OPEN ACCESS Freely available online PLoS GENETICS

The Genographic Project Public Participation Mitochondrial DNA Database

Doron M. Behar[1,2], Saharon Rosset[3], Jason Blue-Smith[4], Oleg Balanovsky[5], Shay Tzur[1], David Comas[6], R. John Mitchell[7], Lluis Quintana-Murci[8,9], Chris Tyler-Smith[10], R. Spencer Wells[4*], The Genographic Consortium

1 Genomics Research Center, Family Tree DNA, Houston, Texas, United States of America, **2** Molecular Medicine Laboratory, Rambam Health Care Campus, Haifa, Israel, **3** Data Analytics Research Group, IBM T. J. Watson Research Center, Yorktown Heights, New York, United States of America, **4** The Genographic Project, National Geographic Society, Washington, District of Columbia, United States of America, **5** Research Centre for Medical Genetics, Russian Academy of Medical Sciences, Moscow, Russia, **6** Unitat de Biologia Evolutiva, Universitat Pompeu Fabra, Barcelona, Spain, **7** Department of Genetics, La Trobe University, Bundoora, Australia, **8** Institut Pasteur, Paris, France, **9** CNRS, URA3012, Paris, France, **10** The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom

The Genographic Project is studying the genetic signatures of ancient human migrations and creating an open-source research database. It allows members of the public to participate in a real-time anthropological genetics study by submitting personal samples for analysis and donating the genetic results to the database. We report our experience from the first 18 months of public participation in the Genographic Project, during which we have created the largest standardized human mitochondrial DNA (mtDNA) database ever collected, comprising 78,590 genotypes. Here, we detail our genotyping and quality assurance protocols including direct sequencing of the mtDNA HVS-I, genotyping of 22 coding-region SNPs, and a series of computational quality checks based on phylogenetic principles. This database is very informative with respect to mtDNA phylogeny and mutational dynamics, and its size allows us to develop a nearest neighbor–based methodology for mtDNA haplogroup prediction based on HVS-I motifs that is superior to classic rule-based approaches. We make available to the scientific community and general public two new resources: a periodically updated database comprising all data donated by participants, and the nearest neighbor haplogroup prediction tool.

Citation: Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, et al. (2007) The Genographic Project public participation mitochondrial DNA database. PLoS Genet 3(6): e104. doi:10.1371/journal.pgen.0030104

Introduction

The plethora of human mitochondrial DNA (mtDNA) studies in recent years has made this molecule one of the most extensively investigated genetic systems. Its abundance in human cells; uniparental, nonrecombining mode of inheritance; and high mutation rate compared to that of the nuclear genome, has made mtDNA attractive to scientists from many disciplines. Knowledge of mtDNA sequence variation is rapidly accumulating, and the field of anthropological genetics, which initially made use of only the first hypervariable segment (HVS-I) of mtDNA, is currently being transformed by complete mtDNA genome analysis [1]. While contemporary combined sources offers approximately 65,000 HVS-I records (Oleg Balanovsky, unpublished data) and over 2,000 complete mtDNA sequences, difficulties remain in standardizing these published data, as they report varying sequence lengths and different coding-region SNPs, and apply any number of methodologies for classifying haplotypes into informative haplogroups (Hgs) [2,3]. For example, some studies have defined the HVS-I range to comprise nucleotides 16093–16383 [4], some 16024–16365 [5], some adhered to the widely accepted definition of 16024–16383 [6], while others extended the reported range to include positions such as 16390 and 16391 due to their predictive value in identifying certain specific clades [7,8]. Even more serious is the problem of Hg assignment, which, in the absence of complete sequence data, is best achieved by genotyping a combination of coding-region biallelic polymorphisms. Forensic studies (which comprise a significant portion of the existing dataset) and many population studies published before 2002 have predicted Hgs based on the HVS-I motif alone, thereby ignoring the occurrence of homoplasy and back mutations [2,9]. Moreover, it has been shown that many published mtDNA databases contain errors that distort phylogenetic and medical conclusions [10–15]. Therefore, it has become abundantly clear that a phylogenetically reliable and systematically quality-controlled database is needed to serve as a standard for the comparison of any newly reported data whether medical, forensic, or anthropological [7].

The Genographic Project, begun in 2005, allows any individual to participate by purchasing a buccal swab kit. Male samples are analyzed for a combination of male specific Y chromosome (MSY) short tandem repeat loci and SNPs. Female samples undergo a standard mtDNA genotyping process that includes direct sequencing of the extended HVS-I (16024–16569) and the typing of a panel of 22 coding-region biallelic sites. Results are returned anonymously through the Internet (http://www.nationalgeographic.com/genographic) after passing a multi-layered quality check process in which phylogenetic principles are applied

Editor: Gil McVean, University of Oxford, United Kingdom

Received February 12, 2007; **Accepted** May 11, 2007; **Published** June 29, 2007

Abbreviations: Hg, haplogroup; HVS-I, first hypervariable segment; MSY, male specific Y chromosome; mtDNA, mitochondrial DNA; NN, nearest neighbor; rCRS, revised Cambridge reference sequence; w-NN, weighted nearest neighbor

* To whom correspondence should be addressed. E-mail: genopubs@ngs.org

Author Summary

The Genographic Project was launched in 2005 to address anthropological questions on a global scale using genetics as a tool. Samples are collected in two ways. First, the project comprises a consortium of ten scientific teams from around the world united by a core ethical and scientific framework that is responsible for sample collection and analysis in their respective region. Second, the project promotes public participation in countries around the world and anyone can participate by purchasing a participation kit (Video S1). The mitochondrial DNA (mtDNA), typed in female participants, is inherited from the mother without recombining, being particularly informative with respect to maternal ancestry. Over the first 18 months of public participation in the project we have built up the largest to date database of mtDNA variants, containing 78,590 entries from around the world. Here, we describe the procedures used to generate, manage, and analyze the genetic data, and the first insights from them. We can understand new aspects of the structure of the mtDNA tree and develop much better ways of classifying mtDNA. We therefore now release this dataset and the new methods we have developed, and will continue to update them as more people join the Genographic Project.

throughout, and which is supported by a specialized laboratory information management system. HVS-I haplotypes are reported based on the direct sequencing results. Hgs are defined by a combined use of the 22-SNP panel results and the HVS-I haplotypes. Following successful typing and reporting of the genotyping results, each participant may elect to donate his or her anonymous genotyping results to Genographic's research database. The magnitude of the project and its worldwide scale offer a unique opportunity to create a large, rapidly expanding, standardized database of HVS-I haplotypes and corresponding coding-region SNPs. Here, we report our experience from genotyping 78,590 public participants' mtDNAs during the first 18 months of the project. First, we describe our genotyping process and quality check measures and our considerations in designing them. Second, we report the unique insights that the standardized database supports with respect to estimation of the frequencies of transversions, transitions, heteroplasmies, indels, back mutations, and homoplasy occurring in both the HVS-I and the coding-region biallelic sites. Third, we present a new nearest neighbor (NN) –based methodology developed for Hg labeling, suggest it as an Hg prediction tool for validation of both new and previously reported databases, and demonstrate its superior performance over rule-based approaches, given a sufficiently large reference database. Finally, we make available to the scientific community and general public two new resources: a database (which will be periodically updated) containing the data donated by participants as an open source research database, and the NN analytical tool, which allows the comparison of any comparable data to the entire expanding Genographic dataset for quality control and predictive purposes.

Results

A total of 78,590 mtDNA samples were analyzed, of which 41,552; 5,046; 15,021; and 16,971, respectively, were genotyped with a panel of 10, 20, 21, and 22 SNPs. We excluded from our analysis samples in which the SNP genotyping result was summarized as "uninformative" and heteroplasmic positions. Therefore, we consider three different versions of the database: (1) The *entire* database: 76,638 samples. (2) The *reference* database made up of the subset of samples genotyped with the panel of 22 SNPs, currently comprising 16,609 samples. This reference database is expanding, as all new samples are genotyped with these 22 SNPs. (3) The *consented* database, released to the public with the participants' consent. So far, data from 21,141 samples (7,174 of which belong to the reference database) have been donated to the scientific community and are reported in Dataset S1 (for future updates of the database please see: http://www.nationalgeographic.com/genographic). Analyses using complete haplotype information are restricted to this dataset. The database presents the following information about each sample: a sequential serial number (different from the anonymous Genographic participant ID number), the number of SNPs genotyped, results of all genotyped SNPs, the Hg inferred from the SNP genotyping, the final Hg assigned in the current study, and the HVS-I haplotype.

