DNA sequence variation in a non-coding region of low recombination on the human X chromosome

Henrik Kaessmann, Florian Heißig, Arndt von Haeseler & Svante Pääbo

DNA sequence variation has become a major source of insight regarding the origin and history of our species[1–5] as well as an important tool for the identification of allelic variants associated with disease. Comparative sequencing of DNA has to date focused mainly on mitochondrial (mt) DNA, which due to its apparent lack of recombination and high evolutionary rate lends itself well to the study of human evolution[1]. These advantages also entail limitations. For example, the high mutation rate of mtDNA results in multiple substitutions that make phylogenetic analysis difficult and, because mtDNA is maternally inherited, it reflects only the history of females. For the history of males, the non-recombining part of the paternally inherited Y chromosome can be studied[2]. The extent of variation on the Y chromosome is so low that variation at particular sites known to be polymorphic rather than entire sequences are typically determined[6]. It is currently unclear how some forms of analysis (such as the coalescent) should be applied to such data. Furthermore, the lack of recombination means that selection at

any locus affects all 59 Mb of DNA. To gauge the extent and pattern of point substitutional variation in non-coding parts of the human genome, we have sequenced 10 kb of non-coding DNA in a region of low recombination at Xq13.3. Analysis of this sequence in 69 individuals representing all major linguistic groups reveals the highest overall diversity in Africa, whereas deep divergences also exist in Asia. The time elapsed since the most recent common ancestor (MRCA) is 535,000±119,000 years. We expect this type of nuclear locus to provide more answers about the genetic origin and history of humans.

Population studies of nuclear DNA sequences have recently been published for genes encoding β-globin[3], lipoprotein lipase[4] and pyruvate dehydrogenase E1 α-subunit[5]. Although these loci have yielded interesting results, their involvement in haemoglobinopathies[7], cardiovascular disease[8] and neurological conditions[9], respectively, make them likely targets for selection. Furthermore, at least in the first two genes, relatively high recombination rates make evolutionary inference difficult. For several

Fig. 1 Map of the world indicating the approximate origins of the individuals studied. The individuals belong to the following language phyla (numbers correspond to those in Fig. 2): Khoisan (8, 65, 66), Niger-Kordofanian (5, 6, 7, 9, 10, 11, 12, 24, 26, 27, 34, 57, 58, 63, 64, 67, 68, 69), Nilo-Saharan (25), Afro-Asiatic (23), Caucasian (4), Indo-Hittite (14, 18, 19, 20, 21, 22, 30, 36, 53, 61), Uralic-Yukaghir (28, 29, 35), Altaic (15, 16, 38, 43, 44, 45, 46, 48, 49, 55, 60), Chukchi-Kamchatkan (3), Eskimo-Aleut (54), Elamo-Dravidian (17), Sino-Tibetan (39, 40, 41, 42), Austric (13, 33, 47, 51), Indo-Pacific (31, 32, 52, 59, 62), Australian (1, 2) and Amerind (37, 50, 56).

Max Planck Institute for Evolutionary Anthropology, Inselstrasse 22, D-04103 Leipzig, Germany. Correspondence should be addressed to H.K. (e-mail: kaessmann@eva.mpg.de).
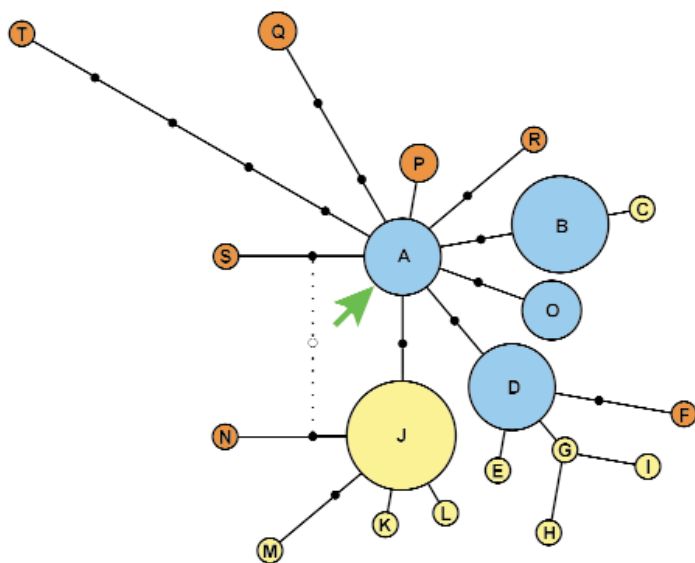
Fig. 2 Variable positions for the individuals studied. Individuals with identical sequences are grouped together. Letters on the right (A–T) denominate the different sequences.

