2008

The Phylogeny of the Four Pan-American MtDNA Haplogroups: Implications for Evolutionary and Disease Studies

Alessandro Achilli
Università di Pavia

Ugo A. Perego
Università di Pavia

Claudio M. Bravi
Instituto Multidisciplinario de Biologıá Celular (IMBICE)

Michael D. Coble
Armed Forces Institute of Pathology

Qing-Peng Kong
Kunming Institute of Zoology

See next page for additional authors

OPEN ACCESS Freely available online

PLoS one

The Phylogeny of the Four Pan-American MtDNA Haplogroups: Implications for Evolutionary and Disease Studies

Alessandro Achilli[1,2], Ugo A. Perego[1,3], Claudio M. Bravi[4], Michael D. Coble[5], Qing-Peng Kong[6,7], Scott R. Woodward[3], Antonio Salas[8], Antonio Torroni[1]*, Hans-Jürgen Bandelt[9]

1 Dipartimento di Genetica e Microbiologia, Università di Pavia, Pavia, Italy, **2** Dipartimento di Biologia Cellulare e Ambientale, Università degli Studi di Perugia, Perugia, Italy, **3** Sorenson Molecular Genealogy Foundation, Salt Lake City, Utah, United States of America, **4** Laboratorio de Genética Molecular Poblacional, Instituto Multidisciplinario de Biología Celular (IMBICE), La Plata, Argentina, **5** Armed Forces DNA Identification Laboratory, Armed Forces Institute of Pathology, Rockville, Maryland, United States of America, **6** Laboratory of Cellular and Molecular Evolution, and Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, China, **7** Laboratory for Conservation and Utilization of Bio-resource, Yunnan University, Kunming, China, **8** Unidade de Xenética, Instituto de Medicina Legal, Facultad de Medicina, Universidad de Santiago de Compostela, Grupo de Medicina Xenómica, Hospital Clínico Universitario, Santiago de Compostela, Galicia, Spain, **9** Department of Mathematics, University of Hamburg, Hamburg, Germany

Abstract

Only a limited number of complete mitochondrial genome sequences belonging to Native American haplogroups were available until recently, which left America as the continent with the least amount of information about sequence variation of entire mitochondrial DNAs. In this study, a comprehensive overview of all available complete mitochondrial DNA (mtDNA) genomes of the four pan-American haplogroups A2, B2, C1, and D1 is provided by revising the information scattered throughout GenBank and the literature, and adding 14 novel mtDNA sequences. The phylogenies of haplogroups A2, B2, C1, and D1 reveal a large number of sub-haplogroups but suggest that the ancestral Beringian population(s) contributed only six (successful) founder haplotypes to these haplogroups. The derived clades are overall starlike with coalescence times ranging from 18,000 to 21,000 years (with one exception) using the conventional calibration. The average of about 19,000 years somewhat contrasts with the corresponding lower age of about 13,500 years that was recently proposed by employing a different calibration and estimation approach. Our estimate indicates a human entry and spread of the pan-American haplogroups into the Americas right after the peak of the Last Glacial Maximum and comfortably agrees with the undisputed ages of the earliest Paleoindians in South America. In addition, the phylogenetic approach also indicates that the pathogenic status proposed for various mtDNA mutations, which actually define branches of Native American haplogroups, was based on insufficient grounds.

Citation: Achilli A, Perego UA, Bravi CM, Coble MD, Kong Q-P, et al (2008) The Phylogeny of the Four Pan-American MtDNA Haplogroups: Implications for Evolutionary and Disease Studies. PLoS ONE 3(3): e1764. doi:10.1371/journal.pone.0001764

Editor: Vincent Macaulay, University of Glasgow, United Kingdom

Received January 9, 2008; **Accepted** February 9, 2008; **Published** March 12, 2008

Funding: This research received support from Progetti Ricerca Interesse Nazionale 2005 (Italian Ministry of the University) (to AT) and Fondazione Cariplo (to AT). Funding agencies had no role in the design and conduct of the study.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: torroni@ipvgen.unipv.it

Introduction

America was the last continent to be colonized by humans, and molecular data provided by different genetic systems [1,2] have been extensively employed to shed light on the routes and times of human arrival and dispersion into the New World. As for mitochondrial DNA (mtDNA), it has been clear, since the early nineties, that mtDNAs of Native Americans could be traced back to four major haplogroups of Asian origin shared by North, Central and South American populations [3–7]. These were initially named A, B, C and D, and are now termed A2, B2, C1 and D1 [8]. Afterwards, a fifth haplogroup – now known as X2a – was described in Native Americans, but in contrast to the four "pan-American" haplogroups, its geographic distribution is restricted to some Amerindian populations of northern North America [8–12]. Later, two more haplogroups – D2a and D3 – were identified: D2a in the Aleuts and Eskimos [13,14] and D3 only in the Eskimos [15,16]. Most recently there were two further (uncommon) additions – D4h3 and C4c [14,17] – bringing the total number of Native American haplogroups to nine.

Since the early studies, the interpretation of mtDNA data has been rather controversial with scenarios postulating one to multiple migrational events from Beringia at very different times (between 11,000 and 40,000 years ago) (for a review, see [7]). Pinpointing an accurate timeframe for the arrival of the Native American founders would be essential to solve such a debate. Yet, accurate ages can only be based on large numbers of complete mitochondrial genomes, and American mtDNA haplogroups were only poorly represented in the total database of >3000 complete mtDNA sequences until very recently. Thus, despite the protagonist role of Native American mtDNAs in high-resolution mtDNA studies 15 years ago [4], America remained the continent from which we had the least information about the sequence variation of entire mtDNAs. Worse, the available information had

to be retrieved from the web in a hit or miss fashion and suffers in part from improper documentation, oversights, and inadvertent nomenclature (Text S1). The overall situation is now beginning to change with some new data available in literature and public databases [14,18,19], but the interpretation of subsets of these data continue to remain controversial. For instance, the work by Tamm et al. [14] suggests that the Asian ancestors of the first Native Americans paused when they reached Beringia and that their (swift) migration southward might have occurred only ~13,500 years ago.

Among the novel mtDNA sequences, there are 265 from "Hispanics" and "African-Americans" that recently became available in GenBank [19]. A survey of their variation reveals that 101 mtDNAs of Native American origin were included (47 belonging to haplogroups A2, 13 to B2, 30 to C1, and 11 to D1). Those mtDNAs are not associated with either a specific Native American population/tribe or a specific geographic region but are undoubtedly of Native American origin. Furthermore, due to the fact that these are all from individuals living in the US, they probably provide a fairly good overview of the mtDNA pool of extant or extinct Native American populations from North and Central America plus the Caribbean (due to the contribution of Mexicans, Puerto Ricans, Cubans, Salvadorans, etc. to the present-day US American population), and their analysis might provide important new clues about the process of human colonization of the Americas and the origin of Native Americans. Thus, the aim of this paper is not only to (i) perform a comprehensive analysis of all available complete (or almost complete) sequences of Native American ancestry belonging to the four major pan-American haplogroups, (ii) identify their internal clades and candidate founder sequences, and (iii) estimate their expansion times into the Americas, but also to (iv) provide a framework on which future phylogeographic studies, which remain scarce, can build upon.

Results

The phylogeny of pan-American haplogroups A2, B2, C1, and D1

To define the phylogeny of A2, B2, C1, and D1 at the highest level of molecular resolution – that of complete mtDNA sequences, it is necessary to evaluate (and possibly to expand) the current data set of published mtDNA sequences in regard to reliability as well as to update and correct the nomenclature (Text S1). Figure 1 displays the roots of A2, B2, C1 and D1, together with the complete sequences belonging to the much less common Native American haplogroups C4c, D2a, D3, D4h3 and X2a [8,9,12–15,20]. Moreover, for a better discrimination from closely related Native American counterparts, some Asian (or Beringian) branches (B4b1a2, A2a, A2b, C1a, C4a, C4b, D2b, and D4h1) are illustrated. As for the phylogeny of haplogroup A2, we maintain the codes A2a and A2b for the circumpolar branches [16]. For branch A2a with the characteristic C16192T transition in HVS-I (which on its own is insufficient to identify a haplogroup because it is highly recurrent throughout the mtDNA phylogeny), coding-region information is now available revealing the additional diagnostic marker C3330T [14,18].

The complete variation of all available mtDNA sequences belonging to haplogroups A2, B2, C1, and D1 is displayed in the phylogenies of Figures 2 and 3. As for the phylogeny of A2 (Figure 2), we rename the "A2a" and "A2b" branches of Accetturo et al. [21] as A2d and A2e, maintaining the definition of A2c for the branch with the motif T12468C-G14364A. Moreover, we define six novel branches (A2f - A2k) based on all available information for haplogroup A2 (Table S1) and [20,22]. Numerous independent back mutations at nucleotide positions (nps) 64, 146, 152, 153, 16111, and 16362 are evident (that on their own do not justify support for subhaplogroup naming). Many HVS-I and HVS-II lineages from haplogroup A2 reflect this seemingly mosaic feature of instability. Some additional information on the population distribution of the subhaplogroups can also be drawn from the early high-resolution RFLP data [5,23] and an extensive database of published control-region sequences (mainly comprising HVS-I) (Text S2).