Table 1. Genotyping Parameters of the Reference Database

Methodology	Result	Proportion
Sequencing	Samples containing polymorphism T16189C	13.9%
	Available bidirectional sequencing (16189T)	99.4%
	Available forward sequencing (16189T)	99.7%
	Available backward sequencing (16189T)	99.7%
SNP genotyping	Overall successful Hg labeling	98.5%
	Complete panel of 22 SNPs available	83.2%
	Panel labeled with a Hg	94.7%
	Panel labeled inconsistent	3.2%
	Panel labeled uninformative	2.1%

doi:10.1371/journal.pgen.0030104.t001

Genotyping Parameters

The genotyping parameters associated with the reference database are presented in Table 1. The overall first pass genotyping success rate for the entire process including DNA extraction, sequencing, and SNP genotyping was 98.5%. The average time needed to complete the first genotyping attempt was 31 days. All samples were attempted with bidirectional sequencing, but 13.9% contained the transition T16189C, which blocks the sequencing reaction beyond this position, and these provided data from only one strand. Of the remaining 86.1% reported samples, forward, backward, and bidirectional sequencing were successful in 99.7%, 99.7%, and 99.4% of the samples, respectively. The alternative forward sequencing primer (Table S1) was used once, while the use of an alternative reverse sequencing primer was mandated in approximately 0.15% of the samples. A total of 83.2% of the samples was successfully genotyped for the complete panel of 22 SNPs. The success rate of inferring an Hg by this SNP panel was 94.7%, while 3.2% and 2.1% of the samples were labeled as inconsistent or uninformative, respectively. The total number of samples from project inception in which post-DNA-extraction sample mix-up was suspected due to clear nonconcordance between Hg labeling, as suggested by the HVS-I motif and the SNP genotyping, was

Table 2. Hg Frequencies

Hg by SNPs	Final Hg	Entire Database	Reference Database	Consented Data	Consented Data (22 SNPs)
L0/1	L0*	1	0	0	0
	L0a*	6	2	2	1
	L0a1	149	49	49	21
	L0a2	46	9	13	5
	L0d	11	2	1	0
	L0f	1	1	1	1
	L1*	62	18	17	6
	L1b	343	73	107	34
	L1c*	41	3	9	1
	L1c1	130	23	32	10
	L1c2	118	29	35	14
	L1c3	44	12	13	8
L0/L1 total		952	221	279	101
L2	L2*	5	0	1	0
	L2a	780	194	212	80
	L2b	133	39	43	16
	L2c	185	39	51	22
	L2d	47	6	14	4
	L5	1	1	0	0
L2 total		1,151	279	321	122
L3*	L3*	141	28	46	12
	L3b	183	35	46	11
	L3d	241	47	66	19
	L3e1*	117	34	34	12
	L3e1a	40	8	6	1
	L3e1b	12	4	5	2
	L3e2*	139	23	41	9
	L3e2b	172	30	45	10
	L3e3	100	26	24	7
	L3e4	30	6	10	4
	L3f	178	33	43	12
	L3g	20	5	4	3
L3* total		1,373	279	370	102
M*	M*	1,025	255	290	110
	M1	98	21	39	13
	Z	24	6	7	4
M* total		1,147	282	336	127
C	C	825	229	228	101
D	D	651	147	193	72
N*	N*	225	18	40	5
	N9a	77	18	23	6
N* total		302	36	63	11
N1*	N1*	6	2	1	1
	N1a	160	29	59	19
	N1b	442	106	119	42
	N1c	57	12	18	7
N1* total		665	149	197	69
I	I	1,769	421	475	170
W	W	1,252	282	407	144
X	X	1,080	237	338	111
A	A	1,413	361	373	130
R*	R*	186	26	53	13
	R1	1	1	0	0
	R2	9	4	1	1
	R5	7	5	2	2
	R6	11	2	4	1
R* total		214	38	60	17
U*	U*	958	186	262	89
	U1	22	5	11	3
	U1a	131	25	43	14
	U1b	71	17	18	5
	U2	885	195	264	92
	U3	437	102	143	51
	U4	1,669	348	460	158
	U5	480	110	150	59
	U5a	907	196	249	83
	U5a1	667	130	183	56
	U5a1a	2,168	483	600	206

Table 2. Continued.

Hg by SNPs	Final Hg	Entire Database	Reference Database	Consented Data	Consented Data (22 SNPs)
	U5b	884	177	236	76
	U5b1	144	35	53	26
	U6	3	0	0	0
	U6a	65	14	18	5
	U6a1	100	27	34	16
	U6b	35	11	12	7
	U7	190	49	70	29
U* total		9,816	2,110	2,806	975
K	K	6,264	1,335	1,755	613
R0*	R0*	9	3	1	1
	R0a	178	42	54	15
R0* total		187	45	55	16
HV*	HV*	1,460	318	378	133
	HV1	281	85	86	35
HV* total		1,741	403	464	168
H	H	29,267	6,232	7,779	2,606
V	V	2,185	471	609	183
J	J*	3,928	771	1,107	351
	J1	596	128	164	57
	J1a	645	152	171	59
	J1b	93	26	29	12
	J1b1	593	113	182	47
	J2	506	122	139	58
J total		6,361	1,312	1,792	584
T	T*	1,481	320	420	151
	T1	1,394	294	395	139
	T2	3,000	631	838	266
	T3	331	59	97	25
	T4	181	40	55	21
	T5	247	44	70	18
T total		6,634	1,388	1,875	620
R9	F	249	43	58	18
	R9	17	8	9	3
R9 total		266	51	67	21
B	B	1123	301	299	111
Grand total		76,638	16,609	21,141	7,174

doi:10.1371/journal.pgen.0030104.t002

19 (0.00024%). The total number of samples from project inception in which the genotyping process could not be completed after attempting genotyping from both buccal swabs provided by the participant was 21 (0.00027%).

General Indices

Hg frequencies observed in the entire database, the reference database, and the consented dataset of 21,141 records are given in Table 2. In the entire database, the most frequent Hg was Hg H (38.2%). When the database was collapsed into macro Hgs L(xM,N) M, and N the following frequencies were observed, respectively, 4.54%, 3.42%, and 92.04%. Table S4 provides the observed transitions, tranversions, insertions, and deletions for the entire database and further delineates their frequencies within each Hg for the reference database. Note that inferences regarding the number of times each mutation occurred within each Hg are impossible to determine from this table. The total numbers of distinct transitions and tranversions observed were 343 and 199, respectively. The total numbers of distinct insertions and deletions observed were 35 and 15, respectively. Table S5 describes, for the entire database, the number of distinct heteroplasmies observed and further delineates within the reference database their distribution within each Hg. The total number of distinct heteroplasmies was 152. As it is difficult to establish the threshold of heteroplasmy detection by direct sequencing with current technologies, it is likely that the heteroplasmies found are an underestimate [16].

Homoplasy and Back Mutations in HVS-I Haplotypes

The results described in this section are from the reference database to provide maximum phylogenetic resolution. Homoplasy is the phenomenon in which the same mutation is found in two distinct phylogenetic branches of the mtDNA tree. Back mutation is defined herein as the phenomenon by which a position considered characteristic or diagnostic to a certain Hg has reverted to the ancestral state. It is clear that the phenomenon can affect any other position as well. The result of both phenomena can be haplotypes that are identical by state but not by descent (Figure 1), and can therefore bias interpretation of databases that make use of HVS-I haplotypes alone to infer Hg labeling or shared ancestry. In addition, these phenomena can also lead to an

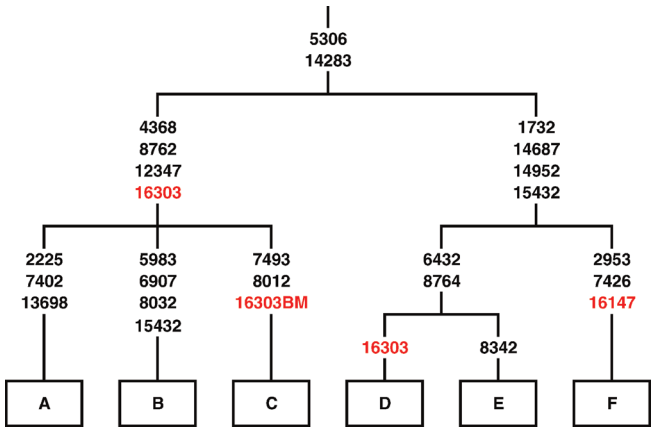

Figure 1. HVS-I Identity by Descent or by State

A theoretically evolving tree is presented. Coding-region polymorphisms are in black. HVS-I polymorphisms are in red. Samples A and B share HVS-I haplotype 16303 by descent. Samples A and D or B and D share HVS-I haplotype 16303 by state and as a result of homoplasy. Samples C and E are identical by state as a result of a back mutation in position 16303 in sample C as marked by the "BM" designation.

doi:10.1371/journal.pgen.0030104.g001

underestimation of population genetic distances. The extensive database presented herein contains numerous examples of these phenomena, of which many are well known while others are previously unreported. Table S6 shows the number of times that all classic HVS-I Hg-defining mutations are present as part of the haplotype motif in all reported Hgs. Table S7 shows the number of times that the same haplotype occurs in different Hgs for the portion of the reference database that overlaps with the consented database. Table S8 shows the number of times that a sample was assigned to an Hg by the SNP genotype, but did not harbor the classic HVS-I motif as defined in Table S2. Unfortunately, the scope of this paper is too limited to describe all examples and, therefore, we focus on a few examples that emphasize the magnitude of these phenomena.

