Software

Open Access

mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences Hwan Young Lee¹, Injee Song², Eunho Ha³, Sung-Bae Cho², Woo Ick Yang¹ and Kyoung-Jin Shin^{*1}

Address: ¹Department of Forensic Medicine and Brain Korea 21 Project for Medical Science, Yonsei University College of Medicine, 250 Seongsanno, Seodaemun-gu, Seoul 120-752, Korea, ²Department of Computer Science, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul 120-749, Korea and ³Department of Information and Statistics, Yonsei University, 234 Maeji-ri, Heungup-myun, Wounju-si, Gangwon-do 220-710, Korea

Email: Hwan Young Lee - hylee192@yuhs.ac; Injee Song - schunya@gmail.com; Eunho Ha - statha@yonsei.ac.kr; Sung-Bae Cho - sbcho@cs.yonsei.ac.kr; Woo Ick Yang - wiyang9660@yuhs.ac; Kyoung-Jin Shin* - kjshin@yuhs.ac * Corresponding author

Published: 17 November 2008

BMC Bioinformatics 2008, 9:483 doi:10.1186/1471-2105-9-483

This article is available from: http://www.biomedcentral.com/1471-2105/9/483

© 2008 Lee et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<u>http://creativecommons.org/licenses/by/2.0</u>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: 18 September 2008 Accepted: 17 November 2008

Abstract

Background: For the past few years, scientific controversy has surrounded the large number of errors in forensic and literature mitochondrial DNA (mtDNA) data. However, recent research has shown that using mtDNA phylogeny and referring to known mtDNA haplotypes can be useful for checking the quality of sequence data.

Results: We developed a Web-based bioinformatics resource "mtDNAmanager" that offers a convenient interface supporting the management and quality analysis of mtDNA sequence data. The mtDNAmanager performs computations on mtDNA control-region sequences to estimate the most-probable mtDNA haplogroups and retrieves similar sequences from a selected database. By the phased designation of the most-probable haplogroups (both expected and estimated haplogroups), mtDNAmanager enables users to systematically detect errors whilst allowing for confirmation of the presence of clear key diagnostic mutations and accompanying mutations. The query tools of mtDNAmanager also facilitate database screening with two options of "match" and "include the queried nucleotide polymorphism". In addition, mtDNAmanager provides Web interfaces for users to manage and analyse their own data in batch mode.

Conclusion: The mtDNAmanager will provide systematic routines for mtDNA sequence data management and analysis via easily accessible Web interfaces, and thus should be very useful for population, medical and forensic studies that employ mtDNA analysis. mtDNAmanager can be accessed at <u>http://mtmanager.yonsei.ac.kr</u>.

Background

The outstanding features of human mitochondrial DNA (mtDNA) – such as its high mutation rate, absence of recombination, stability and the large number of genome copies per cell – have led to its wide utilization in various

disciplines, including population, medical and forensic genetics. For the past few years, scientific controversy has surrounded the large numbers of errors detected in much of the previously published mtDNA data [1,2]. In extreme cases erroneous data can alter the main conclusion of a study [3], requiring confirmation of the absence of errors before proceeding to further analysis or drawing meaningful conclusions. Since phylogenetic investigations and database screening could have detected prevalent errors in published data sets, methodologies based on mtDNA haplogroup determination and comparisons with existing mtDNA haplotypes were proposed for preventing mtDNA errors [4,5]. In particular, the phylogenetic approach which is the key tool used to understand the structure of the mtDNA data under study - was shown to be very useful for systematic reanalysis of an mtDNA data set. According to data and part of the phylogeny, it was reported to detect approximately 50% of all sequence errors [3] and hence has formed a starting point to localizing a sequence to a part of the phylogeny, at least to the level of the haplogroup for systematic error detection. Refinement of mtDNA phylogeny with more diagnostic mutations would facilitate the detection of more errors in mtDNA sequence data since it is based on mutation motifs, and if haplogroup determination fails, a neighbourhood search for sequences in the available database could identify a subset of potentially closely related sequences, thereby allowing researchers to pinpoint errors in the sequence by comparing the sequence in question with a limited subset of the total database [4]. However, manual haplogroup estimation requires a thorough understanding of the worldwide mtDNA phylogeny, and database screening for systematic error detection requires high-quality databases that are publicly available.

The Human Mitochondrial DataBase (HmtDB) has been designed and implemented using automatically running bioinformatics tools to facilitate mtDNA haplogroup determination [6]. The HmtDB is a database of 1255 human mitochondrial genomes annotated with population and variability data that allows researchers to analyse their own mtDNA sequences and to automatically predict their haplogroups, yielding a list of haplogroups that match. However, haplogroup determination is carried out by comparing the complete mitochondrial genome sequences with the updated mtDNA haplogroup classification based on information of the coding-region single nucleotide polymorphisms (SNPs) for about 100 mtDNA haplogroups and subhaplogroups. Accordingly, haplogroup estimation using the HmtDB would be useful for researchers dealing with complete mitochondrial genome sequences, but would not be applicable to the detection of possible errors when researchers have only mtDNA control-region sequences.

As for the database, the EDNAP (European DNA Profiling Group) mtDNA Population Database (EMPOP) is notable because it was established through a collaborative project in order to provide reliable frequency estimates for routine forensic casework [7]. The EMPOP was designed to be a high-quality, Web-based mtDNA database where

primary sequence-lane data are permanently linked to compiled sequences, and phylogenetic quality control analyses are applied to data to check for errors [8]. Currently, the EMPOP contains 5173 high-quality mtDNA haplotypes that are classified into sub-Saharan African, West Eurasian, East Asian and Southeast Asian metapopulations, and thus enables users to assess the rarity of a forensic mtDNA haplotype in various populations. However, due to somewhat narrow query options and inconvenient method used to display the results, its query tool appears to be optimized for calculating frequency estimates for random matches rather than for database screening to detect possible mtDNA errors. Also, the EMPOP does not allow batch analyses. In addition to the accessibility of high-quality databases to generate reliable frequency estimates, the addition of batch analysis of mtDNA sequence data and the construction of a user's database would be greatly beneficial to forensic staff.

Here we present a Web-based bioinformatics resource called mtDNAmanager that provides a convenient interface supporting the management and quality analysis of mtDNA sequence data. The mtDNAmanager performs computations on mtDNA control-region sequences for estimating the most-probable mtDNA haplogroups, and retrieves similar sequences from a selected database. The aims of mtDNAmanager are (1) to allow researchers to automatically estimate the most-probable mtDNA haplogroups of their mtDNA control-region sequences, (2) to facilitate database screening with improved query tools and (3) to provide researchers with a convenient interface for managing and analysing their own data in batch mode. A query system in mtDNAmanager allows researchers to find sequences in the database that include queried nucleotide polymorphisms or to exhibit matches from either a selected population or the entire population. Inputted mtDNA sequences, which are either partial or whole mtDNA control-region sequences, are entered as differences relative to the revised Cambridge Reference Sequence (rCRS) [9]. During sequence searches, mtDNAmanager automatically estimates corresponding haplogroups for submitted data and calculates frequency estimates for random matches. Retrieved sequences are also annotated with the estimated haplogroup affiliation to highlight nucleotide polymorphisms that are specific to a certain group of mtDNA haplotypes. This application provides the first publicly available interface to automatically estimate the most-probable mtDNA haplogroups according to control-region mutation motifs, thereby facilitating data comparisons from a phylogenetic perspective.

Implementation

The mtDNAmanager interface was designed to allow researchers to easily query the database and immediately view results on a single page. The mtDNAmanager Web

interfaces were implemented using PHP and Asynchronous JavaScript and XML (AJAX). AJAX makes Web pages more responsive by exchanging small amounts of data with a server in the background, and thus mtDNAmanager Web pages do not have to be reloaded after each user request. This design aimed at increasing the interactivity, speed, functionality and usability of mtDNAmanager. The mtDNAmanager system is optimized for Internet Explorer version 6.0 or later.

mtDNAmanager has a multithreaded and multiuser SQL database management system designed and implemented using MySQL. The mtDNAmanager database currently comprises seven tables containing human mtDNA control-region sequences, data related to the samples and results of haplogroup estimation obtained by the mtDNA haplogroup-estimating resource that runs automatically (Figure 1).

The most-probable haplogroup of a given mtDNA sequence is estimated using a mathematical algorithm based on propositional logic via hierarchical verification of the presence or absence of haplogroup-specific diagnostic mutations. For that purpose, reliable controlregion mutation motifs (strings of characteristic/diagnostic mutations shared by descent) for the assignment of more than 400 mtDNA haplogroups and subhaplogroups were first identified based on well-characterized mtDNA phylogenies (see the list of mutation motifs at http:// mtmanager.vonsei.ac.kr/help/MutationMotifs.pdf) [10-49]. Mutation motifs of most of the haplogroups could immediately be read from the mtDNA tree. However, since each position of the mutation motif displays different mutation rates and homoplasy mutations are also observed in multiple motifs, individual diagnostic positions were weighted in each haplogroup background. To this end, polymorphisms of representative haplotypes

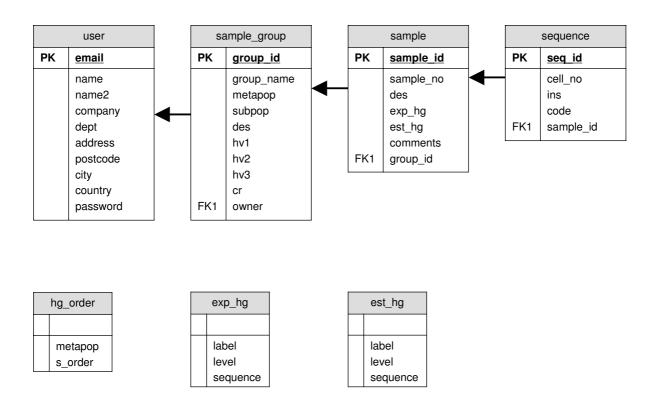


Figure I Relational database structure of mtDNAmanager.

allocated to the corresponding haplogroup or subhaplogroup were screened against other closely related mtDNA haplotypes. According to the mutation stability and specificity in each haplogroup background, individual diagnostic sites were classified into clearer diagnostic mutations and their accompanying mutations. To obtain mutation frequencies, published high-quality data were mostly used, but the data found on Internet resources were also used. The clear key diagnostic mutations of a certain haplogroup could be a single mutation or a combination of multiple mutations. They were selected from the polymorphic sites observed in every haplotype of the corresponding haplogroup (100% specificity) and mostly were not shared with any other haplogroups. On the other hand, accompanying mutations are also observed in almost every haplotype of the corresponding haplogroup (>95% specificity), but could include polymorphic sites observed in another haplogroups. Based on these haplogroup-specific mutation motifs, the bioinformatics tools of mtDNAmanager designates the "expected haplogroup" when a queried data sequence possesses clear diagnostic mutations, and designates the "estimated haplogroup" when the data indicate the presence of accompanying mutations additional to the clear diagnostic mutations.