The phylogeny of haplogroup B2 (Figure 3A) reveals at least four subhaplogroups (B2a - B2d). B2a is defined by the control-region transitions C16111T and G16483A, while the sub-branch B2a1 is defined by the coding-region transition A10895G previously seen as a *Taq*I site at 10893 in haplogroup B mtDNAs from the Navajo, Ojibwa, and Pima [5]. The branches B2b and B2c are based on the presence of transitions G6755A and A7241G, respectively. B2c was also identified as a *Rsa*I site at 7241 in two mtDNA haplotypes from Mexico [23], while its sub-branch B2c1 seems to be defined by a transition at np 9098. The branch B2d (coding-region motif 4122-4123-8875-9682) is probably rather widespread in lower Central America since it was found in the Wayuùs and Ngöbes [14] and (as a *Hae*III gain at np 8872) in several other Chibchan-speaking populations [23,24].

As for haplogroup C1, all sequences appear to fall into one of the three subhaplogroups C1b, C1c, and C1d (Figure 3B). These are most likely spread all over the Americas. Indeed, the transitions at nps 493 and 16051 that define C1b and C1d, respectively, have been observed in haplogroup C1 control-region motifs from a wide range of Native American populations, including some from the southern part of South America. For C1c, which lacks basal salient HVS-I or RFLP motifs, its presence in South America is confirmed by its detection in Colombia [14] and the observation that South American C1 mtDNAs are not fully covered by subhaplogroups C1b and C1d [25], and thus the remaining C1 lineages likely belong to C1c. These findings support the scenario that C1b, C1c and C1d (and their distinguishing mutational motifs) most likely arose early – either in Beringia or at a very initial stage of the Paleoindian southward migration [14].

As for D1 (Figure 3C), the basal mutation of D1a (sequence #134) is based on the comparison with four coding-region sequences (Am02, 10, 11, 14) reported by Kivisild et al. [26]. The three additional sub-clades, D1b, D1c, and D1d have been defined by using either the novel sequences reported in this study or those from Parsons [19].

Overall, the four phylogenies appear to be quite starlike, especially the B2 and D1 trees having high indices (~0.5) of starlikeness (Table 1). In the case of haplogroup C1, the three basal branches (C1b, C1c, and C1d) are themselves starlike, with the exception of C1b where a very low index of starlikeness (influencing also C1) is mainly due to an over-sampling (10 instances) of the root haplotype of the sub-branch C1b2a (sequences #107). The significance of starlike patterns in the Native American haplogroups would be that the successful propagation event of these haplogroups and some of their major branches (in Beringia or later on the move further south) can very well be dated assuming a reliable calibration of the mtDNA mutation rate. The point estimates for the coalescence times of haplogroups A2 (without the branches A2a and A2b), B2, C1 (without the Asian branch C1a), and D1 yield 18.1 ± 1.8, 21.2 ± 2.4, 23.8 ± 4.3, and 18.6 ± 2.3 ky, respectively, based on all 219 coding-region sequences (Table 1) and by employing the calibration of 1 coding-region substitution every 5,140 years [27].

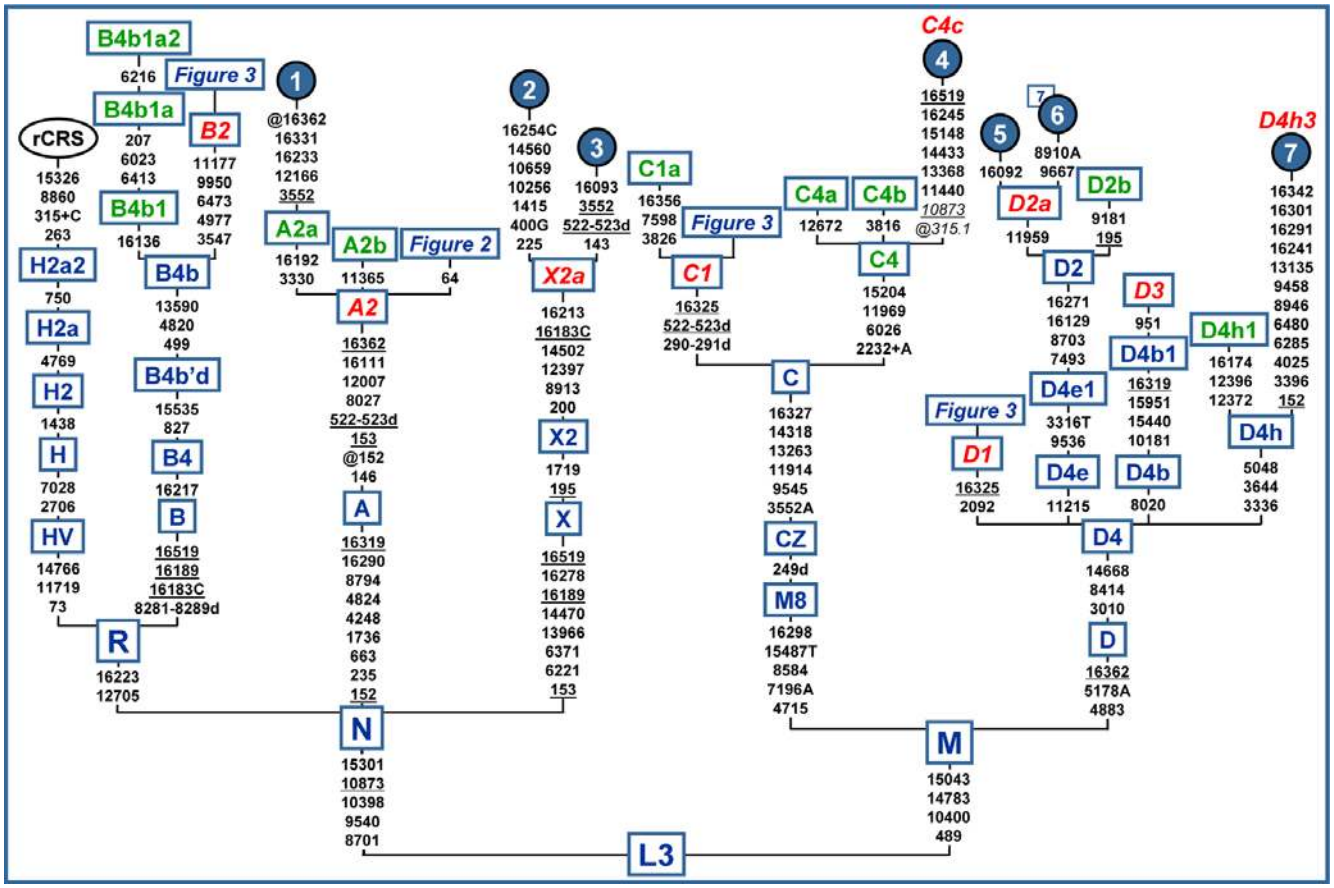

Figure 1. Basal tree encompassing the roots of Native American mtDNA haplogroups. The tree is rooted on the haplogroup L3 founder and the position of the revised Cambridge reference sequence (rCRS) [49] is indicated for reading off sequence motifs. Closely related Asian branches are indicated in green. Detailed phylogenies for the four pan-American haplogroups (A2, B2, C1, and D1, highlighted in red) are shown in the corresponding figures. The complete sequences that are currently available for the other four Native American haplogroups (X2a, C4c, D2a, and D4h3, highlighted in red) are also displayed. Haplogroup D3 is common among Inuit populations [16], but all complete sequences available are from Siberia [13,18]. As for A2a, the HVS-I motif (16111 16192 16223 16233 16290 16319 16331) of the reported sequence (no. 1) is common in Na-Dené groups [5]. Sequence no. 2 has been revised taking into account that the originally reported transitions at 4732 and 5147 [8] were artifacts due to a sample mix-up, while sequence no. 6 represents the shared motif of six Aleutian mitochondrial genomes [13]. Mutations are transitions unless specified: suffixes indicate transversions (to A, G, C, or T) or indels (+, d). Mutations back to the rCRS nucleotide are prefixed with @. Recurrent mutational events are underlined. Mutations in italics are either disease-causing or heteroplasmic or likely erroneous (and do not enter age calculations). We have followed the recent guidelines for standardization of the alignment in long C stretches [50], but disregarded any length variation in the C stretches that would then be scored at 309 or 16193 (which is often subject to considerable heteroplasmy). A number flagging a circled haplotype indicates the number of individuals sharing the corresponding haplotype (if >1). Additional information is provided in Text S4, while Table S1 lists the source of the complete genomes.
doi:10.1371/journal.pone.0001764.g001

The haplogroup ages thus fall into the range of 18–24 ky with an average of about 20.2 ky (Table 1). This value is a little bit lower (~19.0 ky) if the roots of the three branches of C1 (C1b, C1c and C1d), instead of C1 as a whole, are considered as Native American founders. This might be a (slight) underestimation because C1d is clearly under-represented in this study (comprising only eight mtDNAs). Thus, excluding C1d, the time frame is restricted to 18–21 ky and these estimates are about 1.4-fold higher than the larger time frame of 11–17 ky (A2: 13.9±2.0 ky; B2: 16.5±2.7 ky; C1b: 14.7±4.7 ky; C1c: 15.8±4.7 ky; D1: 10.8±2.0 ky) that was recently estimated [14] in a smaller dataset (105 mtDNA sequences) adopting a different calibration [26].

Detrimental mtDNA branches in Native Americans?