The haplotype that shows no polymorphic changes when compared to the revised Cambridge Reference Sequence (rCRS) was well-reported under Hgs R*, U*, HV*, H, V, and their sub-branches [2,5,17,18]. These Hgs, which are frequent in populations of European ancestry, are expected to be frequent among the project's largely North American participants. Indeed, our database can be considered to contain an extensive sampling of European-derived populations. Since it is hard to decide whether the HVS-I haplotype of the ancestor of Hg R was rCRS or 16519C when mtDNA positions 16024–16569 are considered, the existence of the former in these Hgs can either represent identity by descent or identity by state due to homoplasy (Figure 1). Whether identical by descent or by state, the use of the 22-SNP panel allows the accurate placing of each mtDNA genome into a single Hg. Of a total of 463 mtDNA genome sequences that contain the rCRS (16024–16569) as their HVS-I haplotype, 416, 26, 1, and 20 would have been assigned to Hgs H, HV*, U*, and V, respectively, when typed with our 22-SNP panel (Table S6). Likewise, any study of a European population that used HVS-I information only and labeled all rCRS samples as identical or as Hg H, is likely to incorrectly assign about 10% of these samples. The large size of the database allows us to estimate the frequencies of additional examples of homoplasy that were previously reported in the literature. Positions 16343G, 16356C, and 16270T are considered characteristic of Hgs U3, U4, and U5, respectively. These positions actually occur in Hg H in 0.7%, 0.1%, and 3.9% of the cases, respectively. Positions 16224C and 16311C are widely considered to characterize Hg K. Our database, however, shows both a branch within Hg H that carries haplotype 16224C–16278T–16293G–16311C and branches within Hg K that lack positions 16224C or 16311C. The characteristic positions for Hg J and T are 16069T–16126C and 16126C–16294T, respectively. Several samples in our database shared the haplotype 16069T–16126C–16294T that contains both characteristic positions and proved to belong to Hg J. Haplotype 16223T–16519C occurred within Hgs H, M*, N*, U*, and W. More complex haplotypes, such as 16223T–16355T–16519C, occurred under both L3* and M*. Haplotype 16223T–16295T–16519C occurred in Hgs M* and W. The combination of positions 16189C–16217C occurred under both Hg B5 and N*. The important branching point between macroHg N and its daughter, macroHg R, is marked by two transitions, T12705C and T16223C. Our database shows that 2.5% of all preHg R mtDNA genomes have lost polymorphism 16223T and 1.1% of all R mtDNA genomes gained this mutation, mostly in the K1a1b1a lineage [19]. More specifically (Table S8), Hg I is characterized by HVS-I positions 16129A–16223T–16391A. Of the 421 Hg I mtDNA genomes defined by the relevant coding-region SNPs, 1.2%, 1.0%, and 3.3% have lost positions 16129A, 16223T, or 16391A, respectively. Of the 282 Hg W mtDNA genomes defined by the relevant coding-region SNPs, 1.4% and 15.2% have lost positions 16223T and 16292T, respectively. Of the 229 Hg C mtDNA genomes defined by the relevant coding-region SNPs, 2.2%,, 2.6% and 0.4% have lost positions 16223T, 16298C, and 16327T, respectively. These examples of positions that have experienced back mutation cannot indicate the number of times that each position has reverted during the Hg's evolution, as within-Hg resolution is not part of the presented database.

Homoplasy and Back Mutations in Coding-Region SNPs

The results presented below are from the reference database to provide the maximal phylogenetic resolution. Coding-region SNPs, used as reliable markers to define Hgs because they are considered stable evolutionary events, are nevertheless not entirely stable [19–21]. The dataset reported here supports this notion, and the portion of samples in which the SNP genotyping results were shown to be "inconsistent" with the expected phylogenetic hierarchy provides an important opportunity to estimate the extent of this phenomenon. Table S9 gives the number of times each of the tested SNPs occurs in different branches of the phylogeny. The overall frequency of samples in which inconsistency was observed was 3.2%. We note that excluding the 9-bp deletion at position 8280 would decrease the frequency to 2.0%. We highlight a few examples here.

The most trivial is the occurrence of transition A13263G (which we use to identify Hg C) in Hg W. The phylogeny supported by the remaining panel of 21 SNPs correctly places the samples as belonging to Hg W. The occurrence of this transition under Hg W actually defines the samples as belonging to sub-Hg W3 [22]. Hg H, descending from R0, is

expected to harbor transitions A11719G, T14766C, and T7028C. However, 83 of the total 6232 Hg H samples lack transition A11719G, of which 73 share the HVS-I position 16316G. The phylogeny supported by the remaining panel of 21 SNPs correctly places the samples as belonging to Hg H, with this subset probably representing a monophyletic clade characterized by the loss of transition A11719G and gain of position 16316G that has not yet been named. An interesting issue concerns transition T12705C, which is the only coding-region mutation known to separate Hgs R and N [21]. Three samples (0.1%) out of the 2,923 that were labeled by the remaining 21-SNP panel to be pre-R lineages did carry this transition, all of which were in Hg L sub-branches. Conversely, a total of 13,686 samples were labeled by the remaining panel of 21 SNPs to be lineages within Hg R, of which seven (0.05%) did not carry transition T12705C (but all carried SNPs typical of Hgs within R). These findings emphasize the importance of this position as separating Hg N from R.

The NN Methodology

To quantify the effectiveness of the NN/weighted NN (w-NN) method combined with our reference database in mtDNA classification, we tested our ability to recover the classification revealed by the coding-region SNPs in the Genographic database. We consider classification into 23 basal Hgs based on our most extensive SNP typing protocol (22 coding-region SNPs) as a "gold standard" classification (correct with a very high probability), and use it for comparison of the performance of our rule-based and w-NN classification approaches, when classifying based on HVS-I information only (without using the SNPs for classification). For this purpose, we adopted a leave-one-out cross-validation approach, i.e., each of the 16,609 samples for which we have 22 SNPs was left out, and the 16,608 remaining samples were used as a "reference" database for NN/w-NN. The accuracy obtained for recovering the coding-region Hg assignment by the NN/w-NN approaches was 96.72% and 96.73%, respectively (Table S10, last row). While this difference is tiny, we see consistently throughout Table S10 that w-NN does slightly better than NN (win-loss-tie ratio of 35-4-5). We also applied the rule-based approach (Table S2) based on HVS-I only, and obtained an accuracy of 85.3% (Table S10). Our conclusion from this experiment is that the NN-based approaches can support much higher accuracy in classification of our samples (and samples coming from similar populations) based on HVS-I only, when utilizing the Genographic database as reference. Table S10 details the results of repeating the same experiment with a variable number of SNP panels.

Saturation

We studied the level of haplotype saturation with respect to different HVS-I haplotypes and polymorphic sites present in the database by randomizing the order of the samples in the entire database and plotting the number of newly observed HVS-I haplotypes as a function of the accumulated number of samples (Figure 2). We repeated our analysis for the subsets of Hgs known to represent typically African, West Eurasian, East Asia-Americas, and South Asian mtDNA gene pools, and for Hg H haplotypes. Next, we repeated the analysis for the number of polymorphic sites obtained as a function of accumulated number of samples for the same categories.

The entire database of 76,638 samples was included in this analysis, within which 29,267 belonged to Hg H. A total of 11,346 HVS-I haplotypes were observed in this set (Table S3 shows the partial list of the observed haplotypes in the consented database). Note that homoplasy among these haplotypes is ignored, and the total number of phylogenetically independent haplotypes would have been higher if Hg information had been considered. Figure 2A shows the obtained results for all the haplotypes (11,346), for the groups of Hgs grossly affiliated with Africa (1,348), East Asia-Americas (1,663), South Asia (583), West Eurasia (7,684), and Hg H (2,637). Hgs in which geographic affiliation is uncertain (N*, R*) were excluded from the analysis. Figure 2B repeats the analysis for a limited number of samples to allow better comparison with the less-represented geographic groups. Figure 2C and 2D shows the application of the same analysis to the observed HVS-I polymorphic sites.

Searching for Evidence of Neanderthal mtDNA and Recombination

We have utilized our database to search for evidence of Neanderthal origin for any of the samples, and for any discrepancies that might be attributed to recombination.