This haplogroup-estimation workflow gives priority to certain haplogroups according to their degree of specificity to corresponding population groups. Therefore, the bioinformatics tools of mtDNAmanager have a hierarchy consisting of several levels of mutation motifs. Since all of the key diagnostic mutations equally have very high specificity for their corresponding haplogroups or subhaplogroups, the levels of mutation motifs in haplogroup designation were determined by the mutation stability of each mutation motif. Therefore, within a certain haplogroup branch, subhaplogroups have a higher priority than their root haplogroups, and among haplogroups of different branches, haplogroups associated with key diagnostic sites that have a lower mutation frequency in a certain population group have a higher priority. However, since mutation frequencies and specificities differ among population groups, the order of haplogroup designations in a hierarchical analysis of diagnostic mutations varied with the population group represented in the queried sequence. In addition, for two different haplogroups with identical key diagnostic mutations, the haplogroup with the highest prevalence in a certain population group has designation priority.

The data set used to test the bioinformatics tools of mtD-NAmanager contained more than 5000 mtDNA controlregion sequences whose haplogroup affiliations were available from previous publications or on the Internet. Actually, the bioinformatics tools of mtDNAmanager allowed more than 98% of mtDNA to be allocated to an appropriate mtDNA haplogroup or subhaplogroup. For data sets with haplogroup information confirmed by coding-region SNPs, relatively good concordance was also observed between the expected and reference haplogroups (e.g. the concordance of 140 African Americans, 273 Austrians and 593 Koreans was 99.3%, 99.3% and 99.7%, respectively) [34,50,51].

Results

Content and design of the open database

The current open database of mtDNAmanager contains 7090 mtDNA control-region sequences grouped in the following five subsets: African (n = 1388), West Eurasian (n = 2857), East Asian (n = 1557), Oceanian and Admixed (n = 1288) [50-62]. All of the mtDNA control-region sequences were annotated with estimated haplogroup affiliations using the mtDNAmanager bioinformatics tools. In cases where a data sequence had been assigned to a certain haplogroup in a previous study, relevant haplogroup information is provided in the output results.

The query system of mtDNAmanager retrieves sequences that include queried nucleotide polymorphisms from a selected population or the entire population group of its open database by default (exchangeable with the "include" or "match" settings). Since any combination of nucleotide polymorphisms or any partial control-region sequences can be analysed using the include setting, mtD-NAmanager is very useful for analysing partial sequences or comparing similar sequences that share the same nucleotide polymorphisms (Figure 2). With the alternative setting of match, mtDNAmanager also searches sequences that match the queried sequence data from the database. mtDNAmanager provides match options to select specific regions to be analysed [HV1 (hypervariable region 1): np 16024-16365; HV2: np 73-340; HV3: np 438-576; and control-region: np 16024-16569, np 1-576], ignore heteroplasmic insertions in poly C-stretches and permit mismatches in order to overcome differences in data reporting between laboratories due to variability in the analysis region, ambiguities with respect to mtDNA nomenclature and different treatment of length variants (insertion/deletion).

The frequency of a queried nucleotide polymorphism or sequence is estimated from the number of times (x) that it appears in a database of size n (that is generally known as the counting method) while taking into account uncertainty due to sampling errors. This frequency is therefore estimated as (x+2)/(n+2) [63], and is represented as the "match probability".

Design of user database

Upon registration, mtDNAmanager provides Web interfaces through which users can submit and store their data

	🔊 http://mtmanager,yon:	sei,ac,kr/sear	ch_sample,ph	p		~	😽 🗙 Live Search	2
* 4 6	mtDNAmanager – forensio	: mtDNA datab	ase			6	• 🔊 - 🖶 • 🔂 ШОЛ(Р) •	(<u>)</u> 도구(<u>0</u>) •
mtDN/	Amanager					E-mai	i:	Login
Ouery	Home					Passsword	t:	Register
arget Group	A STATE	Searching Optio	on					
Metapopulation:	East Asian	Sequence	16187 235				×	
BI-Asian	Forensic Sci Comm (2002) 4	Region		HV2 🗌 HV3 🗌			2	
Japanese	Int J Legal Med (2003) 117:2	Region	Match	HVZ 🛄 HV3 🗋	U Control Region			
Korean-Yonsei	Int J Legal Med (2006) 120:5	Option	lgnore hete Maximum r	eroplasmic inserti number of mismat queried nucleoti	ched nucleotides	s 0 🕶		
		Search	Reset	0	Frequency	1		
		Expected HG	A5a	Estimated HG	Н	Comments		
		Searching Resu	ılt					
		Searching Resu		No of Tar	oet Samples		Export Print	
		No. of Mate	lt ched Samples 19		get Samples	Match Probability 0.0353	Export Print	
: iroup Informatio		No. of Mate	ched Samples		3	Match Probability 0.0353	Export Print	np 438-576
177 - C		No. of Mate	ched Samples 19	5	93	Match Probability 0.0353		np 438-576 523d 524d
roup Informatio ame	Korean-Yonsei	No. of Mato Sample ID	ched Samples 19 Expected HG	5 Estimated HG	93 np 16024-1656	Match Probability 0.0353 59 6290 16319	np 001-437	A DECEMBER OF STREET
roup Informatio ame letapopulation	n Korean-Yonsei East Asian	No. of Mate Sample ID K017	ched Samples 19 Expected HG A5a	5 Estimated HG A5a	93 np 16024-1656 16187 16223 1 16187 16223 1	Match Probability 0.0353 59 6290 16319	np 001-437 73 235 263 315.1C	523d 524d
roup Informatio ame	Korean-Yonsei East Asian Korean	No. of Mate Sample ID K017 K122	ched Samples 19 Expected HG A5a A5a	5 Estimated HG A5a A5a	93 np 16024-1656 16187 16223 1 16187 16223 1 16187 16189 1	Match Probability 0.0353 59 6290 16319 6290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C	523d 524d 523d 524d
roup Informatio ame etapopulation ubpopulation	n Korean-Yonsei East Asian	No. of Mate Sample ID K017 K122 K178	hed Samples 19 Expected HG A5a A5a A5a	5 Estimated HG A5a A5a A5a	93 np 16024-1656 16187 16223 1 16187 16223 1 16187 16189 1 16187 16223 1	Match Probability 0.0353 69 6290 16319 6290 16319 6290 16319 6223 16290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C	523d 524d 523d 524d 523d 524d 523d 524d 523d 524d
roup Informatio ame letapopulation	Korean-Yonsei East Asian Korean Int J Legal Med (2006) 120:5-14	No. of Mate Sample ID K017 K122 K178 K184	hed Samples 19 Expected HG A5a A5a A5a A5a	5 Estimated HG A5a A5a A5a A5a	93 np 16024-1656 16187 16223 1 16187 16223 1 16187 16189 1 16187 16223 1 16187 16223 1	Match Probability 0.0353 6290 16319 6290 16319 6290 16319 6223 16290 16319 6223 16290 16319 6290 16296 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C	523d 524d 523d 524d 523d 524d 523d 524d 523d 524d
roup Informatic ame ietapopulation ubpopulation escription V1	Korean-Yonsei East Asian Korean Int J Legal Med (2006) 120:5-14	No. of Mato Sample ID K017 K122 K178 K184 K184 K191	hed Samples 19 Expected HG A5a A5a A5a A5a A5a A5a	5 Estimated HG A5a A5a A5a A5a A5a	93 np 16024-1656 16187 16223 1 16187 16223 1 16187 16189 1 16187 16223 1 16187 16223 1	Match Probability 0.0353 6290 16319 6290 16319 6290 16319 6223 16290 16319 6290 16319 6290 16319 6290 16319 6290 16319 6290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 146 195 235 263 309.1C 309.2C 31	523d 524d 523d 524d 523d 524d 523d 524d 523d 524d
roup Informatio ame etapopulation ubpopulation escription V1 V2	Korean-Yonsei East Asian Korean Int J Legal Med (2006) 120:5-14	No. of Mate Sample ID K017 K122 K178 K184 K191 K216	hed Samples 19 Expected HG A5a A5a A5a A5a A5a A5a A5a A5a	5 Estimated HG A5a A5a A5a A5a A5a A5a A5a	93 np 16024-1656 16187 16223 1 16187 16223 1 16187 16123 1 16187 16123 1 16187 16223 1 16187 16223 1 16187 16122 1 16187 16223 1	Match Probability 0.0353 6290 16319 6290 16319 6290 16319 6223 16290 16319 6290 16319 6290 16319 6290 16319 6290 16319 6290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 146 195 235 263 309.1C 309.2C 31 73 235 263 309.1C 315.1C	523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d
ame etapopulation ubpopulation escription V1 V2 V3	Korean-Yonsei East Asian Korean Int J Legal Med (2006) 120:5-14	No. of Mate Sample ID K017 K122 K178 K184 K191 K216 K222	hed Samples 19 Expected HG A5a A5a A5a A5a A5a A5a A5a A5a	5 Estimated HG A5a A5a A5a A5a A5a A5a A5a A5a A5a	93 p 16024-1656 16187 16223 1 16187 16223 1 16086 16187 1	Match Probability 0.0353 6290 16319 6290 16319 6290 16319 6223 16290 16319 6290 16296 16319 6290 16296 16319 6290 16319 16519 6223 16270 16290 16319 6290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 146 195 235 263 309.1C 309.2C 31 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C	523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d
ame etapopulation ubpopulation escription V1 V2 V3	Korean-Yonsei East Asian Korean Int J Legal Med (2006) 120:5-14	No. of Mate Sample ID K017 K122 K178 K184 K191 K216 K222 K263 K271	hed Samples Expected HG A5a A5a A5a A5a A5a A5a A5a A5a	5 Estimated HG A5a A5a A5a A5a A5a A5a A5a A5a A5a A5a	93 pp 16024-1656 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16086 16187 1 16187 16223 1	Match Probability 0.0353 6290 16319 6290 16319 6290 16319 6290 16319 6290 16296 16319 6290 16296 16319 6290 16319 16519 6290 16319 6290 16319 6290 16319 6290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 146 195 235 263 309.1C 309.2C 31. 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C	523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d
etapopulation ubpopulation ubpopulation escription V1 V2 V3 ontrol Region	Korean-Yonsei East Asian Korean Int J Legal Med (2006) 120:5-14	No. of Mate Sample ID K017 K122 K178 K184 K191 K216 K222 K263 K271 K308	hed Samples 19 Expected HG A5a A5a A5a A5a A5a A5a A5a A5a	5 Estimated HG A5a A5a A5a A5a A5a A5a A5a A5a A5a A5a	93 pp 16024-1656 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16187 16223 1 16086 16187 1 16187 16223 1 16187 16223 1	Match Probability 0.0353 39 6290 16319 6290 16319 6223 16290 16319 6220 16296 16319 6290 16296 16319 6290 16319 16519 6223 16270 16290 16319 6290 16319 6290 16319 6290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 195 235 263 309.1C 309.2C 315.1C 73 235 263 315.1C	523d 524d 523d 524d
oup Informatio ame etapopulation ubpopulation escription	Korean-Yonsei East Asian Korean Int J Legal Med (2006) 120:5-14	No. of Mate Sample ID K017 K122 K178 K184 K191 K216 K222 K263 K271	hed Samples Expected HG A5a A5a A5a A5a A5a A5a A5a A5a	5 Estimated HG A5a A5a A5a A5a A5a A5a A5a A5a A5a A5a	93 pp 16024-1656 16187 16223 1 16187 16223 1	Match Probability 0.0353 39 6290 16319 6290 16319 6223 16290 16319 6220 16296 16319 6290 16296 16319 6290 16319 16519 6223 16270 16290 16319 6290 16319 6290 16319 6290 16319	np 001-437 73 235 263 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 146 195 235 263 309.1C 309.2C 31. 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C 73 235 263 309.1C 315.1C	523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d 523d 524d