	Sequence
Gorilla	C T T A C A C G C C G C C C C A G C G C G T C T T A C C A C C T C
Chimpanzee	C T G A C A C A C C A C C C C A G C G C G T C T T A C C A C C T C
Consensus	T T T A C A C A C C G C C C C A G C G C G T C T T A C C A C C T C
1 Abor. Australian	- A
2 Abor. Australian	- -
3 Chuckchi	- -
4 Georgian	- -
5 Biaka	- -
6 Ibo	- -
7 I isongo	- -
8 San	- -
9 Bamileke	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - - B
10 Effik	- - - - - - C - - - - - - - - - - - - - - - - C - - - - - - - - -
11 Mandenka	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
12 Mkamba	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
13 Filipino	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
14 Iranian	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
15 Kyrgyz	- - - - - - C - - - - - - - - - - - - - - - - C - - - - - - - - -
16 Uzbek	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
17 Tamile	- - - - - - C - - - - - - - - - - - - - - - - C - - - - - - - - -
18 Dutch	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
19 French	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
20 Italian	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
21 Orkadian	- - - - - - G - - - - - - - - - - - - - - - - C - - - - - - - - -
22 English	- - - - - C - - - - - - - - - - - - - - - - C - C - - - - - - - - C
23 Hausa	- - - - C - - - - - - - - - - - - - - - - - G - - - - - - - - - - D
24 Kikuyu	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
25 Luo	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
26 Mandenka	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
27 Mbenzele	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
28 Estonian	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
29 Saami	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
30 Swedish	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
31 Nasioi	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
32 PNG (Highl.)	- - - C - - - - - - - - - - - - - - - - - - G - - - - - - - - - -
33 PNG (Lowl.)	- - - C - - - - - - - - - - - - - - - - - - G - - - - - C - - - - E
34 Ewondo	- - - C - - - - - - - - - - - - - T - - - - G - T - - - - - - - - F
35 Saami	- - - C T - - - - - - - - - - - - - - - - - G - - - - - - - - - - G
36 Asian Indian	- - - C I - - - - - - - - - I - - - - - - - G - - - - - - - - - - H
37 Warao Indian	- - - C T - - - - - - - - - - - A - - - - - G - - - - - - - - - - I
38 Burial	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - - J
39 Chinese	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
40 Chinese	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
41 Chinese	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
42 Chinese	- - - - - - - - - - I - - - A - - - - - - - - - - - - - - - - - -
43 Korean	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
44 Japanese	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
45 Japanese	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
46 Evenk	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
47 Thai	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
48 Crimean Tatar	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
49 Uzbek	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
50 Native American	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
51 PNG (Lowl.)	- - - - - - - - - - I - - - A - - - - - - - - - - - - - - - - - -
52 Nasioi	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
53 Swedish	- - - - - - - - - - T - - - A - - - - - - - - - - - - - - - - - -
54 Inuit	- - - - - - - - - - T T - - T - - - - - - - - - - - - - - - - - - K
55 Japanese	- - - - - - G - - - - - - - T - - A - - - - - - - - - - - - - - - L
56 Warao	C - - - - - - - - - - - - - T - - A - - - - - - - - - - - T - - - M
57 Yoruba	- - - - - - - - A - - - T - - A - - - - - - - - - - - - - G - - - N
58 Mbenzele	- - - - - - - - - - - - - A - - - - - - - - C - - - - - - - - - - O
59 Nasioi	- - - - - - - - - - - - - A - - - - - - - - C - - - - - - - - - -
60 Crimean Tatar	- - - - - - - - - - - - - A - - - - - - - - C - - - - - - - - - -
61 German	- - - - - - - - - - - - - A - - - - - - - - C - - - - - - - - - -
62 PNG (Highl.)	- - - - - - - - - - - - - A - - - - - - - - C - - - - - - - - - -
63 Mbuti	- - - - - - - - - - - - - - A - - - - - - - - - - - - - - - - - - P
64 Mbuti	- - - - - - - - - - - - - - A - - - - - - - - - - - - - - - - - -
65 San	- - - - - - G - - - - - - - - - - - C - G - - - - - - - - - - - - Q
66 San	- - - - - - G - - - - - - - - - - - C - G - - - - - - - - - - - -
67 Biaka	- C - - - - - - - T - R
68 Biaka	- - - - - - - - A - G - - - S
69 Yoruba	- - G - - - G - - - - - T - T - - - - - - - - - - - - - - - - G T T

