Phylogenetic Network for European mtDNA

The sequence in the first hypervariable segment (HVS-I) f the control region f as a source of evo**lutionary information in most phylogenetic analyses of mtDNA. Population genetic inference would benefit from a better understanding of the variation in the mtDNA coding region, but, thus far, complete mtDNA sequences have been rare. We determined the nucleotide sequence in the coding region of mtDNA from 121 Finns, by conformation-sensitive gel electrophoresis and subsequent sequencing and by direct sequencing of the D loop. Furthermore, 71 sequences from our previous reports were included, so that the samples represented all the mtDNA haplogroups the Finnish population. We found a total of 297 a.a.** f' is the conding region, which in the coding region, which is the c **allowed the compilation of unambiguous phylogenetic networks. The D loop harbored 104 variable sites, and, in t** a strain $\frac{1}{2}$ is the could be localized with the coding-region networks, with discrepancies of the coding-region networks, with the coding-region $\frac{1}{2}$ is the coding-region of the coding-region of the coding**homoglasies were detected in the coding region.** Nucleotide variation in the region in the reg **that in the third nucleotide positions of structural genes amounted to 22% of that in the HVS-I. The complete networks enabled the relationships between the mtDNA haplogroups to be analyzed. Phylogenetic networks based on the entire coding-region sequence in model provide a** rich source for t , t for t , t is $\frac{d}{dt}$, $a \quad \dot{s} \quad a \quad \dot{s} \quad \text{iff} \quad \dot{s} \quad \dot{a} \quad \dot{c} \quad \text{if} \quad \dot{c} \quad \dot{c} \quad \text{if} \quad$

Introduction

M $\mathcal{N} = e e^{\int \rho} e^{-\rho} \int \rho^2 N A \int_{\mathbf{a}} \rho e^{\mathbf{b}}$ een based on $\frac{1}{2}$ or $\frac{1}{2}$ in the first hypersurface segments of $\frac{1}{2}$ is the first hypersurface segments of $\frac{1}{2}$ $(H-I)$ fecentrely (Richards eq. 1998; Ma**a** $G = 299$, B $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is a $G = 1990$; N
i le $G = 1999$, B $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is a $G = 1970$; N
i le $G = 161$
i le $G = 161$ if $e \oplus f$ is the substitution of the evolution \mathbb{R}^n $\int H -Ic$, $\int e^{i\theta}$ e ϕ ject to a_1 and a_2 is the some sites at some sites at some sites at some sites and some sites and some sites and some sites and some sites at some sites at a_1 e^{c} . It at a control information at other sites $a - e$ $f_{\rm eff}$ and to $f_{\rm eff}$ either low $f_{\rm eff}$ $\begin{array}{ccc}\n\mathbf{a} & \mathbf{a} & \mathbf{c} \\
\mathbf{i} & \mathbf{a} & \mathbf{j} & \mathbf{c} \\
\mathbf{E} & \mathbf{i} & \mathbf{j} & \mathbf{c} \\
\mathbf{E} & \mathbf{j} & \mathbf{k} & \mathbf{c} \\
\mathbf{DNA} & \mathbf{a} & \mathbf{c} & \mathbf{j} \\
\mathbf{A} & \mathbf{E} & \mathbf{G} & \mathbf{j} & \mathbf{k} \\
\mathbf{k} & \mathbf{i} & \mathbf{k} & \mathbf{k} \\
\mathbf{k} & \mathbf{i} & \mathbf{k} & \mathbf{k} \\
\mathbf{k$ FL_{L} l rec^{on} el r_{a} bee Ge^r_l regio DNA $_{a}$ \sim $\frac{G}{H}$ (\sim 1 e $_{a}$ \sim 1996). The et be e^{3} EG \sim e^{3} in EG \sim e^{3} $f \circ f$ \downarrow DNA \downarrow ₂, \downarrow G₁, c₁G₁ e : H['], ^K, \downarrow J, and \mathbb{F}_{\cdot} is most accurate physical phy represents the DNA $a \neq 0$ been constructed by Ge figures by Ge figures $a \rightarrow a$ Ge ce a_a from H -I ($1c_a$ ³ e $a = 1298$), and Gmented with $\frac{1}{2}$ and $\frac{1}{2}$ FL analyses of the coding $\frac{1}{2}$ e l $(M_{\rm A} G_{\rm A} G_{\rm B}$ e a $\sim 1999)$ and from H -II e-Ge ce $(He_{\text{eq}}$ e $a = 2000)$. \degree e Ge \degree H -I se- \int ce $\frac{1}{4}$ a \int ₂ \int ³ de³ led the inconsistent definition of the inconsistent defi

ecel e^2 Dece be 28_h 2000; accepted for publication April 3, 2001; e.ec $\frac{1}{9}$, $\frac{1}{9}$, $\frac{1}{9}$, e^* M_a 10, 2001.