In some of the newly defined Native American branches, one can identify mutations for which a pathogenic role was suggested in the medical literature. The seemingly 'detrimental' status of mutations G3316A and G13708A, defining haplogroups A2f and A2e respectively, has already been questioned and discussed at length in the East Asian mtDNA context [28]. The occurrence of both mutations is not infrequent (also appearing, for instance, in haplogroups B2 and D1) and therefore, not unexpectedly, they participate in the motifs of several haplogroups. A similar case is represented by the transition T1005C, which was proposed as a primary mutation for non-syndromic hearing loss [29,30], and defines for instance the Asian haplogroup F2. In the context of Native American haplogroups, T1005C appears as a basal mutation of C1b5 – a branch of haplogroup C1b. Thus, all of these mutations are old and have been transmitted for at least some hundreds of generations. Although an effect of "old" mtDNA mutations in some multi-factorial/complex (and common) diseases cannot be ruled out *a priori*, a pathogenic role specific for such variants can, however, only be inferred from association studies in which haplogroup frequencies are properly evaluated in both patients and controls [31].

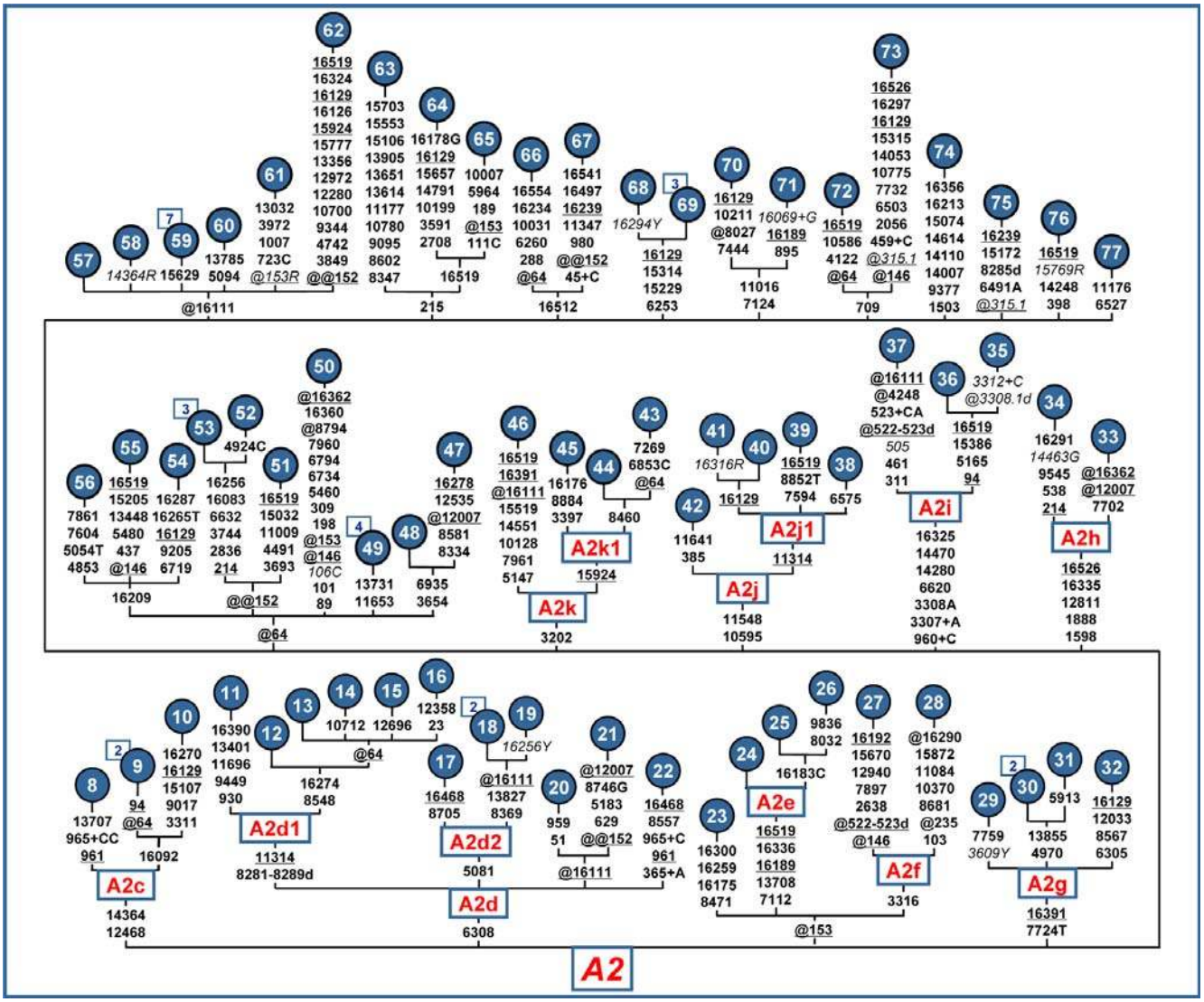

Figure 2. Phylogeny of complete mtDNA sequences belonging to haplogroup A2. The sequencing procedure for the novel complete sequences and the phylogeny construction were performed as described elsewhere [47]. Recurrent mutational events within the haplogroup are underlined, while mutations in italics are either disease-causing or heteroplasmic or likely erroneous, and were not used for age calculations. Table S1 lists the source of the complete genomes. For additional information, see the legend for Figure 1.
doi:10.1371/journal.pone.0001764.g002

An extremely interesting case of a mutational motif marking a Native American branch of the mtDNA phylogeny is represented by the T3308A transversion with a subsequent insertion of one C (3308+C) that characterize haplogroup A2i. The insertion, first reported in a patient with dystonia, leads to a frameshift mutation for which a pathogenic role was proposed [32]. However, the other mutation of the motif – the T3308A transversion – eliminates the starting codon (methionine) of the ND1 subunit by converting it to lysine, thus paralleling the scenario first described for the T3308C transition that marks the African haplogroup L1b [33]. The finding that the elimination of the methionine codon AUA at position 1 of the ND1 subunit is polymorphic in some populations clearly indicates that the maintenance of that codon is not essential in our species, and therefore the insertion of one C at 3308 does not cause a frameshift for the entire gene. This is most likely due to the fact that the third codon (AUG) of the ND1 subunit also encodes for methionine, thus despite the shortening of two amino acids, ND1 could still retain its function.

A different case is the one concerning the homoplasmic mutation T9205C detected in one mtDNA (no. 54) belonging to haplogroup A2 (Figure 2). This mutation converts the termination codon of the ATPase 6 subunit into a glutamine codon, and extends the subunit by ten amino acids (Gln-Trp-Pro-Thr-Asn-His-Met-Pro-Ile-Met) at the carboxyl-terminus. No information is available concerning the health/disease status of the subject harboring this mitochondrial genome. Thus, for the moment it cannot be ruled out that T9205C is a benign or mildly deleterious variant, despite the considerable extension of the amino acid chain. Such a scenario would parallel the situation previously reported for mutations A7443G, G7444A, and A7745C, which erase the termination codon of the COI subunit and whose pathogenic role is unclear [34].

Another illustrative case of hypothesized association between mtDNA mutations and a complex disorder is represented by the