On the Neanderthal question, we first extracted from GenBank all six Neanderthal HVS-I sequences of length at least 300 bp (Table S11). It is now accepted that a combination of five HVS-I mutations (16037G, 16139t, 16244A, 16262T, and 16263.1A), which appears in all of these samples, distinguishes these Neanderthal sequences from modern humans [23]. While all of these five mutations have in fact been observed in our full database of 78,590 samples, no combination of any two of them has appeared in any sample. However, since these six samples may not represent the full diversity of Neanderthal lineages, we have also investigated separately the level of divergence they show from our entire database. No sample in our database is as divergent as these Neanderthal samples, in terms of its distance from its nearest neighbor outside its own Hg, or its distance from the rCRS, which we take to represent a "random" modern human mtDNA (Table S11). We also observe that the most divergent samples in our database all carry well-known HVS-I motifs characteristic of African Hg L branches. While it is difficult to translate these findings into probabilities, it is clear that our results do not support the existence of mtDNA samples of Neanderthal (or other archaic *Homo*) origin in our database.

In the search for recombination, we concentrated on our reference database. If there was a detectable level of recombination in mtDNA, it should lead to phylogenetic inconsistencies in the 22-SNP genotypes. For example, if there was recombination between an Hg H mtDNA and an Hg M mtDNA, where the M sample "donated" its region between nucleotides 9000 and 12000 into the Hg H sample, then positions 10400 (Hg M), 10873 (Hg N), and 11719 (R0), which are in this region and hence in their non-rCRS state, should be "inconsistent" with positions 7028 (Hg H), 14766 (Hg HV), and 12705 (Hg R), which are in their rCRS state. Thus, we extracted all the samples in our reference database that were "inconsistent" (a total of 538 records). Of these, 521 can be explained by a single inconsistency, which can be attributed to a single repeated/back mutation rather than recombination. The remaining 17 require two repeated mutations to explain them. Nine of these 17 cannot be

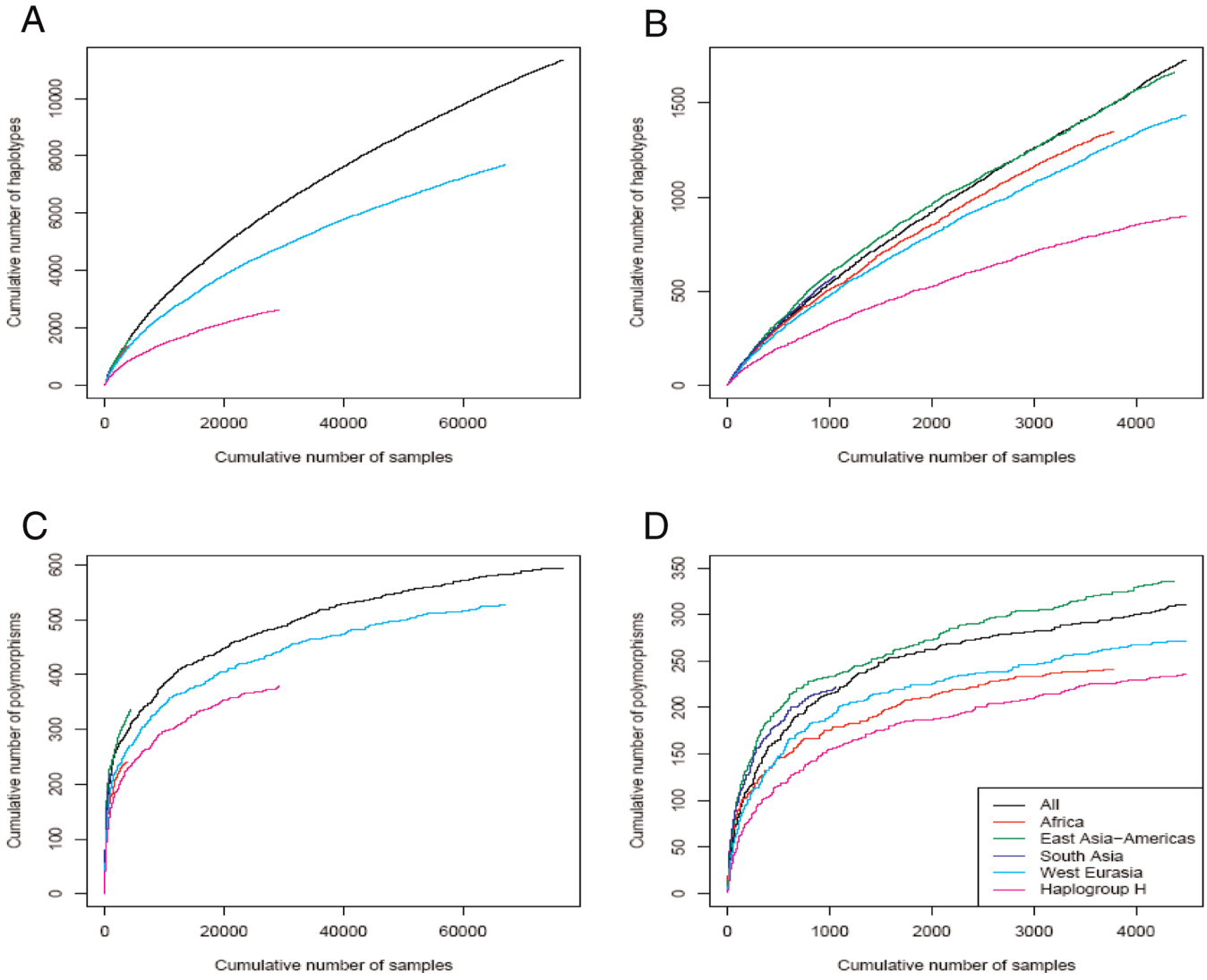

Figure 2. Saturation Curves

The number of accumulated mtDNA HVS-I haplotypes (A and B) and polymorphic sites (C and D) as a function of the number of accumulating samples is shown. The analysis is presented once for the entire database (A and C) and once for a limited number of samples (B and D), allowing a better comparison with the less well-represented geographic groups. The Hgs were grossly divided to represent four different geographic groups as follows. Africa: L, M1, and U6; East Asia-Americas: A, B, C, D, F, N9a, and R9; South Asia: M*, R1, R2, R5, and R6; and West Eurasia: N1, R, W, and X. Saturation curves for Hg H are also presented.

doi:10.1371/journal.pgen.0030104.g002

explained by a single recombination event. The remaining eight fall under the Hg H branch, described above, which is marked by the back mutation at position 11719 and by the HVS-I transition A16316G. The second inconsistency in seven of these samples involves the 9-bp deletion at position 8280 and the eighth sample involves an inconsistency in position 13368 (Hg T). As all eight occurred under a phylogenetically consistent branch, we attribute them to repeated mutation rather than a recombination event. We thus conclude that we can find no evidence of recombination in our reference database.

Discussion

The Genographic Project allows members of the public to participate in a real-time anthropological genetic study. Since its inception in early 2005, over 188,000 individuals have joined the project, of which over 55,000 have submitted their mtDNA or MSY results to the research effort, illustrating the high level of interest. Because any member of the public may participate in the study, rigorously controlled group affiliation data will not typically be obtained for the samples. Furthermore, it is clear that the accumulated database is biased towards countries in which the project is well known and where the kits are economically accessible to a significant fraction of the population. The fact that 95% of the participation kits were ordered in the US and Western Europe is consistent with the Hg frequencies observed, and suggests that the majority of the participants are of West Eurasian (probably European) ancestry.

The importance of improving the quality of the global shared mtDNA database was recently reemphasized and

summarized by Bandelt et al [24]. The strict uniform adherence to standard analytic and genotyping protocols across tens of thousands of samples makes the current study an ideal resource for the scientific community. We tried to consider all previously identified sources of errors while designing our genotyping, analysis, and reporting tools. Our database is unique for a project of this scale in using sequencing of both strands of the HVS-I as a standard procedure to assure high-quality data. The same goals led us to incorporate standard coding-region SNP genotyping on all samples. The entire analysis is "pen-less" to avoid any typographic mistakes, and a series of computational quality control measures are embedded in it. Despite the rigorous quality check procedures implemented, we still anticipate some inaccuracies in the database, but believe that these genotyping standards raise the bar on mtDNA genotyping and represent good progress towards more reliable databases.

A few simple measures can be suggested to facilitate future assembly of mtDNA databases. First, as sequencing procedures have become more efficient and stretches of 600 bp can easily be obtained, we suggest standardizing the reported "HVS-I" range to include positions 16024–16569 as presented herein. Second, it would be worthwhile to create a standard list of coding-region SNPs used by the scientific community for Hg assignment and change to alternative coding-region SNPs defining the same Hg when there is a reason to suspect that the standard SNP is misleading due to homoplasy or back mutation. We make available our quality check measures as a model for any future mtDNA database submitted for publication.