 A^{ss} e for e, se^{ss} e cears e i \pm D. Kari Majamaa, Uni- $\ell \in \{O(\frac{1}{2}) \log_2 e \quad e \quad \text{NeG} \quad \text{O.B. } 5000, \text{FIN-90014},$ $U_{\ell}e$ i $\int O(\frac{C}{\epsilon})$ $O(\frac{C}{\epsilon})$ \int $\frac{C}{\epsilon}$ \int \frac \degree 2001 by The American Society of HG and Genetic And The e.g. 0002-9297/2001/6806-0019\$02.00

 h_a , θ gocg e, since some of the happy subset of the happ $\frac{1}{\sqrt{6}}$ $\frac{1}{\sqrt{6}}$ eq. $\frac{1}{2}$ bee, ibe βc $\frac{1}{2}$ e a $\frac{1}{2}$

 \vec{F} \vec{F}

Figure 1 Physical Late ariation in the coding of the coding sequence in the coding sequence in the coding sequence in \mathbb{R} or G is meaning in \mathbb{R} or G is much in the outgoing is much in \mathbb{R} or G is me from African individual (Ingless 2000; GenBank accession of the AF346980). NG best from the nodes denotes denote samples. Unless the nodes of the nodes of the nodes of the nodes denotes denote samples. Unless the nodes of marked otherwise, shown on the polymorphic variants, shown on the nodes, and the nodes, and the nodes, are transversions or the nodes, and transversions or the nodes, and the nodes, are transversions in the nodes, and the inserted nucleotides. Position 10398 $_{\rm g}$ 3 -dig equal. Applied in the control of the Geodiec positions were equal. Applied to the nucleotide positions were equal. Applied by $_{\rm g}$ in the nucleotide positions were $e^{i\alpha}$ α β at position 15884 it is α as resolved by α in samples α in samples α and α and α following substitutions were commonly substitutions and $\frac{1}{4}$, $14\frac{3}{6}$ SG \rightarrow C, 1436 SG \rightarrow C, and 106 e C. i p in the 1 in the CO ii $\frac{1}{2}$ e compared in the CO II-NAL intervention in the CO II-taggeral mutation in the Co ii $\frac{1}{2}$ in the CO II-NAL intervention; $\frac{1}{2}$ $=$ lee, ic G_n i.

 \leftarrow G. Ieel \rightarrow elee \rightarrow ic G. i 7706 $\ddot{G} \rightarrow A$ i CO^X II₂ ³ 14696A \rightarrow G i NA^GG e e c $e^{-i\theta}e^{i\theta}$ ie W. G . $H_a = G I c + e^2$ (bc $G e$; subclusters; subc 11 m^3 see" i es b 3990C-, 6734G-A, 8616G-9053G-A, 9947G-A, and 10915 \rightarrow C and included included included to $G^{\bullet}e^{\bullet}$ 62% fearse and 62% characterized 5% b $\text{Re }c^{\frac{a}{2}}$ - ei, \Box \Box 15758A \rightarrow G_a $\frac{a}{2}$ iclude^s 38% fe_{ar}, e. O of G a , e belonges $T_a = G^*$. $e^{\int e^a e_a e_b e_b}$. $\hspace{0.1cm} = \hspace{0.1cm} \mathcal{G}_1 Z c \mathfrak{g} = \int_0^\infty e^{\mathfrak{g}} a \cdot \mathfrak{g} \, \mathrm{d} \mathfrak{e} = \mathcal{G} + \int e^{\mathfrak{g}} a \cdot e \, \mathfrak{g} =$ $\operatorname{Ge}^{\bullet_1}$ e, ee $\operatorname{G}^{\bullet}$, it 18 $\operatorname{G}^{\bullet}$ $\operatorname{G}^{\bullet}$, it in the coding region $($ $($ $, 1)$. $H_a = \sqrt{G}$ and K (Find in eq. 2000, and in press) and $\mathcal{G} = \mathcal{G} - a$ and $\mathcal{J}(\mathbb{R}^{-1})$ and $\mathcal{G} = \mathcal{M}$ and a 2001 eeacc a^{s sex}t e $-$ e \mathcal{M}_L in the coding region (fig. 1) ielded a to $p = \frac{1}{4}$ showed distinct compared that \mathcal{G} H, $K_{\nu,+}$ J, and \mathbb{F}_a and \mathbb{F}_a \mathbb{F}_a \mathbb{F}_a and \mathbb{F}_a concordance u^{i} . The EG μ β - DNA tee base x^{i} and i^{i} $\text{Re } H = I \left(\begin{array}{cc} 1 & c_{a} \end{array} \right)$ e a \rightarrow 1998). Specifie complex of e $p = \mu^{11}$ eve^{s s}ee e_1 eq. e_2 eq. a_n , f_n is equal on the matter of the matter on the matter of th $te \alpha + \beta e \alpha$ and α is the status at position 12705, so that α \leftarrow G, H, , , K, and J harbored 12705C compared 128 $h_{a} = \int_{a}^{b}$ I, $\frac{1}{2}$, $\frac{1}{2}$ and $\frac{1}{2}$ and $\frac{1}{2}$ and $\frac{1}{2}$ and $\frac{1}{2}$ and $\frac{1}{2}$