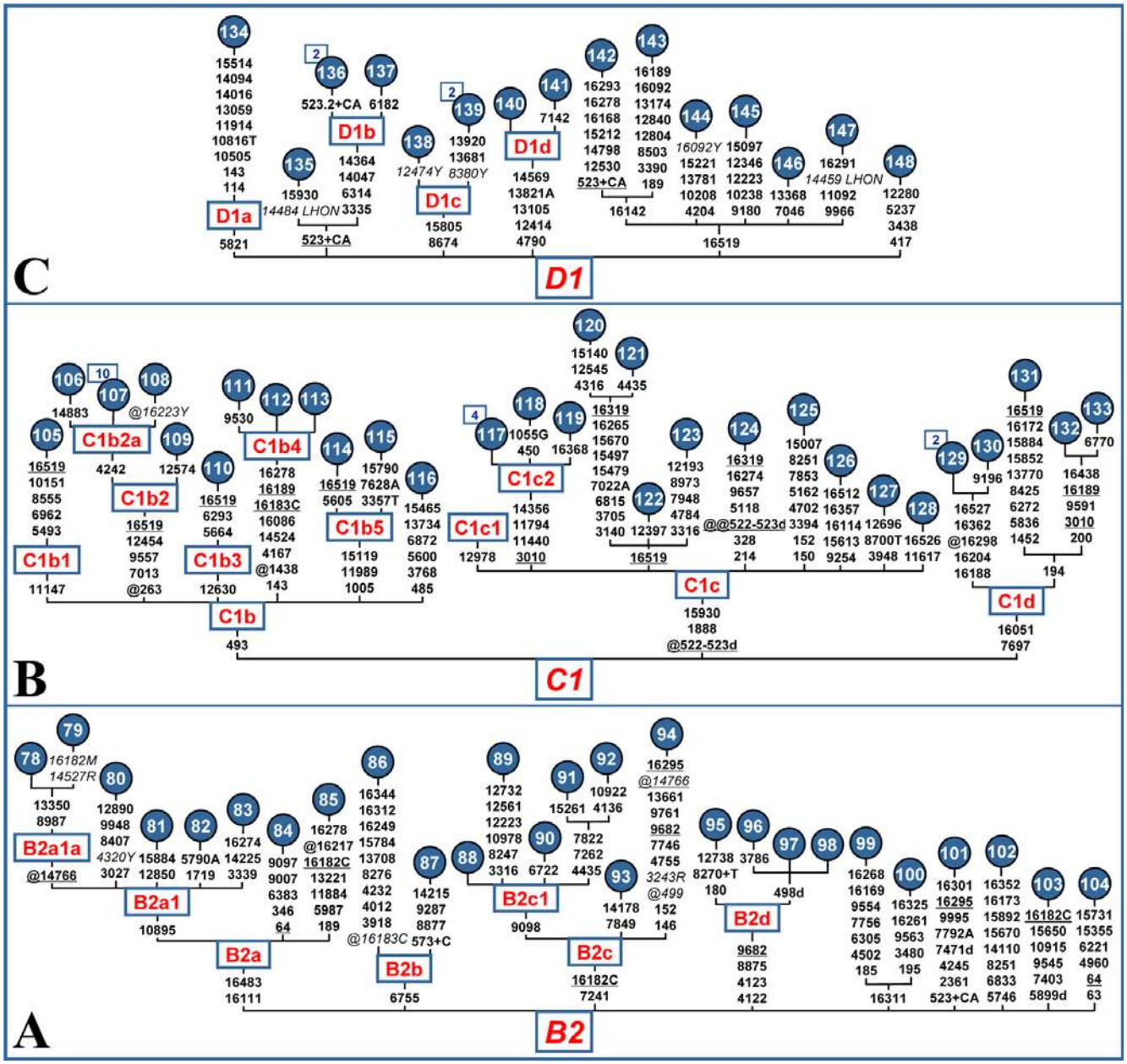

Figure 3. Phylogeny of complete mtDNA sequences belonging to haplogroups B2 (A), C1 (B) and D1 (C). For additional information, see the legends for Figures 1 and 2.
doi:10.1371/journal.pone.0001764.g003

G1888A transition which could play some role in the pathogenesis of Type 2 diabetes [35] – a scenario that would be compatible with the well-known common-disease/common-polymorphism hypothesis. This transition is characteristic of both A2h and C1c, but is also present in West Eurasia, mainly in haplogroup T [36], and in South Asia, mainly on haplogroup M5 [37]. Unfortunately, the study of [35], similar to the most recent work [38], which again implicitly targeted haplogroup T, is absolutely insufficient to shield against population substructure influencing patient cohorts and control subjects in different ways. Especially in a country such as Brazil, matrilineal population substructure matters a lot across the country [39], as well as across social strata, which often correlate with continental matrilineal ancestry. Case-control association studies that do not consider the haplogroup context in which observed mutations are embedded do not allow an objective evaluation of the role played by mtDNA variants in disease expression either, because additional variables (such as social strata and ethnicity) may influence haplogroup frequencies (Text S3) [40].

Discussion

The estimated ages (18–24 ky) of the four pan-American haplogroups A2, B2, C1, and D1 are quite similar with an average value of 20 ky. Thus, if A2, B2, C1, and D1 entered the Americas without variation in the coding region – in other words, each with only a single (successful) founder sequence (the root haplotype), their entry into the Americas would have occurred right after the peak of the Last Glacial Maximum (LGM, centered at ~21.0 kya and extending from 19.0 to at least 23.0 kya [41]), or slightly earlier, so that a coastal (Pacific) route would have been the

Table 1. Haplogroup coalescence time estimates

Haplogroup	No. (n) of mtDNAs[a]	No. of base sub-stitutions[a]	ρ[b]	σ[c]	Star-likeness[d] $\rho/(n\sigma^2)$	T (years)[e]	ΔT (years)
A2	96[f]+1	321+3	3.340	0.322	0.332	17,200	1,700
A2 (w/o A2a, A2b)	86+1	304+3	3.529	0.348	0.335	18,100	1,800
B2	27+16	116+61	4.116	0.463	0.447	21,200	2,400
C1 (w/o C1a)	42+13	198+57	4.636	0.836	0.121	23,800	4,300
C1b	21+4	86+14	4.000	1.150	0.121	20,600	5,900
C1c	15+7	63+23	3.909	0.695	0.368	20,100	3,600
C1d	6+2	13+4	2.125	0.573	0.809	10,900	2,900
D1	17+17	67+56	3.618	0.441	0.547	18,600	2,300
Total [g] (A2,B2,C1,D1)	172+47	684+177	3.932	0.311	0.186	20,200	1,600
Total [g] (A2,B2,C1b,C1c,C1d,D1)	172+47	649+161	3.699	0.274	0.225	19,000	1,400

[a]First summand refers to the complete mtDNA sequences displayed in Figures 2 and 3 and second summand refers to additional entire coding-region sequences [1–3]. Three C to G transversions (at positions 14974, 15439, and 15499) [1] – likely candidates for phantom mutations [2] that went undetected – were disregarded.
[b]The average number of base substitutions in the mtDNA coding region (between positions 577 and 16023) from the root sequence type.
[c]Standard error calculated from an estimate of the genealogy [4].
[d]Starlikeness ("effective star size" [4]) can take values between 1/n (single haplotype representing n mtDNAs) and 1 (perfect star phylogeny).
[e]Estimate of the time to the most recent common ancestor of each cluster, using an evolutionary rate estimate of $1.26\pm0.08\times10^{-8}$ base substitutions per nucleotide per year in the coding region [5], corresponding to 5,140 years per substitution in the whole coding region.
[f]This includes one Apache A2a mtDNA (#1 in Table S1) and 9 Siberian mtDNAs (four A2a and five A2b) [6,7].
[g]Without A2a and A2b mtDNAs.
doi:10.1371/journal.pone.0001764.t001

only option during such glacial periods. On the other hand, it is quite plausible that some intra-haplogroup variation – hardly noticeable at the level of HVS-I motifs – already existed in Beringia and was carried directly further south into the American double-continent. If one assumes that at least the root haplotypes of A2, B2, D1, as well as of C1b, C1c, and C1d were of Beringian origin, then the entry time would come slightly down (19.0 kya), that is, falling exactly at the end of the LGM. Moreover, the relatively lower coalescence time (~17 ky) of the entire haplogroup A2 (Table 1) – including the shared sub-arctic branches A2b (Siberians and Inuits) and A2a (Siberians, Inuits and Na-Dené) [5,14,16,18] – is probably due to secondary expansions of haplogroup A2 from Beringia long after the end of the LGM, which would have averaged the overall internal variation of haplogroup A2 in North America – the main source of the A2 mtDNAs in this study.

In any case, all the abovementioned scenarios do not support the 'Clovis-first' hypothesis, but are well in agreement with the undisputed ages of the earliest Paleoindians in South America [42]. This conclusion would not change if one adopted the effectively faster rate of Kivisild et al. [26] based only on synonymous substitutions, which would generally shrink ages by a factor of ~3/4, as judged from a comparison with both the ages of the Native American haplogroups [14] and those of super-haplogroups L, L3, M, and N [43]. Therefore the main difference between both rates seems to concern only the absolute calibration as manifested in the estimated global coalescence times for super-haplogroup L. It is dubious whether the partial utilization of the coding-region information [14,26] leads to more credible age estimates, taking into account the extremely low amount of synonymous mutation data characterizing younger clades, such as the Amerindian ones, and the extreme discrepancies with ages based on control-region variation of some haplogroups such as H, I, T, and U5 [44]. Moreover, if as suggested [26], the molecular clock did not apply to the entire coding region, but only to the synonymous mutations in the 13 genes coding for protein subunits, it would be rather unlikely that an age overlapping such as that reported for the well represented founder haplogroups (A2, B2, D1, C1b, and C1c) in Table 1 would be observed. In any case, with both clocks, a Beringian stage preceeding the expansion into the Americas – estimated at slightly different starting times and with a different duration depending on the clock employed – most likely took place, thus explaining the differentiation of the pan-American lineages from the Asian sister-clades (Figure 1).

Our snapshot of the phylogenies for haplogroups A2, B2, C1, and D1 is only partially representative of Native American mtDNA variation, since most likely it only marginally includes the variation of Native American populations from Central and South America. However, despite this limitation, it is clear that one has to anticipate a pronounced starlike pattern near the root of each respective founder haplogroup/branch. The starlike pattern enhances the precision of the dating of the human entry into the Americas, but inevitably hinges upon the calibration employed and, perhaps more importantly, on a detailed founder analysis across the double-continent. Therefore it will require major sampling and sequencing efforts in the future for uncovering all of the most basal variation in the Native American mtDNA haplogroups by targeting, if possible, both the general mixed population of national states and autochthonous Native American groups, especially in Central and South America.

A widespread knowledge of the specifics for the Native American haplogroups can also prevent the publishing of effectively mutilated or distorted mtDNA sequences from complete sequencing efforts in clinical studies [45,46], but most importantly, the dissection of pan-American haplogroups into clades of younger age and more limited geographic and ethnic distributions is essential for reliable association studies between mtDNA haplogroups and complex disorders [31].

Materials and Methods

The source of the sequence data (171 complete mtDNA sequences) employed for the phylogeny construction are listed in Table S1 (and Text S5), together with 14 novel Native American

mtDNA sequences (four each belonging to haplogroups A2 and C1; three each belonging to B2 and D1) from the Dominican Republic (N = 4), Canada (N = 3) and United States (N = 7). The latter were completely sequenced as described elsewhere [47]. Additional 47 entire coding-region sequences [20,26] were employed only for time estimation and inference of branching nodes (see also Text S4).

The 101 complete mtDNA sequences [19] represent 13 of the 18 most common HVS-I & II haplotypes among the "Hispanic" component of the SWGDAM database [48]. Anonymous, unrelated samples were identified and obtained from either an internal Armed Forces DNA Identification Laboratory (AFDIL) database, or from 575 regional "Hispanics" living in the southern and northeastern regions of the US. The control region of their mtDNAs was then sequenced in order to determine the common HVS-I & II haplotypes [19].