reasons, non-coding DNA at Xq13.3 might represent a class of nuclear loci ideally suited for unravelling the evolutionary history of the nuclear genome. First, this sequence is non-coding and therefore highly unlikely to be the direct target of positive or negative selection. Second, Xq13.3 displays a rate of recombination of approximately 0.16 cM/Mb, approximately one-eighth of the average on the X chromosome[10] (1.3 cM/Mb). Such a low recombination rate is not expected to cause much reshuffling of ancestral DNA sequences, which would confound many methods of evolutionary analysis. A low recombination rate may in principle be expected to increase the extent by which selection at flanking loci may affect the DNA sequence variation due to selective sweeps and background selection[11,12]. Xq13.3, however, is located 1.1 Mb from the nearest known gene[13] (phosphoglycerate kinase-1, *PGK1*). Over this distance, linkage disequilibrium is unlikely to occur[14] even if *PGK1* was subject to selection in some part(s) of the world. Third, Xq13.3 is advantageous from a practical standpoint because the assignment of phase between variable positions is trivial for the X chromosome when males are studied.

There are at least three possible sampling strategies to gauge the genetic diversity of the human species. First, samples are collected according to current population size; however, recent population expansions, for example those associated with the invention of agriculture and colonizations, will bias such an approach. Second, samples are collected according to geography. In this case, areas with low population density will be overrepresented. Third, major linguistic groups form the basis of the sampling. This appears preferable because population expansions during the last few thousand years are unlikely to influence the results. Thus, 69 male samples were collected such that 16 of Earth's 17 language phyla as defined by Ruhlen[15] were represented by one or more individual(s) (Fig. 1). In each of these individuals, 10,200 consecutive sequence positions were determined at least once from each DNA strand, identifying 33 polymorphic positions (Fig. 2). Together, these variable positions defined 20 sequences. A resampling test[16] indicates that these sequences represent approximately 75% of the total number of sequences present at Xq13.3.

When a network that depicts all equally parsimonious mutational pathways connecting the sequences is constructed (Fig. 3), only one case of either a recurrence of a substitution or a recombination event needs to be invoked. Therefore, a substitutional model that assumes infinite sites[17] represents a reasonable approximation of the evolution of this sequence. To estimate the rate of evolution of Xq13.3 and to test whether the sequences evolve at a constant rate 'clock-like', we determined the orthologous sequence in a chimpanzee and a gorilla, which were found to differ by an average of 94.0 and 143.0 substitutions, respectively, from the human sequences. Assuming a divergence between chimpanzees and humans of 5 million years, a date based on both molecular[18,19] and paleontological[20] evidence, approximately 1 substitution per 100,000 years will

accumulate in these sequences. A test[21] comparing the likelihoods of trees reconstructed with and without a clock-assumption indicated that the sequences evolve clock-like (*P*=0.66).

Assuming a worldwide panmictic population and no selection, a test using the relationship between the number of variable positions and mean pairwise sequence difference was performed to determine if the data were compatible with a constant population size[22]. The result (D=−1.61) shows that the assumption of constant population size cannot be rejected. This may seem unexpected in view of the fact that both mtDNA (ref. 23) and microsatellite[24] data indicate population expansion within the last 50,000 years. Because the mutation rates of both mtDNA and microsatellites are high, however, they carry information of more recent demographic events, whereas the lower mutation rate of Xq13.3 will cause its variation to reflect older events[19]. Assuming constant population size, a coalescent approach[25,26] was used to identify the location of the root in the unrooted tree relating the sequences and to estimate the time to the MRCA for Xq13.3. Sequence A was found to be the most likely root (Fig. 3). It contributes 40% of the sum of the likelihoods of roots at all positions in the tree, consistent with the fact that sequence A carries the most common nucleotide at each sequence position. Four additional roots have likelihoods higher than 5%. These are located

Fig. 3 Phylogenetic network relating Xq13.3 sequences. Letters refer to Fig. 2. Red circles represent sequences observed only in Africa, yellow circles represent sequences observed only outside Africa and blue circles represent sequences observed in Africa as well as in other regions of the world. The area of the circles represents the number of individuals carrying each sequence. The location of the most likely root is indicated by a green arrow. The dotted line and open circle indicate an equally parsimonious mutational pathway not seen in the maximum likelihood tree.