Comparison of Networks Based on Variation in the Coding Region and the D Loop

 $\int \text{ce}_b \text{e}_b$, $\text{e}_b \text{e}_b \text{e}_c$ or $\text{e}_b \text{e}_b$ obtained for Eq. , e_b rely on the H_V-coding-interval comare it is a based on the variation in the $D_{\mathbf{r}}$. (e. 2). Networks were $f(x+h)$ constructed for each $f(x)$. \overline{u} $\overline{$ $\frac{1}{2}$ feed $\frac{1}{2}$ to each $\frac{1}{2}$ to each $\frac{1}{2}$ \mathbf{u}^{\dagger} eee, \mathbf{b} eba \mathbf{G}^{\dagger} ee \mathbf{b} ac \mathbf{G}^{\dagger} + the leg $\int e \cdot e^{-x} dx = \int e^{x} \cdot e$. Google con $f^*e = i^*e i_e e^*i$ is $D-$, $e^+e^-e^-i$, $i^*e^$ a few exceptions: β for β exceptions α in the two subsets of α in the subsets of β and β \oint_C c \oint_C ere partly overlapping, but partly overlapping \int_C but partly overlapping \int_C but partly overlapping \int_C but \int e le \int ee \int a_k \int e concordance between the β subclusters detected in the network based on the network based on the coding $e^{\int_0^t e^{i\theta} \cdot e^{i\theta} \cdot e^{i\theta}}$ is the network based in the network based of $e^{i\theta}$ \mathfrak{e} b \mathfrak{e} b \mathfrak{e} ce.

Homoplasies in the Coding Region The $a = G$ be it $a = G$ is not homoply the set of $a = T e$ 192 DNAs 2 70 (a. be $3a - 4$). In the coding region, 21, 11 eel $\rightarrow \pm 1$ c; 7 flee were $\mathcal{G}, 10$ ee \mathcal{G}_A 6 ee te \mathbb{G}_{p} + \mathbb{L} . H \rightarrow \mathbb{R} + c + e c + $\mathbb{G}e^{2}$ 0.14% of the ie i lec^odei_a ³4.4% of the leD \longrightarrow , \int e_1 \downarrow e_2 e_3 \downarrow f 1:31. $M \cap \{e\}$, here \mathcal{C} or \mathcal{C} or \mathcal{C} in different happens in different happens in different happens in different \mathcal{C} \leftarrow G, O \leftarrow (9.5%) feiei tec in

Discussion

Topology of mtDNA Phylogenetic Networks C , eee e dee f leles \mathfrak{g}_a m DNA \mathfrak{g}_a e j \mathfrak{g}_b be G e e e (H $\frac{1}{2}$ e a -2000; I a e a -2000; Elson et al. 2001). In the present article, we have e e i este DNA e Ge cef 121 Finns, and e complete the G ecent data (Findigate M_k and α $2001; \overline{P}$ in eq., i.e., eegbe c Gc $a \cdot b = e$ ecice $b = e^a$ complete DNA e- \mathfrak{g} cef 192 Fi $\,$, at lite $\,$ e ef $f_{\mathbf{Gq}}$ DNA e Ge ce e^s $f_{\mathbf{q}q}$, e f $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ in the multiple in the multiple in the multiple in the population of \mathbb{R} . e e inc G^*e^s . Certain nuclear nuclear markers (Capallinus) of G eq -1994) and markers on the ϵ of the ϵ on the Y chromosome have have ϵ \mathbf{S}_1 e e³ \mathbf{S}_2 in e \mathbf{R}_1 are \mathbf{G}_1 be in Eq. , e (signal $t_{\rm r}$ e a - 1996; Ze i - e a - 1997), bG ℓ e G be, $\text{DNA} + \mathbf{q}_1$ e_l \mathbf{q}_2 e^s be ee e Finns e EG, $g \perp \mathbb{N}$, G e in f_a ferming, e ce EG, a \perp a e $(1 - 1e_2 - 1988; a_4)$