Electronic database information

Accession numbers and URLs for data presented herein are as follows: GenBank, http://www.ncbi.nlm.nih.gov/Genbank/ (for the 14 novel complete mtDNA sequences [accession numbers EF079873-EF079876; EU431080-EU431089]); (for sequence no. 3 of Figure 1 [accession number EU439939])

Supporting Information

Text S1 Mistakes, phantom mutations and discrepancies in literature and public databases
Found at: doi:10.1371/journal.pone.0001764.s001 (0.06 MB DOC)

Text S2 Further information from mtDNA control-region and RFLP data
Found at: doi:10.1371/journal.pone.0001764.s002 (0.08 MB DOC)

Text S3 Additional information concerning mtDNA disease studies
Found at: doi:10.1371/journal.pone.0001764.s003 (0.04 MB DOC)

Text S4 Additional information for Figures 1–3
Found at: doi:10.1371/journal.pone.0001764.s004 (0.04 MB DOC)

Text S5 Additional references
Found at: doi:10.1371/journal.pone.0001764.s005 (0.04 MB DOC)

Table S1 Source of the complete mtDNA sequences
Found at: doi:10.1371/journal.pone.0001764.s006 (0.39 MB DOC)

Acknowledgments

We would also like to thank all the donors for providing biological specimen and the people involved in their collection.

Author Contributions

Conceived and designed the experiments: AS AT HB AA. Performed the experiments: AA UP. Analyzed the data: AS AT HB CB QK AA MC UP SW. Contributed reagents/materials/analysis tools: AT. Wrote the paper: AS AT HB CB QK AA MC UP SW.

References

1. Schurr TG, Sherry ST (2004) Mitochondrial DNA and Y chromosome diversity and the peopling of the Americas: evolutionary and demographic evidence. Am J Hum Biol 16: 420–439.
2. Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, et al. (2007) Genetic variation and population structure in Native Americans. PLoS Genet 3: e185.
3. Schurr TG, Ballinger SW, Gan YY, Hodge JA, Merriwether DA, et al. (1990) Amerindian mitochondrial DNAs have rare Asian mutations at high frequencies, suggesting they derived from four primary maternal lineages. Am J Hum Genet 46: 613–623.
4. Torroni A, Schurr TG, Yang CC, Szathmary EJE, Williams RC, et al. (1992) Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. Genetics 130: 153–162.
5. Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, et al. (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53: 563–590.
6. Torroni A, Sukernik RI, Schurr TG, Starikorskaya YB, Cabell MF, et al. (1993) mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans. Am J Hum Genet 53: 591–608.
7. Schurr TG (2004) The peopling of the New World: perspectives from molecular anthropology. Annu Rev Anthropol 33: 551–583.
8. Bandelt H-J, Herrnstadt C, Yao Y-G, Kong Q-P, Kivisild T, et al. (2003) Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. Ann Hum Genet 67: 512–524.
9. Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59: 935–945.
10. Scozzari R, Cruciani F, Santolamazza P, Sellitto D, Cole DE, et al. (1997) mtDNA and Y chromosome-specific polymorphisms in modern Ojibwa: implications about the origin of their gene pool. Am J Hum Genet 60: 241–244.
11. Brown MD, Hosseini SH, Torroni A, Bandelt H-J, Allen JC, et al. (1998) mtDNA haplogroup X: An ancient link between Europe/Western Asia and North America? Am J Hum Genet 63: 1852–1861.
12. Reidla M, Kivisild T, Metspalu E, Kaldma K, Tambets K, et al. (2003) Origin and diffusion of mtDNA haplogroup X. Am J Hum Genet 73: 1178–1190.
13. Derbeneva OA, Sukernik RI, Volodko NV, Hosseini SH, Lott MT, et al. (2002) Analysis of mitochondrial DNA diversity in the Aleuts of the Commander islands and its implications for the genetic history of Beringia. Am J Hum Genet 71: 415–421.
14. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, et al. (2007) Beringian standstill and spread of Native American founders. PLoS ONE 2: e829.
15. Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67: 718–726.
16. Helgason A, Pálsson G, Pedersen HS, Angulalik E, Gunnarsdóttir ED, et al. (2006) mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. Am J Phys Anthropol 130: 123–134.
17. Kemp BM, Malhi RS, McDonough J, Bolnick DA, Eshleman JA, et al. (2007) Genetic analysis of early Holocene skeletal remains from Alaska and its implications for the settlement of the Americas. Am J Phys Anthropol 132: 605–621.
18. Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, et al. (2007) Phylogeographic analysis of mitochondrial DNA in northern Asian populations. Am J Hum Genet 81: 1025–1041.
19. Parsons TJ (unpublished) http://www.ncjrs.gov/pdffiles1/nij/grants/213502.pdf.
20. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, et al. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152–1171.
21. Accetturo M, Santamaria M, Lascaro D, Rubino F, Achilli A, et al. (2006) Human mtDNA site-specific variability values can act as haplogroup markers. Hum Mutat 27: 965–974.
22. Herrnstadt C, Preston G, Howell N (2003) Errors, phantoms and otherwise, in human mtDNA sequences. Am J Hum Genet 72: 1585–1586.
23. Torroni A, Chen YS, Semino O, Santachiara-Benecceretti AS, Scott CR, et al. (1994) mtDNA and Y-chromosome polymorphisms in four Native American populations from southern Mexico. Am J Hum Genet 54: 303–318.
24. Torroni A, Neel JV, Barrantes R, Schurr TG, Wallace DC (1994) Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. Proc Natl Acad Sci U S A 91: 1158–1162.
25. Álvarez-Iglesias V, Jaime JC, Carracedo Á, Salas A (2007) Coding region mitochondrial DNA SNPs: Targeting East Asian and Native American haplogroups. Forensic Sci Int Genet 1: 44–55.
26. Kivisild T, Shen P, Wall DP, Do B, Sung R, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373–387.
27. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, et al. (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100: 171–176.
28. Kong Q-P, Bandelt H-J, Sun C, Yao Y-G, Salas A, et al. (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15: 2076–2086.

29. Li Z, Li R, Chen J, Liao Z, Zhu Y, et al. (2005) Mutational analysis of the mitochondrial 12S rRNA gene in Chinese pediatric subjects with aminoglycoside-induced and non-syndromic hearing loss. Hum Genet 117: 9–15.
30. Yao Y-G, Salas A, Bravi CM, Bandelt H-J (2006) A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. Hum Genet 119: 505–515.
31. Carelli V, Achilli A, Valentino ML, Rengo C, Semino O, et al. (2006) Haplogroup effects and recombination of mitochondrial DNA: novel clues from the analysis of Leber hereditary optic neuropathy pedigrees. Am J Hum Genet 78: 564–574.
32. Simon DK, Tarnopolsky MA, Greenamyre JT, Johns DR (2001) A frameshift mitochondrial complex I gene mutation in a patient with dystonia and cataracts: is the mutation pathogenic? J Med Genet 38: 58–61.
33. Rocha H, Flores C, Campos Y, Arenas J, Vilarinho L, et al. (1999) About the "pathological" role of the mtDNA T3308C mutation… Am J Hum Genet 65: 1457–1459.
34. MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org.
35. Crispim D, Canani LH, Gross JL, Carlessi RM, Tschiedel B, et al. (2005) The G1888A variant in the mitochondrial 16S rRNA gene may be associated with Type 2 diabetes in Caucasian-Brazilian patients from southern Brazil. Diabet Med 22: 1683–1689.
36. Palanichamy Mg, Sun C, Agrawal S, Bandelt H-J, Kong Q-P, et al. (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75: 966–978.
37. Sun C, Kong Q-P, Palanichamy Mg, Agrawal S, Bandelt H-J, et al. (2006) The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. Mol Biol Evol 23: 683–690.
38. Crispim D, Canani LH, Gross JL, Tschiedel B, Souto KE, et al. (2006) The European-specific mitochondrial cluster J/T could confer an increased risk of insulin-resistance and type 2 diabetes: an analysis of the m.4216T > C and m.4917A > G variants. Ann Hum Genet 70: 488–495.
39. Alves-Silva J, da Silva Santos M, Guimarães PE, Ferreira AC, Bandelt H-J, et al. (2000) The ancestry of Brazilian mtDNA lineages. Am J Hum Genet 67: 444–461.
40. Mosquera-Miguel A, Álvarez-Iglesias V, Vega A, Milne R, Cabrera de León A, et al. (2008) Is mitochondrial DNA variation associated with sporadic breast cancer risk? Cancer Res, in press.
41. Mix AC, Bard E, Schneider R (2001) Environmental processes of the ice age: land, oceans, glaciers (EPILOG). Quaternary Science Reviews 20: 627–657.
42. Waters MR, Stafford TW Jr. (2007) Redefining the age of Clovis: implications for the peopling of the Americas. Science 315: 1122–1126.
43. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, et al. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308: 1034–1036.
44. Richards M, Macaulay V, Hickey E, Vega E, Sykes B, et al. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67: 1251–1276.
45. Bandelt H-J, Achilli A, Kong Q-P, Salas A, Lutz-Bonengel S, et al. (2005) Low "penetrance" of phylogenetic knowledge in mitochondrial disease studies. Biochem Biophys Res Commun 333: 122–130.
46. Bandelt H-J, Yao Y-G, Salas A, Kivisild T, Bravi CM (2007) High penetrance of sequencing errors and interpretative shortcomings in mtDNA sequence analysis of LHON patients. Biochem Biophys Res Commun 352: 283–291.
47. Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, et al. (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75: 910–918.
48. Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B (2002) The mtDNA Population Database: an integrated software and database resource for forensic comparison. Forensic Sci Commun 4: 2.
49. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, et al. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23: 147.
50. Bandelt H-J, Parson W (2008) Consistent treatment of length variants in the human mtDNA control region: a reappraisal. Int J Legal Med 122: 11–21.