Query page of the mtDNAmanager. The query system – using the include setting by default – retrieves sequences that include the queried nucleotide polymorphisms from a chosen population or the entire population group of its open database. The results are displayed on the same page that the query was entered, and while displaying retrieved sequences, mtDNAmanager shows frequency estimates for random matches from a selected group and automatically estimated haplogroup affiliations for both submitted data and retrieved sequences.

in batch mode and search for sequences that show a match or include queried nucleotide polymorphisms from their databases as well as mtDNAmanager's open database. The sample system allows users to submit their own data in batch mode and store data in groups while simultaneously characterizing them by the automatically running haplogroup-estimating workflow (Figure 3A). The match system permits cross-matches of all sequence data between two selected groups as well as the retrieval of matched sequences for a sample of the user's database from their own database or mtDNAmanager's open database, which will facilitate casework in disasters involving many individuals (Figure 3B). In addition, the query system enables users to search sequences that show a match or include the queried nucleotide polymorphisms from

their own databases or mtDNAmanager's open database, thus assisting the analysis of the increasing amount of mtDNA data available worldwide.

Input

Input queries are entered as differences relative to the rCRS according to ISFG (International Society for Forensic Genetics) guidelines [64]. When a difference between sequence data and the rCRS is observed, only the site (which has a designated number) and nucleotide differing from the reference standard are recorded (e.g. "73G"). Insertions are recorded by first noting the site immediately 5' to the insertion followed by a decimal point and a "1" (for the first insertion), a "2" (if there is a second insertion) and so on, and then the nucleotide that is inserted is

	A	
	Δ	
	-	

(] - []	🔊 http://mtmanager, yon:	sei,ac,kr/samj	ple_manager, php	#	×	← X Naver	<u>ب</u> م
* * 6	mtDNAmanager – forensio	: mtDNA datab	ase			6 • 6	• • • •
mtDNA	manager					Welcome, C	
Sample	Match Query	Home				Logout	Change Profile
Group List		Sample List					
Add Edit I	Delete	Add Edit	Delete Import	Export	Print		
Name	Description	Sample ID	Expected HG	stimated HG	np 16024-16569	np 001-437	np 438-576
Demo-African	FSI: Genetics (2008) 2:e45-	AUT0500121	T2b T	2b	16126 16294 16296 16304 16519	73 263 315.1C	
Demo-Asian	ULM (2006) 120:5-14	AUT0500122	U5a1 U	5a1	16256 16270 16399 16519	73 263 315.1C	524.1A 524.2C
Demo-Casework	Skeletal Remains of Korean	AUT0500123	HV0 H	IV0	16298	64 72 195 263 309.1C 315.1C	
Demo-European	EMPOP	AUT0500124	U5a1 U	15a1	16256 16270 16399	73 263 315.1C	
Demo-Hispanic	FSI: Genetcis (2008) 2:e45-	AUT0500125	H1a H	1a	16051 16162 16291 16304 16519	73 263 315.1C	515
		AUT0500126	H1a H	1a	16162 16519	73 263 315.1C	
		AUT0500127	т т		16126 16241 16294 16304 16519	73 263 309.1C 309.2C 315.1C	
		AUT0500128	J1 J	1	16063 16069 16126	73 263 295 315.1C	462 489
		AUT0500129	Н6 Н	16	16362	239 263 309.1C 309.2C 315.1C	
		AUT0500130	Т1 Т	1	16126 16163 16172 16186 16189 16294 16298	73 263 309.1C 315.1C	
<	>	AUT0500131	K1c K	(1c	16224 16311 16320 16519	73 146 152 263 315.1C	498d
Group Informatio	n	AUT0500132	J1c J	1c	16069 16126 16519	73 185 188 228 263 295 315.1C	462 489
Name	Demo-European	AUT0500133	J1c1 J	1c1	16069 16126	73 228 263 295 315.1C	462 482 489
Metapopulation	West Eurasian	AUT0500134	U4c1 U	l4c1	16179 16356 16519	73 195 263 309.1C 315.1C	499
Subpopulation	Austrian	AUT0500135	H H		16299 16519	150 263 315.1C	
	EMPOP	AUT0500136	н н		16519	263 309.1C 309.2C 315.1C	
Description	And South and So	AUT0500137	U4a3 U	4a3	16265 16356 16362 16519	73 195 247 263 315.1C	499 524.1A 52
HV1		AUT0500138	J1c1 J	101	16069 16126	73 185 228 263 295 315.1C	462 482 489
HV2		AUT0500139	U4a3 U	4a3	16265 16356 16362 16519	73 195 247 263 315.1C	499 524.1A 52
HV3		AUT0500140	н		16239 16519	263 309.1C 315.1C	
Control Region	V				📢 🖣 123456 7 8910 🕨	×	
No. of Samples	273	<				Ϋ́.	

				Cross	-match li	nformatic	on				🚐 Print
				Worki	ng Group	Demo-Casework					
				R	egion	HV1 HV2					
				0	ption	Ignore heteropi	asmic insertions in the poly C-stretches	s, Maximum number of m	ismatched nucleotides: 1		
				Tarrow	t Group	Demo-Asian					
				-	get Samples						
				no. of tar	get samples	000					
				Match	ed Samp	ole List					
				Sample	Expected			Sequence		Comments	Description
SmtDNAmanager – forensic mtl	INA databas	e – Windov	vs Internet E:	Kpl ID B001	HG B4o1o	HG B4o1o	161830 16189 16217 16311 73 146		<u></u>		B4o1o
🕒 🕤 🔹 🙋 http://mtmanager.yon	sei.ac.kr/match	h_group.php#		8001	Matched Sar		K257 K371	0 100 100 214 203 310.1	-		84010
🊖 🔅 🌈 mtDNAmanager – forensio	mtDNA databa	ase			No. of Match Match Proba		2 0.0087				
mtDNAmanager				8002	D4/3	D4/3	18223 18382 73 283 315 10 489 49	00.10			Dde
				0.002	Matched Sar		K038 K044 K058 K082 K085 K149 I	K181 K196 K227 K238 H	240 K241 K255 K273 K305 K309 K342 K3	347 K388 K370 K4	
Sample Match Query	Home				No. of Match		K405 K476 K515 K549 K569 K577 I 30	K579			
/orking Group	Working Sample				Match Proba		0.0538				
lame Description			Estimated HG								
lemo-African FSI Genetics (2008) 2:e45-				161: B003	G2a3 Matched Sar	G2a3	16189 16223 16278 16292 16362 7 K529	73 260 263 309.1C 309.2	C 315.1C 489 524.1A 524.2C		G2#3
lemo-Asian ULM (2006) 120:5-14		D4n		1622	No. of Match		1				
Demo-Casework Skeletal Remains of Korean Demo-European EMPOP	B010 B011	D4b1		1623 1618	Match Proba	ibility:	0.0050				
Demo-Hispanic FSt Genetols (2008) 2:e45-	B011 B012	2 pre-Z		1601 8004	D4/G	D4/G	16223 16362 73 152 263 309 1C 30		244		D4
		BSb		1614	Matched Sample ID :		K056 K058 K065 K134 K149 K181 I				
		B4d1		1611	No. of Match		18				
					Match Proba	ibility:	0.0303				
	<			8005	D4j1	D4j1	16184 16223 16311 16362 73 162 :	263 315.1C 489			D4j1
11 >	Matching Option	1			Matched Sar		K272				
arget Group	Region	1 HV1 1 H	1V2 🗌 HV3 🔲	Con	No. of Match	red Samples:	1				
My mtDNA	Option	Ignore heterop	lasmic insertions in	n the					🏹 😜 인터넷		€ 100%
Open DB: African 🗸	0000	Maximum numi	ber of mismatched	nucleo							
ame Description	Match	Match Al	Worldwide Fr	· · · ·							
eme-African FSI: Genetics (2008) 2:e45-	Matching Result										
Nemo-Asian ULM (2006) 120:5-14	No. of Match	hed Sample.	No. of Targe	15 .4	Match Pro	obability	Export Print				
lemo-Casework Skeletal Remains of Korean lemo-European EMPOP	7	7			0.015	51					
emo-European EMPOP emo-Hispanic FSt Genetois (2008) 2:e45-			Estimated HG					np 438-576			
cine rapatio racidenetos (2000) 2.640-		F1b					73 152 249d 263 309.1C 315.1C	523d 524d			
		F1b					73 152 249d 263 309.1C 309.2C 315.1				
		F1b F1b					73 152 249d 263 315.1C 73 152 249d 263 309.1C 309.2C 315.1	523d 524d			
		F1b F1b					73 152 2496 263 309.1C 309.2C 315.1 73 152 2496 263 315.1C	523d 524d 523d 524d			
	N#33							and a second second			
	K437	E1b									
		F1b F1b					73 152 249d 263 309.1C 309.2C 315.1C 73 152 249d 263 315.1C	523d 524d 523d 524d			