Figure 2 Physical on on DNA, based on variation in the D-loop second on G-loop on a African in the G-loop sequence. The outgroup is made in the G-loop sequence. The outgoin in the outgroup is made in the outgroup in the (Ingman et al. 2000; GenBank accession Griege AF346980). Executive inclusion in the notation of intervals of $i = 1$ incit: e_i is deed to p a p deletion; the supersions of the superscripts indicate transversions of e_i in e_i in e_i is defined on e_i in e_i in e_i is defined in e_i in e_i is defined in e_i in e_i is defined ee te e e Ge 1.

Find \mathbf{F}_k e a \div and \mathbf{F}_k in Eq. e and DNA 1481

Table 3

Parallel Mutations Detected in the Coding Region of mtDNA in 192 Finnish Samples

		AMINO ACID ION В Section	LOG ПA.				
ION ິ	GENE		H	\sim $\overline{11}$			-

Parallel Mutations Detected in the D Loop of mtDNA in 192 Finnish Samples

Table 4

NO_{TE} $f \text{ e } f = i \int_{\Omega} a \cdot e \cdot e \quad e \text{ e } i \int_{\Omega} e^{i \theta} i \int_{\Omega} e \cdot i \int_{\Omega} b \cdot e \cdot i \int_{$ \overline{G} H, 16166A-C in \overline{G} J, 16166⁸e A in Subcluster 4, 16129G-C in Subcluster U4, 16129G Q_2 , 16183A→G in $a_2 = Q_1$, 456⁸e C_2 \bullet 456C $\frac{1}{2}$ in $a_2 = Q_1$ H, 514insCA in $h_a = G$ H, 514 ^se CA in Subcluster U5, 568insC in subcluster U5, 568insCCCC, 568insCCCCCC in \mathcal{G} I. H, e.g. in he, i.i. $303a$ and 16519 were also excluded. \mathbb{N} a $\frac{1}{4}$ is a in a be $\frac{3}{4}$. $^{\circ}$ Has shared ancestration in happy in happy \mathbb{Z} but occurs as a parallel mutation in \mathbb{Z} in interval mutation in \mathbb{Z} $h_{a} = \frac{G}{2} - a$ c_H \mathbf{A}_a c^2 and \mathbf{A}_a in the set of \mathbf{A}_b , \mathbf{A}_b , \mathbf{A}_b , \mathbf{A}_b , \mathbf{A}_b \mathbf{A}_c \mathbf{A}_c and \mathbf{A}_c is a parallel $\mathcal{G}_a + \mathcal{G}_a = \mathcal{G}$ H.

A, Gebit KD (1997) ρ G is elequating general subsets and discrete subse ce $16\perp$: characterization central EG, g_1 , g_2 G by DNA ee G_2 i, correlation \mathbb{R}^2 α b α α α d α in the e. HG Mole 6: 1835–1846