The database reported herein is very informative with respect to the mtDNA phylogeny, including the frequencies of the observed haplotypes, transversions, transitions, indels, and heteroplasmic positions both in the coding and control regions (Tables S3–S10). No highly divergent (e.g., Neanderthal) sequences were observed, despite more than doubling the total number of sequences examined, and no evidence for recombination was found. The database did, however, provide evidence for homoplasy and back mutations affecting a low, but not insignificant, percentage of the samples both at the HVS-I and the coding-region SNPs chosen herein. For the coding-region SNPs, even these phenomena do not usually prevent the correct positioning of an mtDNA genome in the phylogeny, as the latter is based on the identification of a string of positions and not a single one. For the HVS-I, our analysis shows that while the use of Hg labeling techniques based on HVS-I variation have an overall good correlation with coding-region SNP genotyping, caution should be used in general, and, in certain specific cases, prediction is best avoided. In population-based studies of large sample size, these phenomena will likely have a small affect on the overall conclusions. However, for individual genotyping, as studied in genealogical or forensic cases, these percentages may be sufficient to preclude, for example, a firm conclusion regarding the time to most common recent ancestor of a set of samples for which only HVS-I information is available.

The NN methodology presented herein, when jointly used with our reference database, has been shown to assign more mtDNA genomes to their correct Hg than prediction methods based on the classic set of HVS-I motifs. Our genotyping strategy, associating each of the HVS-I unique mutations with an Hg confirmed by a coding-region SNP, supplies the needed infrastructure for developing the NN methodology. It is clear that the high prediction score of NN/w-NN is a function of the size of the reference database collected within the population, in which the NN/w-NN methodology is implemented along with the length of the analyzed fragment in the HVS-I. For this study, and considering the large reference database, it was shown that, when no coding-region genotyping was done and Hg prediction was based solely on HVS-I classic Hg-determining rules, as many as 15% of the predictions were wrong, while the w-NN yielded an accuracy of 96.73%. In the sample set studied, the high rates of failure in predicting the correct Hg using HVS-I based rules alone is likely the result of high prevalence of Hgs for which no satisfactory predictive rules exist (such as Hg H and HV*) and to a lesser extent from phenomena like homoplasy or back mutations. To illustrate how the use of the w-NN methodology requires a joint use of a reliable relevant reference database for the studied population, we applied the w-NN methodology and our current reference database to published databases that are external to the Genographic Project and from various populations. West European and non-West Eurasian sequences, the two extremes, yielded prediction scores at a high and a low of 93.8% and 77.9%, respectively (data not shown). Therefore, we make the NN prediction methodologies available on our Web site (http://www.nationalgeographic.com/genographic) in two forms: a) the NN independent code to be used with any reference database and b) in combination with an upload tool that allows the NN methods to be applied to uploaded samples using the Genographic reference database. As emphasized, we expect that the best prediction scores will currently be obtained in samples of West Eurasian ancestry for the 23 basal Hgs defined here, and that the predictions will gradually improve for other populations as the Genographic Project progresses and worldwide samples are obtained and included in the reference database, and as more coding-region SNPs are used to further resolve the basal Hgs into their sub-clades, a process actively underway in the Genographic research consortium.

An interesting question that can be examined using our database relates to the effect of protocols using variable numbers of coding-region SNPs on the accuracy of Hg assignment when compared with the classification of the reference database using the full 22-SNP protocol as a gold standard (as if 100% accurate). Table S10 gives the results for several coding-region SNP protocols of which the 10-, 20-, and 21-SNP protocols were previously used by the Genographic Project. These data show that a high degree of predictive accuracy was rapidly achieved as SNPs were added. When no SNPs were used, the best prediction methodology was with w-NN and yielded an accuracy of 96.73%. The most important single SNP in our population, 7028 (Hg H), allows 98.18% accuracy (w-NN) on its own. The initial panel of ten SNPs, when combined with the HVS-I information, is responsible for 99.81% (w-NN) of the Hg assignment accuracy achieved, and the last 12 SNPs are needed to resolve the remaining small portion of the samples (Table S10).

Our large database allows us to make some simple measurements of haplotype and polymorphic site saturation. Figure 2 shows that even the large number of samples collected in our study does not reach HVS-I haplotype saturation. The discrepancy between the shapes of the

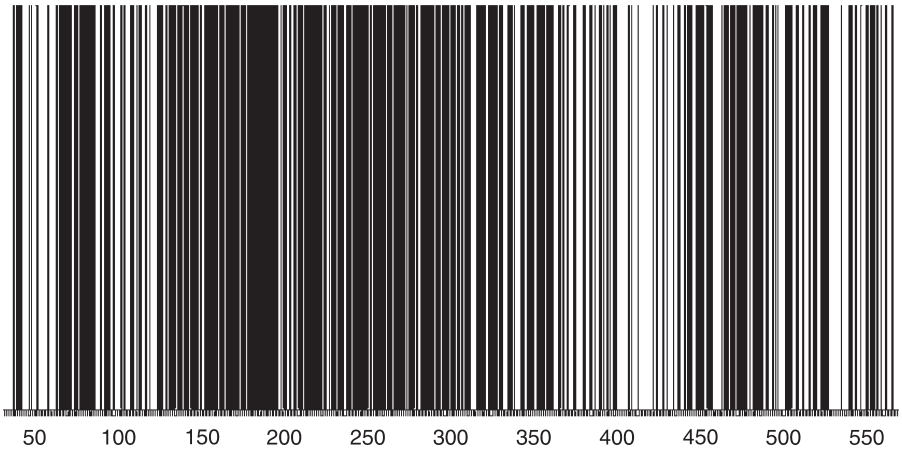

Figure 3. Physical Map of HVS-I

The figure presents a simple map made up from all polymorphic sites observed in the sequenced region 16024–16569 without denoting their frequencies. Conclusions regarding the number of times each observed position was hit during *Homo sapiens'* evolution can not be inferred.
doi:10.1371/journal.pgen.0030104.g003

haplotype and polymorphic site curves probably means that the number of observed polymorphic sites is closer to saturation than the number of observed haplotypes, which in turn suggests that shuffling of the same polymorphic sites, through homoplasy and back mutations, is the dominant mechanism that increases haplotype variation. These results are not surprising in view of the strong signal of expansion observed in human mtDNA [1]. Indeed, given the huge state space for haplotype motifs, we would expect a large number of haplotypes at very low frequencies, keeping the saturation curve of haplotypes steadily rising. In contrast, the space of sites is tiny, and, therefore, presumably closer to saturation.

The function of the control region is not completely understood, but is thought to be involved in mtDNA genome replication and transcription, and possibly contains the origin of heavy- and light-strand mtDNA replication and several transcription binding sites, with the HVS-I depauperate in regions of this kind [25]. One might expect that the parts of the control region in which these sites are found will be more conserved than others. The information obtained from all unique polymorphic transversions, transitions and indels was used to draw a "bar code" of the sequenced region to show all positions in which a mutation was observed (Figure 3). A total of 358 (65.5%) of the possible 546 sequenced positions showed polymorphism. Some variability in density of polymorphic regions is evident, but no "polymorphism-free" regions can be detected. Note that the map does not distinguish between positions that mutated once or multiple times during the mtDNA evolution of *Homo sapiens*. It is also important to note that the database does not represent the worldwide variety of mtDNA and, therefore, mutations typical of other populations may not be represented.

A few considerations unique to a public project should be discussed. Because the current dataset presented in this manuscript comprises members of the public who have joined Genographic's research effort, the samples herein represent a subset of the total global mtDNA diversity. To properly survey the genetic variation in non-Western Eurasian lineages, the Genographic Consortium is actively consulting and engaging with members of indigenous communities from around the world, and conducting anthropological and genetic analysis on those DNA samples. As such data are published they will also be made available anonymously as part of the reported Genographic reference database. The classification, saturation, and analytical techniques will need to be updated accordingly, as is the case with any expanding database. In addition, this manuscript presents a level of Hg resolution based on the current 22-SNP panel and HVS-I information. A stated goal of this research effort is to continue to refine and increase this resolution, which will be achieved by further genotyping or revised analysis incorporating the expanding dataset. Therefore, at present, the participants and scientific community are presented with a solid, but still rather simple, level of analytic resolution and are encouraged to return periodically to the project's Web site to access up-to-date data and analytical tools.

In summary, we report both data and new classification methods developed using by far the largest standardized mtDNA database yet created, and detail the logistic, scientific, and public considerations unique to the Genographic Project. Most importantly, we return to the public a database made possible by their enthusiastic participation in the Genographic Project.