Sample and match pages of the mtDNAmanager. Upon registration, mtDNAmanager provides Web interfaces that allow users to submit and store their own data in batch mode and search sequences that show a match or include queried nucleotide polymorphisms from their own databases as well as mtDNAmanager's open database. (A) The sample system allows users to manage and analyse large amounts of data in batch mode. Data are characterized whilst being imported by the automatically running haplogroup-estimating workflow, and accordingly, each sample is annotated with the most-probable mtDNA haplogroup (both expected and estimated haplogroups). (B) The match system permits cross-matching of all sequence data between two selected groups as well as retrieval of matched sequences for a sample of the own database of the user or mtDNAmanager's open database. Clicking the "Match All" button will display cross-matched results in a new pop-up window.

recorded (e.g. "315.1C"). Deletions are recorded by listing the missing site followed by a "d" (i.e. "249d"). For convenience, transition mutations can be recorded by listing the site and omitting the indication of the nucleotide difference. However, transversion mutations are recorded in every case (e.g. "73" versus "73C") in which the nucleotide differs from the reference standard. Polymorphic sites can be separated using a space, return or comma character. Sequence searches are allowed to show matches even when no data (i.e. no differences relative to the rCRS) have been submitted, since some Europeans possess mtDNA control-region sequences identical to the rCRS. The frequencies of nucleotide polymorphisms that are identical to the rCRS can also be obtained by entering the site and nucleotide polymorphisms of the rCRS or by entering the site with "=" (e.g. "73A" and "73=") using the include setting.

- Input sequence example 1: 16304C 73G 249d 263G 315.1C
- Input sequence example 2: 16304 73 249d 263 315.1C

To import data through the sample system in batch mode, the sample group should first be generated by the user. User-defined sample groups are added to the group list by clicking the "Add" button and entering their names and properties. Then, batch input files are prepared in a text file to be imported into a specific, user-defined group. Input files are initially prepared as Excel files that contain both the mtDNA sequence data and descriptions of the properties of the data (see examples at http://mtman ager.yonsei.ac.kr/help/Examples.xls). The mtDNA sequence data are entered using the same method as input queries. The Excel file is then saved as a text file (separated by tabs) that is imported to a specific user-defined group of the sample system. Input sequences can also be uploaded one by one using the "Add" button on the sample list.

Output

Results from mtDNAmanager are displayed on the same page on which the query was submitted (Figure 2). While showing retrieved sequences, mtDNAmanager shows frequency estimates for random matches from a selected group and the automatically estimated haplogroup affiliations for submitted data. Queried nucleotide polymorphisms that are either identical to the rCRS or entered as IUPAC (International Union of Pure and Applied Chemistry) codes for point heteroplasmy are indicated as such under "Comments". Frequency estimates for all of the population groups in the database can be obtained by clicking the "Worldwide Frequency" button, and the cross-match result can be obtained by clicking the "Match All" button. The retrieved sequences are displayed with estimated haplogroup affiliations (both expected and estimated haplogroups), nucleotide polymorphisms and, if available, the haplogroup affiliations obtained from previous reports. Therefore, mtDNAmanager should facilitate the comparison of sequences that share the same nucleotide polymorphisms from a phylogenetic perspective. In addition to the Web-page presentation tools, retrieved sequences can be exported as an Excel file for user convenience.

Discussion

The mtDNAmanager can be used to manage large amounts of mtDNA data as well as to estimate the quality of mtDNA data and compare such data with similar sequences from a phylogenetic perspective. The application provides systematic routines for error detection and strategies for screening mtDNA databases by enabling researchers to automatically estimate the most-probable mtDNA haplogroups and search the database with two alternative settings (include and match).

In particular, the phased designation of haplogroups (i.e. expected haplogroups and estimated haplogroups) facilitates systematic error detection by allowing the respective confirmation of the presence of clearer key diagnostic mutations and accompanying mutations. Specifically, if a certain mtDNA sequence was annotated with the same expected and estimated haplogroups, this means that the sequence possessed the complete mutation motif for the corresponding haplogroup. Likewise, if a sequence was annotated with only the expected haplogroup, this suggests a lack of accompanying mutations for the expected haplogroup, which was determined by the presence of key diagnostic mutations (Figure 4A). From an mtDNA phylogenetic perspective, this would mean that a given mtDNA haplotype is located at a previously unsampled interior node of the tree such as a back mutation [4]. Accordingly, in this case, it would be necessary to recheck the entire set of haplogroup-specific mutation sites in a given sequence data. More importantly, data without both of the haplogroup affiliations or that showed discordance between expected and estimated haplogroups would also imply a need to recheck the sequence for possible errors due to contamination or a sample mix-up during the sequencing and documentation process (Figure 5). Since mtDNA haplotypes that could only be obtained from separate amplifications of several smaller fragments - such as those found in highly degraded samples - are prone to these errors, using mtDNAmanager to confirm the absence of these errors after data generation will help to authenticate the sequence data in highly degraded samples. For user convenience, currently identified haplogroup-specific control-region mutation motifs for more than 400 haplogroups are available on the mtDNAmanager home page. From a phylogenetic perspective, mtDNA

Δ	mtDNA	mana	ger									me, Collea	
	Sample	Match	Query	Home							Logo	ut Chang	e Profile
	Target Group			Searching Opti	n								
	O My mtDNA				(16226)162	94 16296 16	304 73 263	309.10	315.10		-		
	Open DB:	West Eura	sian 🗸	Sequence	\smile					USA.	CAU.000	272 (20	02)
	Name	Description		Dealer				1000)				
	Austrian		Int (2007) 166:	Region		HV2 HV3	Control Re	gion					
	FBI-Caucasian	Forensic Sci	Comm (2002) 4		Match	17 17 17	-		·	Match op	tions		
	Finnish		Int (2007) 172:	Option	Ignore het	teroplasmic inse	rtions in the po	y C-streto	ches 🕑	•			
	German		ed (2006) 120:3			number of misn			×)				
	Hungarian	Int J Legal Me	ed (2007) 121:3			e queried nucle		ISM					
				Search		Worldw			T2b hap	ologroup	specific	mutati	on m
	(0 0 0)			Expected HG		Estimated HC		-	: 16126	-16294-16	296-163	04	
	<		>	Searching Resu								-	
	Group Information				ched Samples	No. of T			Match Probabil	ty Expor	t Print		
	Name	FBI-Caucasia		Sample ID	10	I Paramateria and	1655	0000	0.0072	np 001-4	-		438-576
	Metapopulation		n	FRA.CAU.004	Expected HG	T2b	16126 1629		10204		97 19.1C 315.1C	np -	+30-576
	Subpopulation			USA.CAU.004		T20	16126 162				9.1C 315.1C		
	Description	Forensic Sci ((2002) 4 Onlin		USA.CAU.014		T2b	16126 1625				9.1C 315.1C		
	HV1	(2002) 4 Onin		USA.CAU.019		T2b	16126 1625				9.10 315.10		
	HV2			USA.CAU.027		T2b	16126 1629				9.10 315.10		
	HV2			USA.CAU.028	8 T2b	T2b	16126 1629	4 16296	16304	73 263 30	9.1C 315.1C		
	Control Region	-		USA.CAU.075		T2b	16126 1629				9.10 315.10		
	No. of Samples			USA.CAU.075		T2b	16126 1629				9.10 315.10		
	No. of Samples	1000		USA.CAU.085	2 T2b	T2b	16126 1629	4 16296	16304	73 263 30	9.10 315.10		
							C mtDNAmar	iager - fi	orensic mtDNA da	atabase - Windows	Internet Explorer		
							nttp://mtma	nager, yons	sel, ac, kr/sample/sam	ple_search_freq_print.ph	o?group_id=2558tgNar	ne=FBI-Caucasian&	seq=162268.hv
							Comments						
	Amanager						Worldwide	Freque	псу				
	Match Query						Population	Group		Description	No. of Matched	No. of Population Samples	Match Prot
rget Group		Searching Op	(16226)				Admixed	name			oungres	1200	0.001
Open DB:	West Eurasian 💌	Sequence					Dubalan, UAE	Dubaian FBI-Hispanio	FSI: Gene	rics (2008) 2 e8-e10	0	249	0.003
ane	Description Forensic Sci Int (2007)	Region	✓ HV1 ✓ HV	/2 🗌 HV3 🔲 Contr	ol Region			US-Hapanio	FSI: Genetics (2008) 2:	Comm (2002) 4 Online e19-e23, FSI: Genetics (2008	de o	253	0.002
ustrian BI-Caucasian	Forensic Sci Comm (20)	02) 4	O Match		· · · ·		Venezuelan	Venezuelan		etics (2008) In Press	0	100	0.019
innish	Forensic Sci Int (2007)			oplasmic insertions in t mber of mismatched nu		4	African				0	1388	0.001
lerman ungarian	Int J Legal Med (2006) 1 Int J Legal Med (2007) 1	121:5	Include the q	ueried nucleotide polyr	norphism		African African American	FBI-African		Comm (2002) 4 Online	•	1148	0.001
		Search	Reset	Worldwide Freque	ncy		American Kenyan	Kenyan		ed (2004) 118-294-305		100	0.019
nuo Informatio		Expected HG Searching Re		Estimated HG H	Comments		East Asian				0	1667	0.001
ame	FBI-Caucasian			No. of Target San	ples Mat	ch Probability	East Aslan Japanese	FBI-Aslan Japanese	Forenaid Sci Int J Legal M	Comm (2002) 4 Online led (2003) 117:218-225	0	763	0.002
	West Eurasian		0	1655		0.0012	Korean	Korean-	Int J Legal	Med (2008) 120.6-14	0	693	0.003
ubpopulation		Sample ID	Expected HG	Estimated HG np 16	024-16569		Oceanian				101	10	1.000
escription	Forensic Sci Comm (2002) 4 Online						Opeanian	Oceanian	Cor	ming Soon	٥	0	1.000
V1 V2							West Eurasian				0	2957	0.000
V3			World	wide F	requen	cy:	Austrian Caucasian	Austrian FBI- Caucasian		Int (2007) 166:164-175 Comm (2002) 4 Online	0	273	0.007
ontrol Region			0 from	n 7090 H	nanlotv	nes	Finnish	Caucasian Finnish	Forenaio Sol	Int (2007) 172 171-178	0	200	0.001
o, of Samples	1655		0 11011	1 1 0 3 0 1	lupioty	pes	German	German	Int J Legal Med (2005 (2007	5) 120 310-314, Forensic Sci I 7) 172 218-224	0 11	313	0.006
							Hungarian	Hungarian	Int J Legal M	led (2007) 121:377-383	٥	418	0.004
							완료				🖬 🚱 2	[터넷	
											Woley	ome, Collea	au 0
	mtDNA	mana	ger										
	Sample	Match	Query	Home							Log	out Chan	ge crone
	Target Group			Searching Opti	on								
	O My mtDNA				(16126)162	94 16296 1	5304 73 263	309.1	C 315.1C 🖊			_	
	Open DB:	West Eura	isian 🗸	Sequence	\smile					Corre	cted US	A.CAU.	00027
	Name	Description		Basias				S					
	Austrian	Forensic Sci	Int (2007) 166:	Region	Match	HV2 🗌 HV3	Control Re	gion					
	FBI-Caucasian		Comm (2002) 4			teroplasmic inse	allows in th		abaa 🔽				
	Finnish		Int (2007) 172:	Option		number of misr							
	German		ed (2006) 120:3			e queried nucle							
	Hungarian	Int J Legal Me	ed (2007) 121:3	Search	Reset		otide polymorpi ide Frequency						
	<		>	Expected HG	Tak	Worldw Estimated H			nments				
			-		120	Counated Ho	120	Con	nmonta				
	Group Information			Searching Res								_	
	Name	Austrian		No. of Mat	ched Samples	No. of T	arget Samples		Match Probabi	ity Expo	rt Print		
	Metapopulation		n	Come: D	3	1 Patro and	273	0.000	0.0182	np 001-4	07		100 570
	Subpopulation			Sample ID 12H6	Expected HG	Estimated H	3 np 16024-1		16304 16519		37 53 309.1C 315.1C	np	438-576
	Description	Forensic Sci (2007) 166:10	Int 84-	f2H6 m3E2	T2b T2b	T2b T2b			16304 16519	100.000	53 309.1C 315.1C 09.1C 315.1C		
	Seachp1011	175		IT YES	T2b	T2b			16304 16362 1651		09.10 315.10 09.10 315.10		
	HV1				120	120	10111101	10294	10200 10304 103	10 203 3	00.10 010.10		
	HV2							_					
	HV2 HV3				T2b	16126	16204	162	06-1620	4			
					T2b:	16126	16294	162	96-1630	4			