Text S1. Mistakes, phantom mutations and discrepancies in literature and public databases

Phylogenetic studies.
The geneticist or anthropologist who wishes to study the mtDNA variation in individuals of matrilineal Native American ancestries would have at the moment a very difficult job. He/she might come across the paper [1] that provides a basal classification scheme and supplementary information on previously published coding-region sequences [2,3], but the corresponding complete mtDNA sequences are difficult to retrieve (see notes to Table S1). The most recent submission (July 14, 2007) of the data from [2] to GenBank (Acc. No. EF657231 to EF657790) may additionally confuse the user as all the phantom mutations that were announced in [3] to be removed are still in place, e.g. the transversions to G at positions 7927, 7985, 11227, 14227, 14385, 14460, and 15040 in the Native American coding-region sequences no. 171, 174, 200, 329, and 377 [2]. In the lack of any properly published revision, it is not even clear whether the transversions at 14463 in no. 171 (haplogroup A2; see #34 in our Table 1 below) and at 14974 in no. 197 (haplogroup B2) present in the "revised" data [3] that we employed would be real or not. Moreover, in the study [1] and in the mtDB database [4] the continental assignments of the mtDNAs are not always correct. In particular, the distinction between Asian and Native American mtDNAs was blurred. As for the most recent data of Tamm et al. [5], there are some minor discrepancies between the sequences read off from their Fig. S1 and from the corresponding sequences stored in GenBank (which we have consistently used).

To exacerbate the situation, a genuine Native American branch (subhaplogroup) of haplogroup C1 slipped into the Asian mtDNA phylogeny under the name "C6" in the recent update of the East Asian mtDNA phylogeny [6] due to the lack of control-region information from the sequence data employed. Finally, although prior to November 2006 no additional complete sequences were published, the naming of new branches has been going on – but based only on the poor information that hypervariable segment I (HVS-I) sequences can provide and, unfortunately, by baptizing nested branches of A2 as A3, A6, A7, and A8 [7] instead of following the standard nomenclature rules for naming nested branches, starting with A2a, etc. [8]. It still remains very unlikely that haplogroup B4b1a2, which was called B1 by [9] at the time, could be a Native American branch. The T16136C transition on a haplogroup B4 is, for instance, rather common in the Japanese, Korean, Chinese, and Guam fractions of the SWGDAM mtDNA database [10], but is completely absent in the US American fractions ("Caucasian", "African-American", and "Hispanic"). Furthermore, at the level of the entire American double-continent, only a single occurrence (sample #AND19, a Peruvian Quechua from Arequipa) is recorded in the literature bearing the transition T16136C diagnostic for the B4b1 clade [11].

Disease studies.
The study of [12] provided two haplogroup A2 sequences (control #7 corresponding to sequence #50 in our Figure 2, and patient #12) without listing the nucleotide variants that exist between the rCRS and the root of super-haplogroup N; moreover, the C8794T change is missing in one case (control #7). Similarly, the study [13] determined only a 2/3 subset of the mutation motif separating rCRS from the root of haplogroup A (but reported well all additional mutations for A2 status).

Text S2. Further information from mtDNA control-region and RFLP data

Haplogroup A2.
The characteristic A2b mutation T11365C is recognizable by a loss of the 11362 *Alu*I site [14]. Despite the fact that it was first described as distinctive of the subhaplogroup named "A2a" [15], we maintained the now commonly used nomenclature for A2a and A2b [16]. The G15172A transition (see #75 in Figure 2) causes the *Hae*III site loss at 15172, a site which was indeed observed in RFLP types AM11 and AM106 from Mexico. Furthermore, the A11548G transition in A2j leads to –11546 *Rsa*I, which was recorded in types AM100, AM103, and AM104 (also sampled in Mexico). One further match between the complete A2 sequences and the high-resolution RFLP haplotypes is found for sequence #27 (A2f) in Figure 2 and haplotype AM62 described in ~36% of the Ojibwa from North America [17]; in fact, transitions at positions 2638, 3316 and 7897 correspond to *Hae*III changes at 2636 (site gain) and 3315 (site loss), and a *Rsa*I site loss at 7897 [17], respectively. Moreover, the two mutations defining sub-branch A2d1 are supported by the Mexican type AM111: the corresponding *Alu*I site gain at 11313, however, was incorrectly assigned 8 bp apart.

Sequence #50 in Figure 2 of [12] is a member of a conspicuous set of lineages present in populations from Lower Central America whose ancestral control region motif can be defined by the presence of C16360T and a 6-bp deletion involving positions 106-111 on top of the A2 root. This deletion abolishes a recognition site for *Msp*I at position 104, which has been reported in the Boruca, Kuna, Guaymi, Bribri-Cabecar [17], and the Teribe [18]. Although moderately recurrent, it has always been seen linked with a C16360T transition in Chibchan-speaking populations from Central America [17,19-21]. Sequence #50 carries a derived motif in its HVS-II segment, namely the co-occurrence of transitions T89C and C198T on top of the A2 plus C16360T motif, a combination that hitherto has only been observed in the Huetars and Ngöbe from Central America [21,22]. However, sequence #50 lacks the otherwise expected 6-bp deletion and offers instead the 106C transversion. Since the reversal of a 6-bp deletion is highly unlikely, an incorrect interpretation might have occurred.

Haplogroup C1.
The transition at 7013 (sub-branch C1b2) was found as 7013 *Rsa*I site loss in several Amazonians [17], while the sub-branch C1b3 has the characteristic 12630 transition, which is also recognized by a restriction site, namely by –12629 *Ava*II, as seen in AM120 from Mexico [23]. Also the variant G11440A of C1c2 has been previously seen in Native Americans from Mexico as +11439 *Mbo*I site [23].

The only C lineage from [2], other than the haplogroup C5 members [1], for which C1 status cannot be inferred by comparison with shared mutations with complete C1 sequences is sample no. 174, having the private transition motif A5894G-G6261A-A10397G-G13813A-T14215C. However, the RFLP data [17,23] provide a hint. The A10397G transition (which appears to be a rather infrequent mutation) would erase both the 10394 *Dde*I and 10397 *Alu*I restriction sites. In fact, three RFLP types found in Mexico (AM30, AM31, and AM122) testify to this reverted RFLP pattern. This justifies our assumption that one can also consider sample no. 174 as of Native American ancestry, which likely is a member of C1b since this branch is only recognized by a control-region mutation (A493G) in contrast to the other two Native American branches of haplogroup C1.

Since haplogroup C1 as a whole and its subhaplogroup C1b are only characterized by mutations in the control region, trees that are exclusively based on the coding region [2,24] are not helpful for uncovering the phylogenetic relationships within haplogroup C1. This also concerns the huge MITOMAP mtDNA tree [25], where two apparently polyphyletic clusters emerged within the C branch due to homoplasy at the highly variable positions 1719 and 15930 (Table 6 in [24]).

Text S3. Additional information concerning mtDNA disease studies

In the context of the common-disease common-polymorphism hypothesis, any variant could be considered as a candidate contributing to some common disease, independent of its age and frequency in the population. However, indiscriminate case-control association studies of mtDNA variation can easily lead to premature claims of pathogenicity if just 'random' control samples are taken that were not controlled for population substructure. This is especially true in the context of mtDNA studies where one must consider the fact that mtDNA variation is strongly structured in populations. Thus, it is puzzling to learn that several basal mutations in A2 and C1 were assumed to have some detrimental effect as inferred from studies with improper designs. In the extreme case, an entire Native American haplogroup, such as haplogroup B2, could easily become the target of disease association. For example, the Ile to Val amino acid change in *ND1* due to A3547G – a marker of all haplogroup B2 lineages – was considered in a study that aimed at identifying mutations predisposing to Parkinson's disease (PD), but finally "*no difference was noted in the frequencies of the 3547 mutation in PD and control subjects*" [26]. Even more extreme is the claim that the mutation A827G, characteristic of haplogroup B4bd in which B2 is nested (Figure 1), is pathogenic [27]; see a reassessment by [28].

Most of the mutations mentioned here that were suspected of disease association (A827G, T1005C, G1888A, T3308A, 3308+C, G3316A, A3547G, T12338C, and G13708A) had their genesis in single case studies because the published record was not consulted and/or MITOMAP [25] had not listed those mutations as polymorphisms at the time. Now (as of December 17, 2007) all those mutations except for T3308A and 3308+C are listed as "MtDNA Coding Region Sequence Polymorphisms" in MITOMAP. In addition, A827G and T1005C also appear there under the category "Reported Mitochondrial DNA Base Substitution Diseases: rRNA/tRNA Mutations" with status "Under Review", G3316A there as "Unclear", and G13708A with the comment "that some published reports have determined the mutation to be a non-pathogenic polymorphism", thus, no green light from MITOMAP either for disease status.