- Ing M, Kaessmann H, aa b, Give e (2000) Mid $\frac{1}{2}$ e early a tell $\frac{1}{2}$ a N_a Ge 408:708, 713
- M_a G_a \rightarrow C_a M , Hicke E, eq. E, C G_a i F, G_1 v_2 C $\mathbf{v_2}$ + \mathbf{R} , B $\mathbf{e}_{\mathbf{r},\mathbf{a}}$ + \mathbf{B} , $\mathbf{e}_{\mathbf{r},\mathbf{b}}$ + $\mathbf{A}_{\mathbf{a}}(1999)$ The eilee of e $\mathsf{E}\mathsf{G}_2$ in DNAs: set for $\tilde{c} = e^{\frac{1}{2}} e^{\frac{c}{2}}$ e Genet e Genet e Genet e Genet 64: 232–249
- Meinign M, Finnight Majamaa K. E. \mathbb{R}^3 e ce for \mathbb{N} NA and \mathbb{R}^3 i Gebe ee le H_a Fe _{aa} i HG Hee^s (in $e \left(\begin{array}{cc} 1 \ 1 \end{array} \right)$
- e $\text{e G}, G$ i C, De Chic A, a cc e C (1999) NG c e i^3 e substitution rate of mammalian mitochondrial genomes. J $\rm M = E / \sim 48:427$, 434
- R_{α} ³ MB, M_{α} \mathcal{G}_{α} $\mathcal{G$ $P = e_{a}$, $\int f d^{a} g_{a}$ DNA in ee EG, e. A HG Ge e 62:241 260
- \mathbf{A} , \mathbf{A} , \mathbf{A} \mathbf{B} \mathbf{C} K, Gei C, \mathbf{A} b (1996) and \mathbf{a} and \mathbf{a} and \mathbf{a} a bot-DNA lineage decay a bot- $\texttt{etc.}$ in the founding of the Finnish population. \mathbb{C} N_{atl}ing \mathbb{C} N_{atlin}g population. $Ag = d \quad A \quad 93:12035 \quad 12039$
- $\mathbf{i} \cdot \mathbf{G} \mathbf{Q}$ \mathbf{A} the A, $\mathbf{G} \mathbf{G}$ central \mathbf{A} , $\mathbf{G} \mathbf{G}$ and \mathbf{G} and \mathbf{H} .

e (2000) e $G_1 + a e_1$ re G_2 DNA c $-e^{i}$. A J HG Gee 66:1599. 1609

- \mathbf{A}, \mathbf{B} is a contributed HJ, D´urbano L, Lahermo P, Ma e \perp D, e C, F e , and a G ML, B $.e.$ B, $c \bullet_{\mathbb{Z}} 1$ (1998) $DNA_{\mathbf{a} \cdot \mathbf{a}} = 1$ e.g. $-\mathbf{a} \cdot \mathbf{a}$ i.e. $p = p$ lithic, $p = p$ is equal to p equal to p equal to p **q** e $E_q^{\tau'}$, e A J H_um Ge e 62:1137 1152
- $T(A, H_G^G, T e K, F_a \hat{q}_\bullet \hat{cd}$, $e \bullet \text{rel } M, M \hat{cd}$ L, c \mathbf{v}_2 , i., ObingD, \mathbf{q} , \mathbf{v}_2 and \mathbf{q} , \mathbf{p} and \mathbf{p} (1996) C_{m} ight f EG on α DNAs from an analysis of C_{m} the EG, $\frac{1}{2}$, $\frac{1}{2}$, $\frac{1}{2}$. Genetic 144:1835, 1850
- $V = \iint_R a \cdot r \cdot a \cdot G$ ML, $N = \iint_R e$ EK (1988) H G_R id $\frac{1}{2}$ DNA e in Find Huggers 30:
- 317–321 a_1 E, Fe, ichards M, Bandele HJ (1997) Mi-
- chondrial footprints of $G_{\mathbf{a}}$, expansions in Africa. Am J HG Genet 61:691 704 \cdot – \cdot – \cdot M, \cdot – \cdot M, \cdot e \cdot M, \cdot e \cdot
- ere E, d'el^{ss}e C, e e M, \equiv e M, Fie $^{\prime}$ (1994) Convenient single-step, one of \mathbb{G} , \mathbb{G} purification of PC, \mathbb{F}_2 \mathbf{G} f³ \mathbf{u} ece \mathbf{G} ed. N \mathbf{G} ec \mathbf{A} ds e 22:4354–4355 Ze' \mathbf{i} \mathbf{D} \mathbf{i} \mathbf{a} B, \mathbf{a} \mathbf{a} A, \mathbf{K} e M, e e $\mathbf{L}_{\mathbf{i}}$ \mathbf{a}
- F , chefenhovel, Free ∞ , Jobling MA, Harihara S, \mathbb{R} I I \mathbb{G} K, e jihaa D, a jihaa \mathbb{R} a \mathbb{R} , \mathbb{R} a \mathbb{R} , Ca i P, MH, Gine_l EK, E_{sta}r G₁O₂, $\frac{1}{2}$, $\frac{1}{2}$ C (1997) Genetic e- \mathbf{r} ii, \mathbf{A} ians \mathbf{a} is Europeans, revealed by \mathscr{L} and \mathscr{L} and 1174–1183