An example of clerical error detection using mtDNAmanager. (A) Using the match setting, USA.CAU.000272 [1,52] was analysed with the option that permits mismatch. The results showed that the sequence had no match with European populations, but the retrieved similar sequences containing mismatches all belonged to the T2b haplogroup. In addition, the original sequence was annotated with only the expected haplogroup, T2b. It was suspected that the mutation at 16226 resulted from a clerical error because the mutation motif for the T2b haplogroup suggested that the sequence lacks the 16126 mutation. (B) By clicking "Worldwide Frequency", the rarity of the 16226 mutation was investigated and the results showed that none of the 7090 mtDNA sequences bears this mutation. (C) The corrected sequence [52] was annotated with both the expected and estimated haplogroups of T2b.

	A	mtDN	Amar	nager										come, Colleague
		Sample	Match	Que		lome								Change Profile
		Target Group			Se	arching Optic	n							
		 My mtDN/ Open DB: 		ו ו	s	equence	16126 161 195 263 2	87 16189 16 64 315.10	223 16264 16	270 16278	3 16311 73	A		.000074 (200
		Name FBI-African	Descript Forensic	tion c Sci Comm (J	2002) 4	egion	HV1 Match	HV2 □ HV3	Control Region	1)			.000074 (200
		Kenyan 4 US-African		al Med (2004 netics (2008)		otion	ignore het		tions in the poly C atched nucleotide:			Match	options	
		Targ	et dat	abase		Search	O Include the Reset		tide polymorphisn le Frequency	1				
						spected HG	R1 X2 M3	Estimated HG	R1 X2 M3	Comments				
		<			> Se	arching Resu								
		Group Informa			2		hed Samples		arget Samples	Ma	o.0017	/	Export Prin	nt
		Name	FBI-Afric	an	s				np 16024-165	39	0.0017	no	001-437	np 438-576
		Metapopulatio					1 - 14	1				1.15		
ntDN/	Aman	ager							W	elcome, C	olleague			
	Match	-	Home							Logout	Change Profile			
et Group			Searching Optio					HV1	sequer	ice				
My mtDNA Open DB:	African	×	Sequence	16126 1618	87 16189 1	6223 16264	16270 16278 16	531 1. 🥒	~					
10	Description		Region		HV2 HV2	Control Re	aion		2					
African yan		ci Comm (2002) 4 Med (2004) 118:2		O Match										
A frican	FSI: Genet	ics (2008) 2:e45-	Option			ertions in the po matched nucleot	ly C-stretches 🗹							
				Include the	queried nucl	otide polymorph		Include	e option					
			Search Expected HG	Reset		G R1 X M3	Comments							
			Searching Resul		conduct R		DNAmai	100.02						Welco
p Informatio				thed Samples	No. of	Target 1	ple Match	0	11.000					Log
e population	US-African		1 Sample ID	Expected HG	Estimated H	G np Target 0		Query	Home Searching Option				HV2 sec	uence
	African An		US.AFR-007	L1b1	L161	18 🔿 My	mtDNA		Sequence	3 143 146	152 195 263	264 315.10	J	<u></u>
iption	FSt Geneti (2008) 2:e-		US.AFR-008	L161	L161	16 💿 Op 16 Name	Descript					20. 2012/ 11		2
			US.AFR-008	L161	L161	FBLAF	ican Forensid	Sci Comm (2002) 4	1	HV1 W HV	/2 🗌 н∨з 🔲	Control Region		
			US.AFR-009	L161	L161	16 16 US-Afr	i Int J Leg ican FSt Ger	al Med (2004) 118:2 etics (2008) 2:e45		Ignore hetero	oplasmic insertio			latch options
ol Region			US.AFR-098 US.AFR-099	L161 L161	L161 L161	16					mber of mismatcl ueried nucleotide			-
Samples	140		US.AFR-100	L161	L1b1	16			Search	Reset	Worldwide F	requency		
									Expected HG		Estimated HG		Comments	
					B	Group In	nformation	>	Searching Result No. of Matcheo	Samples	No. of Targe	et Samples	Match Probability	Export Print
						Name	US-Afric	an	7		14)	0.0634	
							ulation African		Sample ID E: US.AFR-021		Estimated HG		294 16309 16390	np 001-437 73 143 146 152 195 263 31
						Descrip	FSI: Gen	etics			L2a1	16223 16278 16	294 16309 16390 16399	73 143 146 152 195 263 31
							(2008) 2	:e45-e48			L2a1 L2a1		294 16309 16390 6223 16278 16294 16309	73 143 146 152 195 263 31 16390 73 143 146 152 195 263 31
									US.AFR-025	al	L2a1	16223 16278 16	294 16309 16390	73 143 146 152 195 263 31
						HV1 HV2	0							
						HV2 HV3			US.AFR-026	2a1	L2a1 L2a1		294 16309 16390 16519 223 16278 16294 16309 16	73 143 146 152 195 263 31
						HV2 HV3 Control			US.AFR-026	2a1	L2a1		294 16309 16390 16519	73 143 146 152 195 263 31
						HV2 HV3 Control	Region		US.AFR-026	2a1	L2a1		294 16309 16390 16519 223 16278 16294 16309 16	73 143 146 152 195 263 31 3390 73 143 146 152 195 263 31
	\mathbf{C}	mtDN	Amai	ıager		HV2 HV3 Control	Region		US.AFR-026	2a1	L2a1		294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 3390 73 143 146 152 195 263 31 come, Colleague
	0	Sample	Amar	nager Quer		HV2 HV3 Control No. of S	Region C		US.AFR-026	2a1	L2a1		294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 3390 73 143 146 152 195 263 31
	0	Sample Target Group	Match			HV2 HV3 Control No. of S	Region IV amples 140		US.AFR-026	2a1 2a1	L2a1 L2a1	16189 18192 16	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 3390 73 143 146 152 195 263 31 come, Colleague
		Sample Target Group	Match	Que	St	HV2 HV3 Control No. of S	Region 66 iamples 140		US.AFR-026 US.AFR-027 2	2a1 2a1	12a1 12a1 3 16311 73	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
	0	Sample Target Group O My mtDN/ Open DB	Match A African	Quer	se s	HV2 HV3 Control No. of S arching Optic	Region 07 iamples 140	47 263 315.	US AFR-026 US AFR-027 223 16264 16 10 357	201	12a1 12a1 3 16311 73	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group	Match A African Descript	Quer	Se Se	HV2 HV3 Control No. of S Iome arching Optic	Region Ø amples 140	47 263 315.	US.AFR-026 US.AFR-027 2	201	12a1 12a1 3 16311 73	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group My mtDNJ Open DB: Name FBI-African Kenyan	Match A A Descript Forensic Int J Leg	Quer tion to Sci Comm (2 jai Med (2004	St S	HV2 HV3 Control No. of 5 arching Optic equence	Region 000 iampica 140 16126 161 189 195 2 00 Hv1 00 00 Match	47 263 315. нv2 🗌 нv3	US.AFR-026 US.AFR-027 223 16264 16 10 357	270 16276	16311 73 Corre	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group My mtDNA Open DB Name FBI-African	Match A A Descript Forensic Int J Leg	Quer	St S	HV2 HV3 Control No. of S arching Optic	Region 06 iamples 140 16126 161 189 195 2 19 Hv1 1 6 Match Ignore het	47 263 315. HV2 🔲 HV3 eroplasmic inser	US_AFR-026 US_AFR-027 223 16264 16 10 357 Control Regio: tions in the poly 0	270 16276	16311 73 Corre	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group My mtDNJ Open DB: Name FBI-African Kenyan	Match A A Descript Forensic Int J Leg	Quer tion to Sci Comm (2 jai Med (2004	St S	HV2 HV3 Control No. of 5 arching Optic equence	Region (27) Region (27)	47 263 315. HV2 HV3 HV3 eroplasmic inser number of misma	US.AFR-026 US.AFR-027 223 16264 16 10 357	270 16276	16311 73 Corre	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group My mtDNJ Open DB: Name FBI-African Kenyan	Match A A Descript Forensic Int J Leg	Quer tion to Sci Comm (2 jai Med (2004	St S	HV2 HV3 Control No. of 5 arching Optic equence	Region (27) Region (27)	47 263 315. HV2 HV3 HV3 eroplasmic inser number of misma e queried nucleo	US_AFR-026 US_AFR-027 223 16264 16 10 357 Control Region tions in the poly 0 atched nucleotider	270 16276	16311 73 Corre	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group My mtDNJ Open DB: Name FBI-African Kenyan	Match A A Descript Forensic Int J Leg	Quer tion to Sci Comm (2 jai Med (2004	2002) 4 2002) 4 2118:2 2:e45 0	HV2 HV3 Control No. of S arching Optic squence sgion Search	Region 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	47 263 315. HV2 HV3 HV3 eroplasmic inser number of misma e queried nucleo	US_AFR-026 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-026 LUS_AFR	270 16276	2a1 2a1 3 16311 73 Corre	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group My mtDNJ Open DB: Name FBI-African Kenyan	Match A A Descript Forensic Int J Leg	Quer tion to Sci Comm (2 jai Med (2004	✓ St 2002) 4 3) 118:2 2:e45- 0	HV2 HV3 Control No. of S arching Optic squence sgion Search	Region (Kr amples 140 16126 151 189 195 2 V HV1 (V Match Ignore het Ignore het Ignore het L1b	47 263 315. HV2 HV3 eroplasmic inser number of misma e queried nucleo Worldwid	US_AFR-026 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-027 LUS_AFR-026 LUS_AFR	270 16276	2at 2at 3 16311 73 Corre	152 182 1	294 16309 16390 16519 223 16278 16294 16309 16 Weld	73 143 146 152 195 263 31 73 143 146 152 195 263 31 come, Colleague ggout Change Profile