on the branches to sequences T and O. In addition, a phylogenetic analysis with a maximum likelihood approach[21] also identified sequence A as the root when either the chimpanzee and/or the gorilla sequences were used as outgroups. The coalescent analysis allows the human effective population size to be estimated at 16,000 and the age of the MRCA to 535,000±119,000 years. Because the effective population size of the X chromosome is three times that of mtDNA and the Y chromosome and three-quarters that of autosomes, this age agrees well with the estimates of the age of the MRCAs for mtDNA and Y chromosomal sequences, which fall 100,000–200,000 years ago[1,2,27], and for β-globin gene sequences, which are 750,000±210,000 years[3].

The resulting phylogenetic tree reveals genetic patterns of early human history. Sequence A, which is likely to be ancestral to all other sequences in the tree, is found in four Africans, two Aboriginal Australians, one Chukchi and one Georgian. Thus, in addition to Africans, Aboriginal Australians as well as other Asians carry ancient sequences, an observation consistent with other work[3,28]. Africans, however, are more widely distributed in the tree than non-Africans. Focusing on the nine branches originating from the ancestral sequence A (Fig. 3), one or more Africans are represented in all nine branches, whereas non-Africans are found in four. The wider distribution of African sequences is reflected in higher genetic diversity. For example, the mean pairwise sequence difference among the Africans (3.6±2.0 differences per 10,163 bp) is almost as large as that of the entire data set (3.7±2.0), whereas it tends to be lower in non-Africans (3.1±1.9). Moreover, 24 of 33 variable positions were found in Africa, whereas 17 were found in the rest of the world, in spite of the fact that more than twice as many non-Africans (n=47) than Africans (n=22) were sequenced. The age of the MRCA excludes a divergence of one million years or more (*P*=0.0028), a time depth expected if genetic continuity between *Homo erectus* and modern human populations had occurred for this locus[29] in several regions of the world. On balance, the overall pattern is most easily reconciled with an African origin for the sequence variation observed at Xq13.3, even if other scenarios cannot be excluded. The older sequences that exist in Asia warrant further work in light of recent suggestions of an early coastal migration from the Horn of Africa to Australia approximately 120,000 years ago[30]. Another notable feature of the network is that a large proportion of individuals from Asia carry

sequence J, something that requires further investigation with regard to its possible association with past demographic events, such as the spread of rice agriculture.

On a general note, loci such as Xq13.3 represent a class of nuclear loci that are very useful for evolutionary analyses. Since their substitution and recombination rates are low, their history within humans can be reconstructed with less ambiguity than for mtDNA, where the mutation rate results in multiple substitutions, or for nuclear loci, where recombination rates are higher. Furthermore, because they can be chosen to be located far away from genes, the distorting effects of selection are minimized. As the general problem that selection may affect any particular locus cannot be excluded, several such loci need to be studied to arrive at a comprehensive picture of the origin and evolutionary history of the human gene pool.

Methods

for 40 s, 61 °C for 1 min and 68 °C for 40 s. Sequencing reactions (5 μl) were run on Alf-Express sequencers (Pharmacia Biotech) using 5.5% Long-Ranger (FMC Bio Products) gels. Running conditions were as recommended by the supplier.

Analyses. AlfWIN software (Pharmacia) was used for base-calling. Sequences and trace-data were transferred to the SEQMAN program (DNASTAR), which was used for sequence assembly. SEQMAN was also used for the final alignment of the complete sequences (10,163 bp) and subsequent identification of variable nucleotide positions. The program AvH2 (unpublished data) was used for calculation of mean pairwise sequence differences. The clock test was performed using the program PUZZLE 4.0 (ref. 21). The time to the MRCA was estimated by the Genetree package[26] (http://www.maths.monash.edu.au/˜mbahlo/mpg/gtree.html), excluding variable position 11. The time back to the MRCA was calculated assuming a global panmictic population and the maximum likelihood value for θ. The root depicted was determined using the Genetree package and the human sequences. The location of the root was confirmed phylogenetically, assuming a TN model with γ-distributed evolutionary rates as implemented in PUZZLE 4.0 (ref. 21). The absence of putative exons in the

sequence studied was confirmed using the Grail program (http://comp-bio.ornl.gov/manuals/grail-genquest.9407.shtml).