Materials and Methods

Sampling and sample handling. The Genographic Project's Web site allows members of the public to order a buccal swab kit (containing two buccal swabs) and undergo genotyping for either mtDNA or MSY analysis. To ensure anonymity, each participation kit is encoded with a randomly generated, nonsequential, Genographic Participant ID number. All samples are genotyped with informed consent according to procedures approved by the Institutional Review Boards of the University of Pennsylvania and the United States Department of Health and Human Services. Once results are obtained, the participants may consent to contribute their genetic data anonymously to the Genographic research database, to be used for anthropological studies and made available to the scientific community. The participants are also asked to provide genealogical information relevant to their deep ancestry.

Nomenclature. We use the term haplotype to describe HVS-I

variation. The reported HVS-I is "extended" and covers 16024–16569 for all samples. Absolute numbers are used to describe nucleotide position (1–16569) in the mitochondrial genome, and refer to the position of the polymorphism compared with the rCRS [26]. It is common practice to label by letters the nucleotide change only for transversions (e.g., 16318t) and to avoid labeling by letter transitions (e.g., 16093), since the changed nucleotide can be inferred from the rCRS [27]. As this study also addresses the general public, who may not be familiar with rCRS nomenclature conventions, we note here that we deviate from the common practice, and to facilitate reading and the use of the released database we label both. Transitions are labeled by capital letters (16093C), transversions by small letters (16318t), and heteroplasmies by the letter "N" (16189N). Sequencing alignment always prefers 3′ gap placement for indels. Deletions are marked by the letter "D" (16166D) and insertions by the point (.) sign (16188.1C).

We use the term Hg to describe haplotype groups ("haplogroups") [28] that usually coalesce tens of thousands of years ago and are best defined by a combination of coding-region SNPs. The Hgs currently reported by the Genographic Project are listed in Table 2. We adopt a standard Hg nomenclature scheme [27]. Since we have noted that the asterisk (*) suffix used in this scheme leads to some confusion among public participants, we elaborate here on this point by giving an example. A label such as M* means that a sample belongs to Hg M, but not to any of the known subclades within M. It is temporary, and should mean that the Hg is one of many paraphyletic clades falling under the monophyletic Hg M but is not any of the *known* single-letter (e.g., Hg D) or letter-number (e.g., Hg M1) coded Hg M sub-branches. It is therefore clear that even if all reported databases abided by this definition and labeled M* by excluding all known sub-branches at time of publication, it would be impossible to compare samples that fell into this cluster in different publications, because new sub-branches are continually defined. The solution suggested for Y chromosomal nomenclature [29], which clearly specifies which sub-branches were excluded (e.g., M*(xCZ, M1, M3, M51)), might ease database comparisons, especially, when phylogenetic knowledge enlarges and it becomes harder to exclude all known sub-branches of each given Hg in each study. Therefore, we suggest a slight modification to the use of the asterisk suffix. Herein, its use denotes that the sample was excluded from all sub-Hgs reported in this study only (Table 2), whether defined by a coding-region SNP or an HVS-I defining motif (Table S2). Therefore, in this study, the label M* means that the sample belongs to Hg M and was excluded only from the sub M branches reported in this study; namely, C and D by coding-region SNPs, and M1 and Z by HVS-I defining motifs. The sample could still belong, for example, to the well-defined M5 or M8 branches that are not part of the Hgs reported in this study.

HVS-I sequencing. Sequences of an extended HVS-I (16024–16569) are determined from positions 16024 to 16569, by use of the ABI Prism Dye Terminator cycle-sequencing protocols developed by Applied Biosystems (http://www.appliedbiosystems.com). Sequencing is performed on a 3730xl DNA Analyzer (Applied Biosystems). Mutations are scored relative to the rCRS [26]. The primary amplification is achieved by primers 15876F and 639R (Table S1). PCR products are cleaned using magnetic-particle technology (BioSprint 96; Qiagen, http://www.qiagen.com). Following the primary amplifications, all samples are subject to bidirectional sequencing using primers 15946F and 132R (Table S1). In cases of template polymorphism at the annealing site(s) and failed sequencing due to primer/template mismatch, alternative primers are used (Table S1). High quality is assured by the following procedures: (1) All sequences are aligned by the software Sequencher (Gene Codes Corporation) and observed by an operator. (2) All positions with Phred score <30 are directly inspected by an operator [30,31]. (3) All positions that differ from the rCRS are recorded electronically. (4) Forward and backward sequences of all samples are electronically checked for consistency. (5) All scenarios noted herein are highlighted for review: failed samples, inconsistencies in forward and backward sequencing, successful sequencing in one direction only, sequences that contain indels or heteroplasmy, and sequences that are shorter than the required length. (6) All highlighted samples are observed again by a second operator. (7) All sequences containing two or more heteroplasmies are regarded as contaminated and DNA is re-extracted from the second swab of the participant. (8) The list of HVS-I haplotypes observed among the lab staff is presented as part of Dataset S1. (9) All reported variant positions are digitally checked for consistency of the expected order of the mutations (i.e., 16093C followed by 16126C and not 16126C followed by 16093C). (10) All reported variants are verified to represent a real polymorphism by direct comparison to the rCRS. (11) All variants reported for the first time when compared to the entire database are highlighted and re-observed. (12) All data donated to the scientific world with consent are released. Any comments and remarks raised by external investigators after release will be addressed by re-observing the original sequences for accuracy. Following that, any unresolved result will be further examined by re-genotyping and, if necessary, immediately corrected by publishing an erratum.

Coding-region biallelic site genotyping. The biallelic sites are genotyped by means of KASPar assays [32] and are independent of the sequencing, thus playing an additional role in the quality check. Twenty one SNPs and the 9-bp deletion make up the total of 22 biallelic sites. For simplicity, we will refer to all biallelic sites as SNPs. The number of SNPs tested was gradually increased from ten at inception of the project to the 22 currently used. The ten initial SNPs were 3594, 4580, 5178, 7028, 10400, 10873, 11467, 11719, 12705, and 14766 (numbers refer to the nucleotide position in the mitochondrial genome). The panel was augmented to a total of 20 coding-region SNPs by including the following additional ten SNPs: 4248, 6371, 8994, 10034, 10238, 10550, 12612, 13263, 13368, and 13928. The panel was further augmented by the addition of SNP 2758, to a total of 21 coding-region SNPs and finally by including the 9-bp deletion at position 8280 to a total of 22 coding-region SNPs (Figure 4). Two further changes were made: positions 8994 and 13928 used in some early work were respectively replaced with their phylogenetic equivalents 1243 and 3970. Therefore, the current panel includes the following SNPs, with their respective gene locations shown in brackets [33]: 2758 (16S), 3594 (ND1), 4248 (M), 4580 (ND2), 5178 (ND2), 6371 (COI), 7028 (COI), 8280 (9-bp deletion) (NC7), 8994 (ATPase6), 10034 (G), 10238 (ND3), 10400 (R), 10550 (NDRL), 10873 (ND4), 11467 (ND4), 11719 (ND4), 12612 (ND5), 12705 (ND5), 13263 (ND5), 13368 (ND5), 13928 (ND5), and 14766 (Cytb). The coding-region SNPs were chosen based on the following considerations: (1) Major branching points in the mtDNA phylogeny obtained using complete mtDNA sequences [21]. (2) Hgs known to be frequent among the current populations in which the project is advertised [2,34,35]. For example, the R0 clade within macroHgs R and N is over-represented. (3) Hgs in which the HVS-I predictive value is known to be unsatisfactory [18]. (4) SNPs reported in previous publications that have been commonly used to identify a particular Hg [2]. For example, we choose polymorphism 7028 and not 2706 to identify Hg H. (5) Technical issues concerning the ability to validate any given assay.

The SNP genotyping results are obtained digitally and analyzed automatically to suggest the appropriate Hg consistent with the mtDNA phylogenetic tree. Two possible scenarios can prevent the reliable assignment of an Hg by SNPs. First, when SNP genotyping in critical positions for labeling a particular Hg has failed due to technical problems, the genotyping result is rendered "uninformative." Note that most of the information might still exist with only the terminal SNP in the mtDNA phylogeny missing. Second, when SNP genotyping is complete but the reported mutations deviate in a particular SNP from the accepted mtDNA phylogeny, the genotyping result is labeled as "inconsistent" and can result from homoplasy, back mutation, a new unknown SNP next to the checked SNP that distorts the reaction, or a genotyping error.

Hg assignment. Hg labeling is achieved by combining the information obtained from (1) the coding-region SNPs, and (2) the HVS-I motifs. A third means of Hg labeling, based on NN methodology, is developed herein.