		Sample Target Group My mtDNJ Open DB: Name FBI-African Kenyan	Match A Descript Forensic Int J Leg FSt Gen	Quer tion to Sci Comm (2 jai Med (2004	✓ St 2002) 4 3) 118:2 2:e45- 0	HV2 HV3 Control No. of S arching Optic equence agion Search spected HG arching Resu	Region V V iampies 140	47 263 315. HV2 HV3 eroplasmic inser number of misme e queried nucleo Worldwid Estimated HG No. of Ta	US AFR-020 US AFR-027 US AFR-026	270 16276	22a1 3 16311 73 Corre	152 182 1 ected	294 16309 16390 16519 223 16278 16294 16309 16 Weld	72 143 146 152 155 263 31 72 143 146 152 155 263 31 come, Colleague goout Change Profile
		Sample Target Group My mtDNJ Open DJ PBI-African Kenyan US-African	Match A Descript Forensic Int J Leg FSt Gen	Quer	2002) 4 2002) 4) 118:2 2 e45 0	HV2 HV3 Control No. of S arching Optic equence egion otion Search pected HG arching Result No. of Matc	Repen Kr amples 140	47 263 315. HV2 HV3 eroplasmic inser number of misma s queried nucleo Worldwid Estimated HG No. of Ta	US AFR-027	ai 270 16276 -stretches [a 1 Comments	2 16311 73 Corre	152 182 1 152 182 1 ected	284 16309 16396 16519 223 16378 16284 16309 16 Weld USA.AFR USA.AFR	17 143 146 152 155 263 31 173 143 146 152 155 263 31 come, Colleague gout Change Profile .0000074
		Sample Target Group My mDNU Target Group Open DB Name FBLAfrican Kenyan US-African Group Informe Name Netapopulatio	Match A A C A C A C A C A C A C A C A C A C	Quer	✓ S ✓ S 2002) 4) 118:2 2045 0 E S S S	HV2 HV3 Control No of S Arching Optice arching Optice aguence agion Search pected HG arching Resu No. of Mate ample ID	Report KY ampes 140 (16126, 161 (189, 195, 2) (2) HV1 (2) (3) HV1 (2) (3) HV1 (2) (4) HV1 (2) HV1 (2) (4) HV1 (2)	47 263 315. HV2 HV3 eroplasmic inser number of misma s queried nucleo Worldwid Estimated HG No. of Ta Estimated HG	US_AFR-027	270 1627E	22a1 2 16311 73 Corre 2 2 2 2 2 2 2 2 2 2 2 2 2	152 182 1 ected	244 16309 16390 16519 222 16272 16284 16309 16 Welk 1857 0 USA.AFR USA.AFR	27 143 146 152 155 283 31 27 143 146 152 155 283 31 come, Colleague agout Change Profile .0000074
		Sample Target Group My mDNJ O Open DB Name FBI-A frican Kenyan US-A frican Group Informa Name	Match A A C A C A C A C A C A C A C A C A C	Quer a Quer tion 5 Sci Comm () jai Med (2004) hetics (2008) an can American	✓ S ✓ S 2002) 4) 118:2 ✓ C S S U	HV2 HV3 Control No. of S arching Optic equence egion otion Search pected HG arching Result No. of Matc	Repen Kr amples 140	47 263 315. HV2 HV3 eroplasmic inser number of misma s queried nucleo Worldwid Estimated HG No. of Ta	US_AFR-027 US_AFR-027 US_AFR-027 US_AFR-027 US_AFR-027 US_AFR-027 US_AFR-027 US_AFR-027 US_AFR-026	270 16276	22a1 3 16311 73 COrrec 2 4 4 4 4 4 4 4 4 4 4 4 4 4	152 182 1 ected	224 16309 16390 16519 223 16372 16284 16309 16 223 16372 16284 16309 16 21851 USA.AFR USA.AFR	17 143 146 152 155 263 31 173 143 146 152 155 263 31 come, Colleague gout Change Profile .0000074
		Sample Target Group My mDNU Target Group Open DB Name FBLAfrican Kenyan US-African Group Informe Name Netapopulatio	Match A A C A C A C A C A C A C A C A C A C	Quer tion c Sci Comm (2 sci	 Second second sec	HV2 HV3 Control No. of S Arching Optic squence option Search hipected HG arching Resu No. of Mato S.AFR-100	Repain KY amples 140 16126 161 189 155 161 180 161 180 155 161 180 155 161 180 155 161 16	47 263 315. HV2 HV3 eroplasmic inser number of misma e queried nucleo Worldwid Estimated HG No. of Ta Estimated HG L1b1	US AFR-027	270 16276 	22a1 23 16311 73 Corre Corre 24 25 26 26 26 27 26 26 27 26 26 27 26 26 26 26 26 26 26 26 26 26	152 182 1 ected 1675	Export Prime 122 16209 16390 16516 Weld USA.AFR USA.AFR	72 143 146 152 155 283 31 73 143 146 152 155 283 31 come, Colleague gout Change Profile
		Sample Target Group My mDNU O Open DB Name FBI-African Kenyan US-African Group Informa Name Metapopulatio Subpopulatio	Hatch A A A A A A A A A A A A A A A A A A A	Quer tion c Sci Comm (2 sci	Second secon	HV2 HV3 Control No. of S arching Optic aquence agoin Search Search No. of Mark Search No. of Mark Search No. of Mark Search S. AFR-100 S. AFR-103 S. AFR-103	Region (K) Region (K) Region (K) Region (K) (K) (K) (K) (K) (K) (K) (K)	47 263 315. HV2 HV3 eroplasmic inser- number of misma ourreid nucleo Vordowid Estimated HG L1b1 L1b1 L1b1 L1b1	US_AFR-027	270 16276 -stretches [Comments 1 39 6189 16223 6189 16223 6189 16223 6189 16223 6189 16223 6189 16223	231 15311 73 Correct C	152 182 1 ected / / / / / / / / / / / / / / / / / / /	Export Print 001-437 152 162 165 169 195 USA.AFR 001-437 152 162 165 169 195 152 162 165 169 195 152 162 165 169 195	173 143 146 152 155 263 31 273 143 146 152 155 263 31 COME, Colleague gout Change Profile .0000074 101 1247 263 31 5236 5246 1247 263 31 5236 5246 1247 263 31 5236 5246 1247 263 31 5236 5246
		Sample Target Group Target Group O My mDNu O My mDNu O My mDNu FBLAfrican Kenyan US-African Kenyan US-African Name Metapopulaton Subpopulaton Description HV1 HV2	Match A A Construction African African African African African African African Cools 2 C C C C C C C C C C C C C C C C C C C	Quer tion c Sci Comm (2 sci	> Second	HV2 HV3 Control Result	Report KGY ampies 140 16126.161 189.195.2 V HV1 V Match Ignore het Maximum O Include th Include th Includ	47 263 315. HV2 HV3 HV3 eroplasmic inser number of misma s queried nucleo Worldwick Estimated HG No. of Ta L1b1 L1b1 L1b1 L1b1 L1b1 L1b1	US_AFR-027	270 1627ec stretches 1 commenta 19 10 10 10 10 10 10 10 10 10 10	2.2a1 3 15311 73 Correct w atch Probabilit 0.0775 16264 16270 16264 16270 16264 16270 16264 16270	152 182 1 ected / / / / / / / / / / / / / / / / / / /	Export 199 199 199 199 199 199 199 199 199 19	17 143 146 152 155 253 31 27 143 146 152 155 253 31 come, Colleague poput Change Profile .0000074 .1000074
		Sample Target Group O MymdDNL O O Open DB Name FBLAfrican Kenyan US-African Group Informa Netapopulation Subpopulation Description HV1 HV2 HV2 HV3	Match A C C African Descript Forenic Int J Leg FSit Gen US-Afric US-African African African African African African C2008) 2 C C C C C C C C C C C C C C C C C C C	Quer tion c Sci Comm (2 sci	> Second	HV2 HV3 Control No. of S arching Optic aquence agoin Search Search No. of Mark Search No. of Mark Search No. of Mark Search S. AFR-100 S. AFR-103 S. AFR-103	Region (K) Region (K) Region (K) Region (K) (K) (K) (K) (K) (K) (K) (K)	47 263 315. HV2 HV3 eroplasmic inser- number of misma ourreid nucleo Vordowid Estimated HG L1b1 L1b1 L1b1 L1b1	US.AFR-020 US.AFR-027	ar 270 16275 -stretches 1 -stretches 6169 16223 6169 16223 6169 16223 6169 16223	231 [231] 73 Correct C	152 182 1 ected / / / / / / / / / / / / / / / / / / /	Export Print 122 1827 81234 18309 16 Welk L USA.AFR USA.AFR 001-437 152 182 1051 109 195 152 182 1051 109 195 152 182 1051 109 195 152 182 1851 195 195 152 182 1851 195 247 152 182 1851 195 247	173 143 146 152 155 263 31 273 143 146 152 155 263 31 COME, Colleague gout Change Profile .0000074 101 1247 263 31 5236 5246 1247 263 31 5236 5246 1247 263 31 5236 5246 1247 263 31 5236 5246
		Sample Target Group Target Group O My mDNu O My mDNu O My mDNu FBLAfrican Kenyan US-African Kenyan US-African Name Metapopulaton Subpopulaton Description HV1 HV2	Match A A A A A A A A A A A A A A A A A A A	Quer tion c Sci Comm (2 sci	Second Se	H/2 H/2 H/2 H/2 H/2 H/2 H/2 H/2 H/2 H/2	Region 007 Region 007 amples 140 16126 161 189 195 2 V HV1 (V) Maximum Maxim	47 263 315. HV2 HV3 eroplasmic inser- number of misma eroplasmic inser- number of misma eroplasmic inser- Worldwid Estimated HG L1b1 L1b1 L1b1 L1b1 L1b1 L1b1 L1b1 L1b1	US.AFR.027 US.AFR.027 US.AFR.027 Control Region tions in the poly C tabled nucleotide tide polymorphism Re Frequency L1b Tig 18024-165 16128 16187 16128 16187 16128 16187 16128 16187	at at 270 162702 -stretches 1 - 1 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0	2.2a1 3 16311 73 Correc 2 3 16311 73 Correc 2 3 16311 73 Correc 4 4 4 4 4 4 4 4 4 4 4 4 4	152 182 1 ected , m 1627673 1627673 1627673 1627673 1627673	Export Price 199 199 199 199 199 199 199 199 199 19	12 143 146 152 155 283 31 23 143 146 152 155 283 31 COME, Colleague soout Change Profile .0000074 tt 1247 283 31 5236 5246 1247 283 31 5236 5246 1247 283 31 5236 5246 1247 283 31 5236 5246 1247 283 31 5236 5246 1283 5151 5236 5246