In fact, several Native American haplogroups bear mutations that could increase the suspicion of disease status while also paralleling that of sub-Saharan African haplogroups, such as L1b as a whole in regard to MELAS [29] or a specific (unnamed) branch of L2a1 with elevated 'risk' of prostate cancer [30]. Without a detailed knowledge of all haplogroups – major or minor – thriving in the geographic region or ethnic population of interest, one cannot evaluate whether a mutation defining a particular minor subhaplogroup is responsible for the observed statistical association. Simply screening single coding-region sites believed to participate in the etiology of a disease without properly monitoring confounding variables such as population stratification could easily become a "lottery" since the frequency of a certain haplogroup in the area could actually constitute the major variable determining a strong correlation with the disease in a case-control association study.

Text S4. Additional information for Figures 1-3

In few instances, the classification trees are expanded by incorporating branching nodes inferred from the additional coding-region sequences of [2,3,22,24], for which in most cases control-region information is unfortunately unavailable. In naming the sub-branches, we followed the scheme of [8] and use the principle that a valid code introduced in a publication first is preferred over a second name given later (in ignorance of the former). An exception would be made only in the case that a name given later has become most widely applied in the meantime, so that a shift back to the first name would rather create more confusion (as is e.g. the case with the generally employed name H2a of the subhaplogroup of H2 harboring the rCRS). For example, we attach the name C1a to the single known Asian branch of haplogroup C1 (Figure 1) following [22]. Many other haplogroup names for Asian and Native American haplogroups given in [22], however, violated previously published nomenclature without warning the reader of the change. For example, haplogroups B4, B4d, C4, D1, D4, D5, M8, and M8a (see [6] for earlier references) have been *de novo* baptized in [22], but the new codes have apparently not been followed elsewhere.

Text S5. Additional references

1. Bandelt H-J, Herrnstadt C, Yao Y-G, Kong Q-P, Kivisild T, et al. (2003) Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. Ann Hum Genet 67: 512-524.
2. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, et al. (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am J Hum Genet 70: 1152-1171.
3. Herrnstadt C, Preston G, Howell N (2003) Errors, phantoms and otherwise, in human mtDNA sequences. Am J Hum Genet 72: 1585-1586.
4. Ingman M, Gyllensten U (2006) mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34: D749-D751.
5. Tamm E, Kivisild T, Reidla M, Metspalu M, Smith DG, et al. (2007) Beringian standstill and spread of Native American founders. PLoS ONE 2: e829.
6. Kong Q-P, Bandelt H-J, Sun C, Yao Y-G, Salas A, et al. (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. Hum Mol Genet 15: 2076-2086.
7. Zlojutro M, Rubicz R, Devor EJ, Spitsyn VA, Makarov SV, et al. (2006) Genetic structure of the Aleuts and Circumpolar populations based on mitochondrial DNA sequences: a synthesis. Am J Phys Anthropol 129: 446-464.
8. Richards MB, Macaulay VA, Bandelt H-J, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62: 241-260.
9. Herrnstadt C, Howell N (2004) An evolutionary perspective on pathogenic mtDNA mutations: haplogroup associations of clinical disorders. Mitochondrion 4: 791-798.
10. Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B (2002) The mtDNA Population Database: an integrated software and database resource for forensic comparison. Forensic Sci Commun 4: 2.
11. Fuselli S, Tarazona-Santos E, Dupanloup I, Soto A, Luiselli D, et al. (2003) Mitochondrial DNA diversity in South America and the genetic history of Andean highlanders. Mol Biol Evol 20: 1682-1691.
12. Shin MG, Kajigaya S, Levin BC, Young NS (2003) Mitochondrial DNA mutations in patients with myelodysplastic syndromes. Blood 101: 3118-3125.
13. Glanowski S (2003) Computational Analysis of Mitochondrial Sequence Diversity in Shotgun Sequence Data. *PhD thesis*, George Mason University.
14. Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67: 718-726.
15. Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, et al. (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 14: 1832-1850.
16. Helgason A, Pálsson G, Pedersen HS, Angulalik E, Gunnarsdóttir ED, et al. (2006) mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. Am J Phys Anthropol 130: 123-134.
17. Torroni A, Schurr TG, Cabell MF, Brown MD, Neel JV, et al. (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. Am J Hum Genet 53: 563-590.
18. Torroni A, Neel JV, Barrantes R, Schurr TG, Wallace DC (1994) Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. Proc Natl Acad Sci U S A 91: 1158-1162.
19. Batista O, Kolman CJ, Bermingham E (1995) Mitochondrial DNA diversity in the Kuna Amerinds of Panama. Hum Mol Genet 4: 921-929.
20. Kolman CJ, Bermingham E, Cooke R, Ward RH, Arias TD, et al. (1995) Reduced mtDNA diversity in the Ngobe Amerinds of Panama. Genetics 140: 275-283.

21. Santos M, Ward RH, Barrantes R (1994) mtDNA variation in the Chibcha Amerindian Huetar from Costa Rica. Hum Biol 66: 963-977.
22. Starikovskaya EB, Sukernik RI, Derbeneva OA, Volodko NV, Ruiz-Pesini E, et al. (2005) Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups. Ann Hum Genet 69: 67-89.
23. Torroni A, Chen YS, Semino O, Santachiara-Benecerretti AS, Scott CR, et al. (1994) mtDNA and Y-chromosome polymorphisms in four Native American populations from southern Mexico. Am J Hum Genet 54: 303-318.
24. Kivisild T, Shen P, Wall DP, Do B, Sung R, et al. (2006) The role of selection in the evolution of human mitochondrial genomes. Genetics 172: 373-387.
25. MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org.
26. Simon DK, Mayeux R, Marder K, Kowall NW, Beal MF, et al. (2000) Mitochondrial DNA mutations in complex I and tRNA genes in Parkinson's disease. Neurology 54: 703-709.
27. Li Z, Li R, Chen J, Liao Z, Zhu Y, et al. (2005) Mutational analysis of the mitochondrial 12S rRNA gene in Chinese pediatric subjects with aminoglycoside-induced and non-syndromic hearing loss. Hum Genet 117: 9-15.
28. Yao Y-G, Salas A, Bravi CM, Bandelt H-J (2006) A reappraisal of complete mtDNA variation in East Asian families with hearing impairment. Hum Genet 119: 505-515.
29. Rocha H, Flores C, Campos Y, Arenas J, Vilarinho L, et al. (1999) About the "pathological" role of the mtDNA T3308C mutation... Am J Hum Genet 65: 1457-1459.
30. Bandelt H-J, Salas A, Bravi CM (2006) What is a 'novel' mtDNA mutation--and does 'novelty' really matter? J Hum Genet 51: 1073-1082.
31. Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, et al. (2003) Natural selection shaped regional mtDNA variation in humans. Proc Natl Acad Sci U S A 100: 171-176.
32. Derbeneva OA, Sukernik RI, Volodko NV, Hosseini SH, Lott MT, et al. (2002) Analysis of mitochondrial DNA diversity in the Aleuts of the Commander islands and its implications for the genetic history of Beringia. Am J Hum Genet 71: 415-421.
33. Parsons TJ (unpublished) http://www.ncjrs.gov/pdffiles1/nij/grants/213502.pdf.
34. Simon DK, Tarnopolsky MA, Greenamyre JT, Johns DR (2001) A frameshift mitochondrial complex I gene mutation in a patient with dystonia and cataracts: is the mutation pathogenic? J Med Genet 38: 58-61.
35. Maca-Meyer N, González AM, Larruga JM, Flores C, Cabrera VM (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genet 2: 13.
36. Ingman M, Kaessmann H, Pääbo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408: 708-713.
37. Delgado-Sánchez R, Zárate-Moysen A, Monsalvo-Reyes A, Herrero MD, Ruiz-Pesini E, et al. (2007) [Mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS) with the A3243G mutation of the tRNALeu(UUR) gene of mtDNA in Native American haplogroup B2]. Rev Neurol 44: 18-22.

Table S1. Source of the complete mtDNA sequences

Sample ID[a]	Haplogroup	Original ID	GenBank ID	Reference
1	A2a	Apache514	EU095526	[5]
2	X2a	Na3x	AY195787	[31]
3	X2a	Oj2	EU439939	[1]; present study
4	C4c	Ijka72	EU095543	[5]
5	D2a	Aleut (#IX)	not available	[32]
6	D2a	Aleut (#I)	not available	[32]
6	D2a	Aleut (#II)	not available	[32]
6	D2a	Aleut (#III)	not available	[32]
6	D2a	Aleut (#IV)	not available	[32]
6	D2a	Aleut (#V)	not available	[32]
6	D2a	Aleut (#VI)	not available	[32]
6	D2a	Aleut (#VII)	not available	[32]
7	D4h3	Cayapa600	EU095531	[5]
8	A2c	332	not available	[1]
9	A2c	200	not available	[1]
9	A2c	241	not available	[1]
10	A2c	411	not available	[1]
11	A2d1	418	not available	[1]
12	A2d1	A2-6-05	DQ282433	[33]
13	A2d1	A2-6-03	DQ282431	[33]
14	A2d1	A2-6-01	DQ282429	[33]
15	A2d1	A2-6-02	DQ282430	[33]
16	A2d1	A2-6-04	DQ282432	[33]
17	A2d2	A2-1-02	DQ282388	[33]
18	A2d2	244	not available	[1]
18	A2d2	A2-3-14	DQ282422	[33]
19	A2d2	A2-3-13	DQ282421	[33]
20	A2d	A2-3-10	DQ282418	[33]
21	A2d	331	not available	[1]
22	A2d	376	not available	[1]
23	A2	Tor22	EF079873	Dominican Rep.; present study
24	A2e	139	not available	[1]
25	A2e	400	not available	[1]
26	A2e	374	not available	[1]
27	A2f	12	not available	[1]
28	A2f	532	not available	[1]
29	A2g	A2-1-06	DQ282392	[33]
30	A2g	112	not available	[1]
30	A2g	A2-1-05	DQ282391	[33]
31	A2g	329	not available	[1]
32	A2g	A2-2-05	DQ282406	[33]
33	A2h	247	not available	[1]
34	A2h	171	not available	[1]
35	A2i	170	not available	[1]
36	A2i	IAB_D6	EU431080	USA; present study
37	A2i	Patient	not available	[34]
38	A2j1	A2-1-12	DQ282398	[33]
39	A2j1	A2-1-07	DQ282393	[33]
40	A2j1	459	not available	[1]
41	A2j1	A2-2-03	DQ282404	[33]
42	A2j	A2-1-01	DQ282387	[33]
43	A2k1	A2-1-14	DQ282400	[33]
44	A2k1	A2-1-04	DQ282390	[33]
45	A2k1	Wayuu24	EU095552	[5]
46	A2k	A2-3-09	DQ282417	[33]
47	A2	439	not available	[1]
48	A2	A2-1-24	DQ282401	[33]