Hg assignment by coding-region SNPs. The standard panel of 22 coding-region SNPs allows a reliable, deep-rooted analysis of the mtDNA phylogeny for each sample as presented in Figure 4. The SNP panel contains a diagnostic SNP for each of the following major bifurcations in the mtDNA phylogeny: L2′3 ′4 ′5 ′6 ′7, L3′4 ′7, M, D, C, N, N1, I, A, W, X, R, R9, B, J, T, U, K, R0, HV, V, and H [1,21]. Therefore, a total of 23 Hg clusters can be inferred from the SNP resolution: L0 or L1, L2 or L5 or L6, L4 or L7 or L3(xM,N), M(xC,D), C, D, N(xN1,A,W,X,R), N1(xI), I, A, W, X, R(xU,R0,J,T,R9,B), U(xK), K, R0(xHV), HV(xH,V), H, V, J, T, R9, and B (Figure 4). To facilitate reading, these 23 Hg clusters are labeled more simply as follows: L0/L1, L2, L3*, M*, C, D, N*, N1*, I, A, W, X, R*, U*, K, R0*, HV*, H, V, J, T, R9, and B. We emphasize that these labels are not equivalent to the final Hg definitions. It is important to note that the use of the coding-region SNPs is very accurate but still prone to errors. For example, under a theoretical scenario in which a sample that belongs to Hg H has a back mutation in position 7028, the panel will label it as HV. We have no way of estimating the frequency with which such a scenario might occur, as we test only one coding-region SNP per branch, but we expect that this phenomenon is very rare.

Final Hg labeling. Final resolution to Hgs and sub-Hgs is achieved by

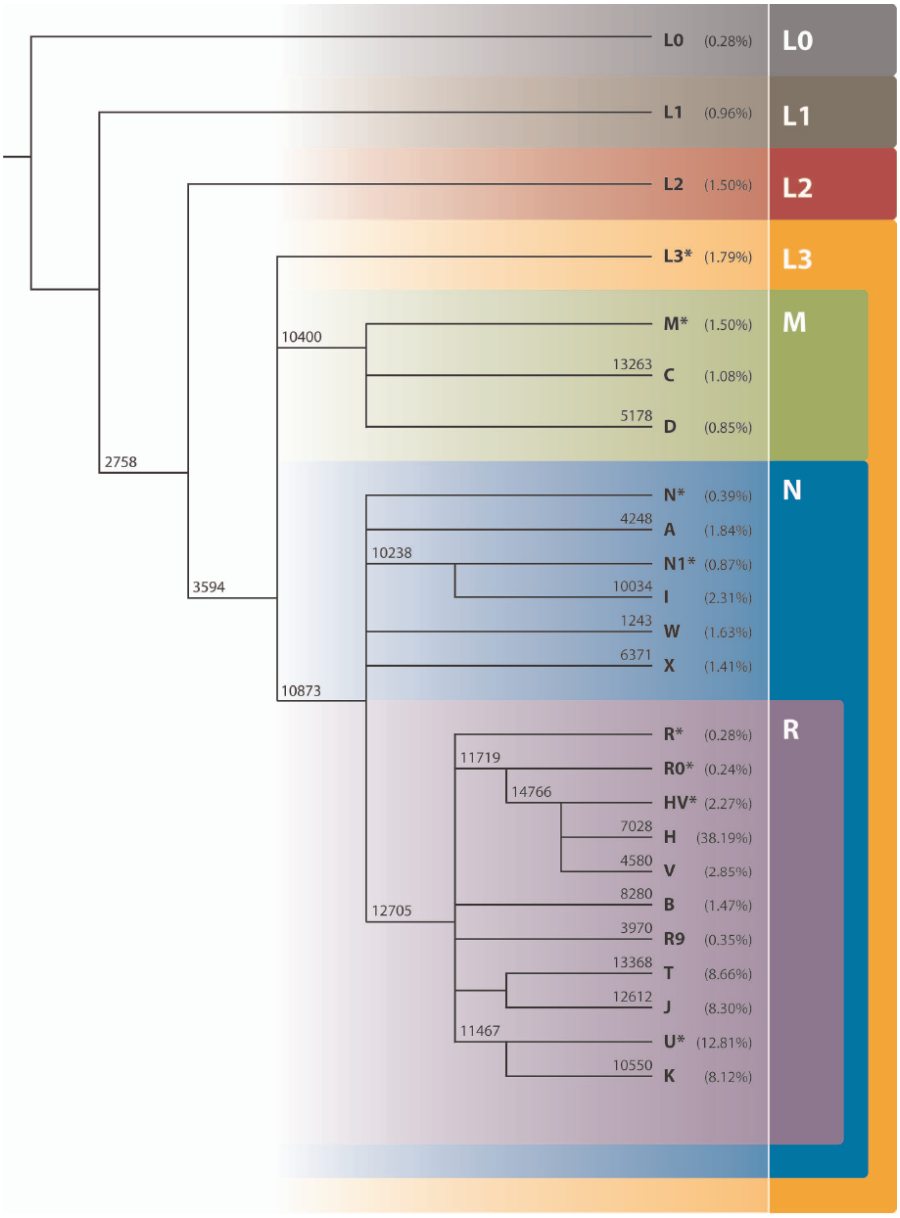

Figure 4. The Phylogeny of mtDNA Haplogroups Inferred from the Panel of 22 Coding-Region SNPs Used in the Genographic Project

The coding-region mutations are shown on the branches. The frequencies of the haplogroups found among the Genographic participants are shown in brackets beside the Hgs assignments and correspond to Table 2. Note that the figure discriminates between haplogroups L0 and L1 while the coding-region SNPs used during genotyping do not distinguish the two and therefore they are labeled throughout the paper as L0/L1.

doi:10.1371/journal.pgen.0030104.g004

comparing and combining the information obtained from the SNP genotyping with the HVS-I motifs. All HVS-I haplotypes obtained following sequencing are digitally screened for possible Hg and sub-Hg definitions by use of accepted HVS-I diagnostic motifs (Table S2) [2,3,36]. The list presents only the motifs used herein for prediction purposes and should not be treated as comprehensive for all Hg suggestions that might rise from HVS-I variation or as representing the Hg basal HVS-I motifs. First, a screen in the *a priori* defined order presented in Table S2 is run and stopped at the first Hg where the sample matches the motif. A second screen for all possible Hgs the sample can fit in is then conducted. It is clear that relying on the HVS-I variation alone to infer Hgs and sub-Hgs such as M1, Z, U5, U6, and HV1 is prone to inaccuracies. In addition, HVS-I haplotypes alone cannot identify Hgs or sub-Hgs that have no defining motifs and ignores the possibility of homoplasy and back mutations. For example, it is clear that some of the mtDNA genomes appearing in our database as U* might actually belong to Hg U4 but, as they did not contain the diagnostic HVS-I position 16356C and in the absence of additional coding-region genotyping, we could not label them as such. Moreover, some of the sub-Hg definitions inferred from the HVS-I, for example within Hgs J and T, will have to be revised in the future as studies using complete mtDNA sequences prove they do not represent monophyletic clades [20,37]. Therefore, whenever analysis is done within an Hg, we refer only to one of the 23 Hgs directly inferred from the 22 SNPs genotyping to avoid any HVS-I based Hg labeling misinterpretations.

Quality checking of Hg labeling is as follows: (1) All discrepancies between HVS-I and SNP labeling are observed by an operator. These discrepancies usually derive from well-known cases of homoplasy and are easily resolved by adhering to the SNP genotyping that correctly assigns the sample to a single Hg in the mtDNA phylogeny. In cases of inability to resolve the discrepancies, the genotyping process is

repeated. (2) All samples in which the SNP information is uninformative are observed by an operator. An attempt to label the final Hg is made by direct observation from the partial list of SNPs available and the HVS-I motif. In case of any persisting doubt, the sample is re-genotyped. (3) All samples in which the SNP information is inconsistent are observed by an operator. The Hg assignment is accomplished after studying the entire string of available mutations and by applying the principle of parsimony. The final Hg can be further supported by the HVS-I information.

Statistical analysis. *General indices.* Success rates of each of the genotyping processes, Hg frequencies, and distributions including the frequencies of transversions, transitions, heteroplasmies, indels, back mutations, and homoplasy occurring in the HVS-I after taking into account the checked coding-region SNPs are determined by direct counting. We report the heteroplasmic positions in Table S5 but excluded them from all other analyses.

NN classification methodology. The common practice of classifying samples into Hgs based on HVS-I information relies on a set of rules that define certain HVS-I backbone haplotypes as characteristic of specific Hgs by using the state of the art knowledge in the literature [2,3,36]. These characteristic motifs, implemented by us here as one of the Hg labeling techniques, are best if previously proven to be associated with particular coding-region SNPs identifying the suggested Hg, and then used to classify newly obtained HVS-I data into Hgs. The weakness of this approach is its sensitivity to phenomena such as homoplasy or back mutations in the motif's HVS-I positions, which may occur between Hgs or within sub-branches of the same Hg. Since parallel evolution is rampant in HVS-I, this issue casts doubt on the ability of rule-based classification to reach high levels of accuracy in certain cases [1,17–19].