Accession numbers. PAC-clone sequence 333E23, GenBank Z82200; 69 human sequences (order as in Fig. 2), and chimpanzee and gorilla sequences, EBI AJ241023–AJ241093.

Note added in proof: One Navajo representing the single language phylum missing in Fig. 1 (Na-DeNe) has now been sequenced. He carries sequence G.

1. Stoneking, M. Recent African origin of human mitochondrial DNA: review of the current status of the hypothesis. in *Progress in Population Genetics and Human Evolution* (eds Donnely, P. & Tavaré, S.) 1–13 (Springer, New York, 1997).
2. Hammer, M.F. *et al.* The geographic distribution of human Y chromosome variation. *Genetics* **145**, 787–805 (1997).
3. Harding, R.M. *et al.* Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* **60**, 772–789 (1997).
4. Nickerson, D.A. *et al.* DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
5. Harris, E.E. & Hey, J. X chromosome evidence for ancient human histories. *Proc. Natl Acad. Sci. USA* **96**, 3320–3324 (1999).
6. Underhill, P.A. *et al.* Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography (DHPLC). *Genome Res.* **7**, 996–1005 (1997).
7. Weatherall, D.J. & Clegg, J.B. *The Thalassaemia Syndromes* (Blackwell Scientific, Oxford, 1981).
8. Brunzell, J. Familial lipoprotein lipase deficiency and other causes of the chylomicronemia syndrome. in *The Metabolic and Molecular Bases of Inherited Disease* (eds Scriver, C., Beaudet, A., Sly, W. & Valle, D.) 1913–1933 (Mc Graw-Hill, New York, 1995).
9. Robinson, B.H., MacKay, N., Chun, K. & Ling, M. Disorders of pyruvate carboxylase and the pyruvate dehydrogenase complex. *J. Inherit. Metab. Dis.* **19**, 452–462 (1996).
10. Nagaraja, R. *et al.* X Chromosome map at 75-kb sts resolution, revealing extremes of recombination and GC content. *Genome Res.* **7**, 210–222 (1997).
11. Charlesworth, B., Morgan, M.T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
12. Nachman, M.W., Bauer, V.L., Crowell, S.L. & Aquadro, C.F. DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**, 1133–1141 (1998).
13. Report of the sixth international workshop on X chromosome mapping. *Cytogenet. Cell Genet.* **71**, 308–342 (1995).
14. Satta, Y., Li Y. & Takahata, N. The neutral theory and natural selection in the HLA region. *Front. Biosci.* **3**, 459–467 (1998).
15. Ruhlen, M.A. *Guide to the World´s Languages* (Edward Arnold, Kent, 1991).
16. Ward, R.H., Redd, A., Valencia, D., Frazier, B. & Pääbo, S. Genetic and linguistic differentiation in the Americas. *Proc. Natl Acad. Sci. USA* **90**, 10663–10667 (1993).
17. Watterson, G.A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
18. Kumar, S. & Hedges, B. A molecular timescale for vertebrate evolution. *Nature* **392**, 917–919 (1998).
19. Takahata, N. A genetic perspective on the origin and history of humans. *Annu. Rev. Ecol. Syst.* **26**, 342–372 (1995).
20. Andrews, P. Evolution and environment in the Hominoidea. *Nature* **360**, 641–646 (1992).
21. Strimmer, K.S. & von Haeseler, A. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964–969 (1996).
22. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
23. Sherry, S.T. *et al.* Mismatch distributions of mtDNA reveal recent human population expansions. *Hum. Biol.* **66**, 761–775 (1994).
24. Kimmel, M. *et al.* Signatures of population expansion in microsatellite repeat data. *Genetics* **148**, 1921–1930 (1998).
25. Griffith, R.C. & Tavaré, S. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* **46**, 131–159 (1994).
26. Griffith, R.C. & Tavaré, S. Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosci.* **127**, 77–98 (1995).
27. Weiss, G. & von Haeseler, A. Estimating the age of the common ancestor of men from the ZFY intron. *Science* **272**, 1358–1360 (1996).
28. Stoneking, M. *et al.* Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**, 1061–1071 (1997).
29. Wolpoff, M. & Caspari, R. *Race and Human Evolution* (Simon & Schuster, New York, 1997).
30. Foley, R. The context of human genetic evolution. *Genome Res.* **8**, 339–347 (1998).
31. Wilson, J. & Erlandsson, R. Sexing of human and other primate DNA. *Biol. Chem.* **379**, 1287–1288 (1998).