49	A2	A2-1-10	DQ282396	[33]
49	A2	A2-1-13	DQ282399	[33]
49	A2	A2-1-09	DQ282395	[33]
49	A2	A2-1-11	DQ282397	[33]
50	A2	#7 (Control)	not available	[12]
51	A2	A2-5-01	DQ282428	[33]
52	A2	A2-4-01	DQ282424	[33]
53	A2	A2-4-03	DQ282426	[33]
53	A2	A2-4-02	DQ282425	[33]
53	A2	A2-4-04	DQ282427	[33]
54	A2	IA_C3	EU431081	USA; present study
55	A2	CanarAF_10	AF382010	[35]
56	A2	415	not available	[1]
57	A2	A2-3-03	DQ282411	[33]
58	A2	A2-3-05	DQ282413	[33]
59	A2	A2-3-04	DQ282412	[33]
59	A2	A2-3-06	DQ282414	[33]
59	A2	A2-3-08	DQ282416	[33]
59	A2	A2-3-12	DQ282420	[33]
59	A2	A2-3-01	DQ282409	[33]
59	A2	A2-3-02	DQ282410	[33]
59	A2	A2-3-07	DQ282415	[33]
60	A2	A2-3-15	DQ282423	[33]
61	A2	A2-3-11	DQ282419	[33]
62	A2	IA_H4	EU431082	Canada; present study
63	A2	Cayapa522	EU095530	[5]
64	A2	330	not available	[1]
65	A2	120	not available	[1]
66	A2	176	not available	[1]
67	A2	184	not available	[1]
68	A2	A2-2-01	DQ282402	[33]
69	A2	A2-2-02	DQ282403	[33]
69	A2	A2-2-04	DQ282405	[33]
69	A2	A2-2-06	DQ282407	[33]
70	A2	A2-2-07	DQ282408	[33]
71	A2	Arsario20	EU095528	[5]
72	A2	A2-1-03	DQ282389	[33]
73	A2	Kogui39	EU095545	[5]
74	A2	Cayapa511	EU095529	[5]
75	A2	Na5A	AY195786	[31]
76	A2	A2-1-08	DQ282394	[33]
77	A2	Dogrib39	EU095538	[5]
78	B2a1a	B2-2-04	DQ282444	[33]
79	B2a1a	B2-2-06	DQ282446	[33]
80	B2a1	B2-2-05	DQ282445	[33]
81	B2a1	B2-2-03	DQ282443	[33]
82	B2a1	B2-2-02	DQ282442	[33]
83	B2a1	IA_E2	EU431083	USA; present study
84	B2a	B2-2-01	DQ282441	[33]
85	B2a	(4) Pi_26_27	AF347001	[36]
86	B2b	419	not available	[2,5]
87	B2b	Cayapa602	EU095532	[5]
88	B2c1	B2-1-05	DQ282438	[33]
89	B2c1	B2-1-06	DQ282439	[33]
90	B2c1	B2-1-01	DQ282434	[33]
91	B2c1	B2-1-07	DQ282440	[33]
92	B2c1	B2-1-04	DQ282437	[33]
93	B2c	B2-1-03	DQ282436	[33]
94	B2c	Patient	not available	[37]
95	B2d	Na1B	AY195749	[31]

96	B2d	Ngoebe14	EU095546	[5]
97	B2d	Wayuu7	EU095550	[5]
98	B2d	Wayuu17	EU095551	[5]
99	B2	Coreguaje1-30	EU095535	[5]
100	B2	IA_G1	EU431084	USA; present study
101	B2	Waunana2-8	EU095548	[5]
102	B2	Tor23	EF079874	Dominican Rep.; present study
103	B2	B2-1-02	DQ282435	[33]
104	B2	Native - Sinixt	EF648602	Direct Submission
105	C1b1	Na4C	AY195759	[31]
106	C1b2a	C1-1-03	DQ282449	[33]
107	C1b2a	C1-1-02	DQ282448	[33]
107	C1b2a	C1-1-04	DQ282450	[33]
107	C1b2a	C1-1-05	DQ282451	[33]
107	C1b2a	C1-1-06	DQ282452	[33]
107	C1b2a	C1-1-08	DQ282454	[33]
107	C1b2a	C1-1-09	DQ282455	[33]
107	C1b2a	C1-1-10	DQ282456	[33]
107	C1b2a	C1-1-11	DQ282457	[33]
107	C1b2a	C1-1-12	DQ282458b	[33]
107	C1b2a	C1-1-01	DQ282447	[33]
108	C1b2a	C1-1-07	DQ282453	[33]
109	C1b2	CanarAF_09	AF382009	[35]
110	C1b3	C1-2-06	DQ282464	[33]
111	C1b4	C1-4-02	DQ282475	[33]
112	C1b4	C1-4-03	DQ282476	[33]
113	C1b4	IA_F1	EU431085	USA; present study
114	C1b5	C1-2-03	DQ282461	[33]
115	C1b5	C1-2-12	DQ282469	[33]
116	C1b	Wayuu4	EU095549	[5]
117	C1c2	C1-2-11	DQ282468	[33]
117	C1c2	C1-2-14	DQ282471	[33]
117	C1c2	C1-2-08	DQ282466	[33]
117	C1c2	C1-2-10	DQ282467	[33]
118	C1c2	C1-2-04	DQ282462	[33]
119	C1c2	C1-2-13	DQ282470	[33]
120	C1c	Arsario5	EU095527	[5]
121	C1c	Kogui12	EU095544	[5]
122	C1c	C1-2-01	DQ282459	[33]
123	C1c	C1-2-02	DQ282460	[33]
124	C1c	Tor24	EF079875	Dominican Rep.; present study
125	C1c	IA_A3	EU431086	Canada; present study
126	C1c	IA_A7	EU431087	USA; present study
127	C1c	C1-2-05	DQ282463	[33]
128	C1c	C1-2-07	DQ282465	[33]
129	C1d	C1-3-02	DQ282473	[33]
129	C1d	C1-3-03	DQ282474	[33]
130	C1d	C1-3-01	DQ282472	[33]
131	C1d	Coreguaje1-54	EU095537	[5]
132	C1d	(27) Wa_RML	AF347013	[36]
133	C1d	(28) Wa_SPACH	AF347012	[36]
134	D1a	(30) G_GRC150	AF346984	[36]
135	D1	IA_G4	EU431088	Canada; present study
136	D1b	Tor25	EF079876	Dominican Rep.; present study
136	D1b	D1-1-03	DQ282479	[33]
137	D1b	D1-1-08	DQ282484	[33]
138	D1c	D1-1-02	DQ282478	[33]
139	D1c	D1-1-01	DQ282477	[33]
139	D1c	D1-1-05	DQ282481	[33]
140	D1d	D1-1-04	DQ282480	[33]

141	D1d	D1-1-09	DQ282485	[33]
142	D1	Coreguaje1-31	EU095536	[5]
143	D1	IA_F2	EU431089	USA; present study
144	D1	D1-1-10	DQ282486	[33]
145	D1	D1-1-06	DQ282482	[33]
146	D1	D1-1-07	DQ282483	[33]
147	D1	Na2D	AY195748	[31]
148	D1	D1-1-11	DQ282487	[33]

[a] These ID numbers correspond to those in Figures 1-3.
[b] Coding-region information from [2,3]. Note that the indicated website (at the former company MitoKor), for downloading the complete mtDNA sequences, does no longer exist—and the user does not get redirected to the site where the data are now retrievable. The relocated source is http://mito546.securesites.net/science/30asianmtdnas.php. After downloading these sequences one is confronted with another obstacle: all 30 sequences have been misedited in the C stretches around position 310 in that one or two nucleotides C have been misplaced. Variant 317+C should be turned into 315+C instead, variant 315+CC into 309+C plus 315+C, and 315+CCC should become 309+CC plus 315+C. It is unclear whether these data represent the corrected coding-region sequences [3] since the general file at http://mito546.securesites.net/science/560mtdnasrevision.php still presents the flawed data [2], which thus also went into the mtDB database [4].