genet Res Camb* 1974, **23:**23-35.
8. Charlesworth D, Charlesworth B, Morgan MT: **The pattern of neutral molecular variation under the background selection model.** *Genetics* 1995, **141:**1605-1617.
9. Wayne ML, Simonson KL: **Statistical tests of neutrality in the age of weak selection.** *Trends Ecol Evol* 1998, **13:**236-240.
10. Giannelli F, Green PM, Sommer SS, Poon M, Ludwig M, Schwaab R, *et al.*: **Haemophilia B: database of point mutations and short additions and deletions–eighth edition**. *Nucleic Acids Res* 1998, **26:**265-268.
11. Koike K, Urata Y, Matsuo S, Koike M: **Characterization and nucleotide sequence of the gene encoding the human pyruvate dehydrogenase alpha-subunit.** *Gene* 1990, **93:**307-311.
12. Li WH: *Molecular Evolution*. Sunderland, Massachusetts: Sinauer Associates; 1997.
13. Przeworski M, Charlesworth B, Wall JD: **Genealogies and weak purifying selection**. *Mol Biol Evol* 1999, **16:**246-252.
14. Tachida H: **DNA evolution under weak selection.** *Gene* 2000, **261:**3-9.
15. Kaessmann H, Heibig F, von Haeseler A, Paabo S: **DNA sequence variation in a con-coding region of low recombination on the human X chromosome.** *Nat Gen* 1999, **22:**78-81.
16. Nachman MW, Crowell SL: **Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, Dmd, in humans.** *Genetics* 2000, **155:**1855-1864.
17. Nachman M, Bauer V, Crowell SL, Aquadro C: **DNA variability and recombination rates at X-linked loci in humans.** *Genetics* 1998, **150:**1133-1141.
18. Ingman M, Kaessmann K, Pääbo S, Gyllensten U: **Mitochondrial genome variation and the origin of modern humans.** *Nature* 2000, **408:**708-713.
19. Zietkiewicz E, Yotova V, Jarnik M, Korab-Laskowska M, Kidd KK, Modiano D, *et al.*: **Genetic structure of the ancestral population of modern humans**. *J Mol Evol* 1998, **47:**146-155.
20. Jaruzelska J, Zietkiewicz E, Batzer M, Cole DEC, Moisan JP, Scozzari R, *et al.*: **Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy.** *Genetics* 1999, **152:**1091-1101.
21. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox M, Schneider JA, *et al.*: **Archaic African and Asian lineages in the genetic ancestry of modern humans**. *Am J Hum Genet* 1997, **60:**772-789.
22. Zhao Z, Li J, Fu YX, Ramsay M, Jenkins T, Leskinen E, *et al.*: **Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22.** *Proc Natl Acad Sci USA* 2000, **97:**11354-11358.
23. Alonso S, Armour JA: **A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa.** *Proc Natl Acad Sci USA* 2001, **98:**864-869.
24. Yu N, Zhao Z, Fu YX, Sambuughin N, Ramsay M, Jenkins T, *et al.*: **Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1**. *Mol Biol Evol* 2001, **18:**214-222.
25. Rieder M, Taylor S, Clark A, Nickerson D: **Sequencing variation in the human angiotensin converting enzyme.** *Nat Genet* 1999, **22:**59-62.
26. Clark A, Weiss K, Nickerson D, Taylor S, Buchanan A, Stengard J, *et al.*: **Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase.** *Am J Hum Genet* 1998, **63:**595-612.
27. Shen P, Frank W, Underhill PA, Franco C, Yang WH, Roxas A, *et al.*: **Population genetic implications from sequence variation in four Y chromosome genes**. *Proc Natl Acad Sci USA* 2000, **97:**7354-7359.
28. Thompson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW: **Recent common ancestry of human Y chromosomes: evidence from DNA sequence data.** *Proc Natl Acad Sci USA* 2000, **97:**7360-7365.