An example of artificial recombination error detection using mtDNAmanager. (A) USA.AFR.000074 [52,65] did not show a match with African populations or affiliations with either haplogroup. (B) Since the data set was known to be prepared from separate amplification of hypervariable regions, a possible artificial recombination was checked using query tools. The results showed that the HVI sequence was only evident in the LIbI haplogroup, and the HV2 sequence was only evident in the L2aI haplogroup. (C) The corrected sequence [52] was annotated with both the expected and estimated haplogroups of LIb. control-region sequence information might not be sufficient for assigning certain mtDNAs into respective haplogroups as reliably as the coding-region information, but the control-region mutation motifs of mtDNAmanager will at least suggest candidate sites or regions that need reinvestigation.

Frequency estimates and sequences retrieved using the include setting indicate the rarity of a nucleotide polymorphism in databases and show similar sequences that share queried nucleotide polymorphisms. Accordingly, mtD-NAmanager can reveal unusual, private mutations (Figure 4B) and suggest a subset of potentially close relatives annotated with estimated haplogroup affiliations even when the haplogroup estimation of a queried sequence data fails (Figure 5B). This will highlight nucleotide polymorphisms that are specific to the retrieved group of mtDNA haplotypes and help to distinguish sites that should be analysed further. In other cases, retrieved sequences with estimated haplogroup affiliations will contribute to completing and refining haplogroup classification by revealing mutation sites that are specific to a new branch of phylogeny. Therefore, to improve mtDNA database screening, we will continue to collect and integrate high-quality mtDNA control-region sequence data that are publicly available.

In addition, mtDNAmanager provides a convenient interface that allows users to construct and analyse their own databases. Therefore, users can collect high-quality data from public databases (e.g. EMPOP) or direct sequencing results to construct their own databases. mtDNAmanager will suggest the most-probable mtDNA haplogroups for all of the sequences in the database, allowing users to also easily estimate the quality of the database. Researchers will therefore be able to select and use the most appropriate database for error detection based on their own evaluation of the quality of the available databases.

Conclusion

The mtDNAmanager supports the management and quality analysis of mtDNA sequence data using software that performs computations on mtDNA control-region sequences for estimating the most-probable mtDNA haplogroups. mtDNAmanager will help in checking the quality of data and facilitate data comparisons from a phylogenetic perspective by displaying information – estimated haplogroup affiliations and nucleotide polymorphisms – of all sequences on a single page. In addition, mtDNAmanager provides researchers with a convenient interface for managing and analysing their own data in batch mode. Therefore, this tool could be very useful for population, medical and forensic studies that involve mtDNA analysis.

Availability and requirements

Project name: A Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences

Project home page: http://mtmanager.yonsei.ac.kr

Operating system(s): Microsoft Windows

Programming language: PHP, Asynchronous JavaScript and XML

Other requirements: Optimized for Internet Explorer version 6.0 or later

Any restrictions to use by non-academics: None

Authors' contributions

KJS initiated the concept of mtDNAmanager. IS and KJS developed the major modules of mtDNAmanager, and IS, HYL and KJS designed the graphical user interface. KJS reviewed and tested the software. HYL wrote this manuscript, and EH, SBC and WIY helped to draft it.

Acknowledgements

This work was supported by a Korean Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No.

M10740030002-07N4003-00210), and grants from the Ministry of National Defense Agency for Killed In Action Recovery and Identification (MAKRI).

References

- 1. Bandelt HJ, Salas A, Bravi C: **Problems in FBI mtDNA database.** Science 2004, **305:**1402-1404.
- Budowle B, Polanskey D: FBI mtDNA database: a cogent perspective. Science 2005, 307:845-847.
- Salas A, Carracedo A, Macaulay V, Richards M, Bandelt HJ: A practical guide to mitochondrial DNA error prevention in clinical, forensic, and population genetics. *Biochem Biophys Res Commun* 2005, 335:891-899.
- Bandelt HJ, Lahermo P, Richards M, Macaulay V: Detecting errors in mtDNA data by phylogenetic analysis. Int J Legal Med 2001, 115:64-69.
- 5. Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V: The fingerprint of phantom mutations in mitochondrial DNA data. *Am J Hum Genet* 2002, **71:**1150-1160.
- Attimonelli M, Accetturo M, Santamaria M, Lascaro D, Scioscia G, Pappadà G, Russo L, Zanchetta L, Tommaseo-Ponzetta M: HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. BMC Bioinformatics 2005, 6:S4.
- 7. The European DNA Profiling Group (EDNAP) MtDNA Population Database (EMPOP) [http://www.empop.org]
- Parson W, Dür A: EMPOP-A forensic mtDNA database. Forensic Sci Int Gene 2007, 1:88-92.
- Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N: Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 1999, 23:147.
- Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonné-Tamir B, Sykes B, Torroni A: The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 1999, 64:232-249.
- 11. Helgason A, Hickey E, Goodacre S, Bosnes V, Stefánsson K, Ward R, Sykes B: mtDNA and the islands of the North Atlantic: esti-

mating the proportions of Norse and Gaelic ancestry. Am J Hum Genet 2001, 68:723-737.

- Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R: The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol 2002, 19:1737-1751.
- Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A: The making of the African mtDNA landscape. Am J Hum Genet 2002, 71:1082-1111.
- Kong QP, Yao YG, Liu M, Shen SP, Chen C, Zhu CL, Palanichamy MG, Zhang YP: Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. Hum Genet 2003, 113:391-405.
- Kong QP, Yao YG, Sun C, Bandelt HJ, Zhu CL, Zhang YP: Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. Am J Hum Genet 2003, 73:671-676.
- Reidla M, Kivisild T, Metspalu E, Kaldma K, Tambets K, Tolk HV, Parik J, Loogväli EL, Derenko M, Malyarchuk B, Bermisheva M, Zhadanov S, Pennarun E, Gubina M, Golubenko M, Damba L, Fedorova S, Gusar V, Grechanina E, Mikerezi I, Moisan JP, Chaventré A, Khusnutdinova E, Osipova L, Stepanov V, Voevoda M, Achilli A, Rengo C, Rickards O, De Stefano GF, Papiha S, Beckman L, Janicijevic B, Rudan P, Anagnou N, Michalodimitrakis E, Koziel S, Usanga E, Geberhiwot T, Herrnstadt C, Howell N, Torroni A, Villems R: Origin and diffusion of mtDNA haplogroup X. Am J Hum Genet 2003, 73:1178-1190.
- Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogväli EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A: 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 2004, 75:910-918.
- Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R: Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. Am J Hum Genet 2004, 75:752-770.
- Loogväli EL, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, Metspalu E, Tambets K, Reidla M, Tolk HV, Parik J, Pennarun E, Laos S, Lunkina A, Golubenko M, Barac L, Pericic M, Balanovsky OP, Gusar V, Khusnutdinova EK, Stepanov V, Puzyrev V, Rudan P, Balanovska EV, Grechanina E, Richard C, Moisan JP, Chaventre A, Anagnou NP, Pappa KI, Michalodimitrakis EN, Claustres M, Golge M, Mikerezi I, Usanga E, Villems R: Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia. *Mol Biol Evol* 2004, 21:2012-2021.
- Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MT, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R: Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet 2004, 5:26.
- Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, Zhang YP: Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 2004, 75:966-978.
- Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti AS, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Qasim Mehdi S, Torroni A, McElreavey K: Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. Am J Hum Genet 2004, 74:827-845.
- Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A: The African diaspora: mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 2004, 74:454-465.
- Tanaka M, Cabrera VM, González AM, Larruga JM, Takeyasu T, Fuku N, Guo LJ, Hirose R, Fujita Y, Kurata M, Shinoda K, Umetsu K, Yamada Y, Oshida Y, Sato Y, Hattori N, Mizuno Y, Arai Y, Hirose N, Ohta S, Ogawa O, Tanaka Y, Kawamori R, Shamoto-Nagai M, Maruyama W, Shimokata H, Suzuki R, Shimodaira H: Mitochondrial genome variation in eastern Asia and the peopling of Japan. Genome Res 2004, 14:1832-1850.
- 25. Yao YG, Kong QP, Wang CY, Zhu CL, Zhang YP: Different matrilineal contributions to genetic structure of ethnic groups in the silk road region in china. *Mol Biol Evol* 2004, **21**:2265-2280.

- Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, Fornarino S, Magri C, Scozzari R, Babudri N, Santachiara-Benerecetti AS, Bandelt HJ, Semino O, Torroni A: Saami and Berbers – an unexpected mitochondrial DNA link. Am J Hum Genet 2005, 76:883-886.
- Friedlaender J, Schurr T, Gentz F, Koki G, Friedlaender F, Horvat G, Babb P, Cerchio S, Kaestle F, Schanfield M, Deka R, Yanagihara R, Merriwether DA: Expanding Southwest Pacific mitochondrial haplogroups P and Q. *Mol Biol Evol* 2005, 22:1506-1517.
- Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M: Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 2005, 308:1034-1036.
- Merriwether DA, Hodgson JA, Friedlaender FR, Allaby R, Cerchio S, Koki G, Friedlaender JS: Ancient mitochondrial M haplogroups identified in the Southwest Pacific. Proc Natl Acad Sci USA 2005, 102:13034-13039.
- Wen B, Li H, Gao S, Mao X, Gao Y, Li F, Zhang F, He Y, Dong Y, Zhang Y, Huang W, Jin J, Xiao C, Lu D, Chakraborty R, Su B, Deka R, Jin L: Genetic structure of Hmong-Mien speaking populations in East Asia as revealed by mtDNA lineages. *Mol Biol Evol* 2005, 22:725-734.
- Behar DM, Metspalu E, Kivisild T, Achilli A, Hadid Y, Tzur S, Pereira L, Amorim A, Quintana-Murci L, Majamaa K, Herrnstadt C, Howell N, Balanovsky O, Kutuev I, Pshenichnov A, Gurwitz D, Bonne-Tamir B, Torroni A, Villems R, Skorecki K: The matrilineal ancestry of Ashkenazi Jewry: portrait of a recent founder event. Am J Hum Genet 2006, 78:487-497.
- Hill C, Soares P, Mormina M, Macaulay V, Meehan W, Blackburn J, Clarke D, Raja JM, Ismail P, Bulbeck D, Oppenheimer S, Richards M: Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol 2006, 23:2480-2491.
- Kong QP, Bandelt HJ, Sun C, Yao YG, Salas A, Achilli A, Wang CY, Zhong L, Zhu CL, Wu SF, Torroni A, Zhang YP: Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. *Hum Mol Genet* 2006, 15:2076-2086.
- Lee HY, Yoo JE, Park MJ, Chung U, Kim CY, Shin KJ: East Asian mtDNA haplogroup determination in Koreans: haplogrouplevel coding region SNP analysis and subhaplogroup-level control region sequence analysis. *Electrophoresis* 2006, 27:4408-4418.
- Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM, Coudray C, Santachiara-Benerecetti AS, Semino O, Bandelt HJ, Torroni A: The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. Science 2006, 314:1767-1770.
- Pierson MJ, Martinez-Arias R, Holland BR, Gemmell NJ, Hurles ME, Penny D: Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. Mol Biol Evol 2006, 23:1966-1975.
- Sun C, Kong QP, Palanichamy MG, Agrawal S, Bandelt HJ, Yao YG, Khan F, Zhu CL, Chaudhuri TK, Zhang YP: The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol Biol Evol* 2006, 23:683-690.
- Thangaraj K, Chaubey G, Singh VK, Vanniarajan A, Thanseem I, Reddy AG, Singh L: In situ origin of deep rooting lineages of mitochondrial macrohaplogroup 'M' in India. BMC Genomics 2006, 7:151.
- Thanseem I, Thangaraj K, Chaubey G, Singh VK, Bhaskar LV, Reddy BM, Reddy AG, Singh L: Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. BMC Genet 2006, 7:42.
- Derenko M, Malyarchuk B, Grzybowski T, Denisova G, Dambueva I, Perkova M, Dorzhu C, Luzina F, Lee HK, Vanecek T, Villems R, Zakharov I: Phylogeographic analysis of mitochondrial DNA in northern Asian populations. Am J Hum Genet 2007, 81:1025-1041.
- Friedlaender JS, Friedlaender FR, Hodgson JA, Stoltz M, Koki G, Horvat G, Zhadanov S, Schurr TG, Merriwether DA: Melanesian mtDNA complexity. PLoS ONE 2007, 2:e248.
- 42. Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S, Richards

M: A mitochondrial stratigraphy for island southeast Asia. Am J Hum Genet 2007, **80:**29-43.

- 43. Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA, Shen P, Oefner P, Renfrew C, Villems R, Forster P: **Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis.** *Proc Natl Acad Sci USA* 2007, **104**:8726-8730.
- 44. Reddy BM, Langstieh BT, Kumar V, Nagaraja T, Reddy AN, Meka A, Reddy AG, Thangaraj K, Singh L: Austro-Asiatic tribes of Northeast India provide hitherto missing genetic link between South and Southeast Asia. PLoS ONE 2007, 2:e1141.
- 45. Roostalu U, Kutuev I, Loogväli EL, Metspalu E, Tambets K, Reidla M, Khusnutdinova EK, Usanga E, Kivisild T, Villems R: Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. Mol Biol Evol 2007, 24:436-448.
- Achilli A, Perego UA, Bravi CM, Coble MD, Kong QP, Woodward SR, Salas A, Torroni A, Bandelt HJ: The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. *PLoS ONE* 2008, 3:e1764.
- Malyarchuk B, Grzybowski T, Derenko M, Perkova M, Vanecek T, Lazur J, Gomolcak P, Tsybovsky I: Mitochondrial DNA phylogeny in Eastern and Western Slavs. Mol Biol Evol 2008, 25:1651-1658.
- Malyarchuk BA, Perkova MA, Derenko MV, Vanecek T, Lazur J, Gomolcak P: Mitochondrial DNA variability in Slovaks, with application to the Roma origin. Ann Hum Genet 2008, 72:228-40.
- Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, Macaulay V, Bulbeck D, Oppenheimer S, Lin M, Richards MB: Climate change and post-glacial human dispersals in southeast Asia. *Mol Biol Evol* 2008, 25:1209-1218.
- 50. Brandstätter A, Niederstätter H, Pavlic M, Grubwieser P, Parson W: Generating population data for the EMPOP database-an overview of the mtDNA sequencing and data evaluation processes considering 273 Austrian control region sequences as example. Forensic Sci Int 2007, 166:164-175.
- Just RS, Diegoli TM, Saunier JL, Irwin JA, Parsons TJ: Complete mitochondrial genome sequences for 265 African American and U.S. "Hispanic" individuals. Forensic Sci Int Gene 2008, 2:e45-e48.
- Monson KL, Miller KWP, Wilson MR, DiZinno JA, Budowle B: The mtDNA population database: an integrated software and database resource for forensic comparison. Forensic Sci Commun 2002, 4:.
- 53. Maruyama S, Minaguchi K, Saitou N: Sequence polymorphisms of the mitochondrial DNA control region and phylogenetic analysis of mtDNA lineages in the Japanese population. Int J Legal Med 2003, 117:218-225.
- Brandstätter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, Parsons TJ: Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. Int J Legal Med 2004, 118:294-306.
- 55. Brandstätter A, Klein R, Duftner N, Wiegand P, Parson W: Application of a quasi-median network analysis for the visualization of character conflicts to a population sample of mitochondrial DNA control region sequences from southern Germany (UIm). Int J Legal Med 2006, 120:310-34.
- 56. Lee HY, Yoo JE, Park MJ, Chung U, Shin KJ: Mitochondrial DNA control region sequences in Koreans: identification of useful variable sites and phylogenetic analysis for mtDNA data quality control. Int J Legal Med 2006, 120:5-14.
- 57. Alshamali F, Brandstätter A, Zimmermann B, Parson W: Mitochondrial DNA control region variation in Dubai, United Arab Emirates. Forensic Sci Int Gene 2008, 2:e9-e10.
- Hedman M, Brandstätter A, Pimenoff V, Sistonen P, Palo JU, Parson W, Sajantila A: Finnish mitochondrial DNA HVS-I and HVS-II population data. Forensic Sci Int 2007, 172:171-178.
- 59. Irwin J, Egyed B, Saunier J, Szamosi G, O'callaghan J, Padar Z, Parsons TJ: Hungarian mtDNA population databases from Budapest and the Baranya county Roma. Int J Legal Med 2007, 121:377-383.
- Tetzlaff S, Brandstätter A, Wegener R, Parson W, Weirich V: Mitochondrial DNA population data of HVS-I and HVS-II sequences from a northeast German sample. Forensic Sci Int 2007, 172:218-224.
- 61. Lander N, Rojas MG, Chiurillo MA, Ramirez JL: Haplotype diversity in human mitochondrial DNA hypervariable regions I-III

in the city of Caracas (Venezuela). Forensic Sci Int Gene 2008, 2:e61-e64.

- 62. Saunier JL, Irwin JA, Just RS, O'Callaghan J, Parsons TJ: Mitochondrial control region sequences from a U.S. "Hispanic" population sample. Forensic Sci Int Gene 2008, 2:e19-23.
- Balding DJ, Nichols RA: DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int 1994, 64:125-140.
- Carracedo A, Bär W, Lincoln P, Mayr W, Morling N, Olaisen B, Schneider P, Budowle B, Brinkmann B, Gill P, Holland M, Tully G, Wilson M: DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. Forensic Sci Int 2000, 110:79-85.
- 65. Bandelt HJ, Salas A, Lutz-Bonengel S: Artificial recombination in forensic mtDNA population databases. Int J Legal Med 2004, 118:267-273.