Given a large enough "reference" data base of correctly labeled samples (for example, if all samples are verified by coding-region SNPs), we are likely to better assign Hgs for HVS-I haplotypes of new samples if we compare them to all available records in the reference database by identifying their "nearest neighbor," i.e., the most similar sample we have already classified with confidence. This allows us to use *all of the HVS-I information* in each classification decision, rather than simply counting on the rule-defining sites. Thus, *any* mutations within Hgs that have appeared in the samples in the reference database will be useful for classification, and recent homoplasy in a single HVS-I locus will have a more minor effect on our classification, because other loci within the HVS-I will still support the correct classification.

Given a backbone database D comprising correctly classified HVS-I samples $s_1,\ldots,s_n$ and a new HVS-I sample t, we define the pair-wise distance as $d(s_i, t) = \Sigma_{j\in J}\, w_j\, I\{t_j \neq s_{ij}\}$, where J is the set of HVS-I loci (defined as 16024–16569 in our case), and w_j is a locus-dependent weight. In a simple application (unweighted NN) we would simply take $w_j = 1\ \forall j$ and get the (unweighted) Hamming distance, often used in neighbor-joining algorithms. A more reasonable approach would be to down-weigh the loci with a higher mutation rate (such as 16311). Denote these mutation rates (in units of "mutations per year") as $p_1,\ldots p_J$. Then a weighting of $w_j = \log(20{,}000 \times p_j)$ can lead to an interpretation of NN Hg classification based on $d(s_i, t)$ as an approximate maximum likelihood estimate of the Hg, using the following logic: Assume that the "average" sample has a NN with coalescent time of about 10,000 years. Then the number of mutations separating the sample from its NN in site i has a Poisson $(20{,}000 \times p_i)$ distribution, under sufficiently simple substitution models. If we assume that $20{,}000 \times p_i$ is still very small, as would be the case for practically all sites, then we can approximate the Poisson by a Bernoulli $(20\,000 \times p_i)$ (which is 1 if the samples differ in site i). Now, if we treat the identity of the NN as the parameter to be estimated, we can see that a maximum likelihood estimate would lead us to choose the one minimizing $d(s_i, t) = \Sigma_{j\in J}\, w_j\, I\{t_j \neq s_{ij}\}$.

The w-NN analysis requires calculation of site-specific mutation rates, like the ones recently proposed by Bandelt et al. [38] We were limited in our ability to use these published rates, as they only apply to the region 16051–16365, rather than our HVS-I definition. Thus, in our experiments below we use a set of probabilities we derived using a novel methodology (Rosset et al., in preparation). We verified that these estimates are consistent with Bandelt et al. [38] for the region in common, and use them here since they are the only complete set we could obtain. An improved set of probability estimates may improve the results further.

?In applying the NN methodology, we are bound to encounter many "ties," when there are two equally close NNs in two different Hgs. In our implementation, we assign the new sample to the Hg in which the most similar haplotypes are most prevalent in the reference database.

Supporting Information

Dataset S1. The Genographic Project Open Resource Mitochondrial DNA Database (Consented Database)

Found at doi:10.1371/journal.pgen.0030104.sd001 (8.1 MB XLS).

Table S1. Amplification and Sequencing Primers

Found at doi:10.1371/journal.pgen.0030104.st001 (4 KB XLS).

Table S2. Hg-Predicting Motifs

Found at doi:10.1371/journal.pgen.0030104.st002 (9 KB XLS).

Table S3. Haplotypes Observed in the Database (Consented Database)

Found at doi:10.1371/journal.pgen.0030104.st003 (457 KB XLS).

Table S4. Polymorphic Sites Observed in the Database (Entire and Reference Database)

Found at doi:10.1371/journal.pgen.0030104.st004 (189 KB XLS).

Table S5. Heteroplasmic Sites Observed in the Database (Entire and Reference Database)

Found at doi:10.1371/journal.pgen.0030104.st005 (48 KB XLS).

Table S6. Classic HVS-I Motif Homoplasy (Reference Database)

Found at doi:10.1371/journal.pgen.0030104.st006 (17 KB XLS).

Table S7. Haplotypes Homoplasy (Consented Database, 22-SNP Group)

Found at doi:10.1371/journal.pgen.0030104.st007 (745 KB XLS).

Table S8. Classic HVS-I Haplotypes Back Mutation Events (Reference Database)

Found at doi:10.1371/journal.pgen.0030104.st008 (16 KB XLS).

Table S9. Coding-region SNPs Homoplasy (Reference Database)

Found at doi:10.1371/journal.pgen.0030104.st009 (8 KB XLS).

Table S10. Hg Prediction by Classic HVS-I Rules and NN Using Variable Sets of Coding-Region SNPs (Entire Database)

Found at doi:10.1371/journal.pgen.0030104.st010 (10 KB XLS).

Table S11. Divergent Measures of Neanderthals (Entire Database)

Found at doi:10.1371/journal.pgen.0030104.st011 (4 KB XLS).

Video S1. Introductory Video of the Genographic Project

Found at doi:10.1371/journal.pgen.0030104.sv001 (27 MB MOV).

Acknowledgments

We gratefully acknowledge the National Geographic Society, IBM, Family Tree DNA, and Arizona Research Labs for their support of the project. CTS is supported by The Wellcome Trust. We would like to thank all participants, whose collaboration has made this study possible.

The Genographic Consortium

Theodore G. Schurr, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America; Fabricio R. Santos, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; Lluis Quintana-Murci, Institut Pasteur, Institut Pasteur, Paris, France; Jaume Bertranpetit, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain; David Comas, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain; Chris Tyler-Smith, The Wellcome Trust Sanger Institute, Hinxton, United Kingdom; Elena Balanovska, Russian Academy of Medical Sciences, Moscow, Russia; Oleg Balanovsky, Russian Academy of Medical Sciences, Moscow, Russia; Doron M. Behar, Genomics Research Center, Family Tree DNA, Houston, Texas, United States of America and Rambam Health Care Campus, Haifa, Israel; R. John Mitchell, La Trobe University, Melbourne, Victoria, Australia; Li Jin, Fudan University, Shanghai, China; Himla Soodyall, National Health Laboratory Service, Johannesburg, South Africa; Ramasamy Pitchappan, Madurai Kamaraj University, Madurai, Tamil Nadu, India; Alan Cooper, University of Adelaide, South Australia, Australia; Ajay K. Royyuru, IBM, Yorktown Heights, New York, United States of America; Saharon Rosset, IBM T. J. Watson Research Center, New York, United States of America; Jason Blue-Smith, National Geographic Society, Washington, District of Columbia, United States of America; and R. Spencer Wells, National

Geographic Society, Washington, District of Columbia, United States of America.

Author contributions. RSW conceived and designed the Genographic Project. DMB performed the experiments and wrote the paper. SR designed and applied the modeling methodology and statistical analysis. All of the authors and the other members of the Genographic Consortium contributed to the analysis of the data, the writing of the manuscript, and approved the final version.

Funding. Field research is funded by a grant from the Waitt Family Foundation.

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

References

1. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ (2006) Harvesting the fruit of the human mtDNA tree. Trends Genet 22: 339–345.
2. 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.
3. Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, et al. (2004) Where west meets east: The complex mtDNA landscape of the southwest and Central Asian corridor. Am J Hum Genet 74: 827–845.
4. Thomas MG, Weale ME, Jones AL, Richards M, Smith A, et al. (2002) Founding mothers of Jewish communities: Geographically separated Jewish groups were independently founded by very few female ancestors. Am J Hum Genet 70: 1411–1420.
5. Pereira L, Richards M, Goios A, Alonso A, Albarran C, et al. (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. Genome Res 15: 19–24.
6. Behar DM, Hammer MF, Garrigan D, Villems R, Bonne-Tamir B, et al. (2004) MtDNA evidence for a genetic bottleneck in the early history of the Ashkenazi Jewish population. Eur J Hum Genet 12: 355–364.
7. Brandstatter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, et al. (2004) Mitochondrial DNA control region sequences from Nairobi (Kenya): Inferring phylogenetic parameters for the establishment of a forensic database. Int J Legal Med 118: 294–306.
8. Comas D, Plaza S, Wells RS, Yuldaseva N, Lao O, et al. (2004) Admixture, migrations, and dispersals in Central Asia: Evidence from maternal DNA lineages. Eur J Hum Genet 12: 495–504.
9. Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, et al. (2001) Origins and divergence of the Roma (gypsies). Am J Hum Genet 69: 1314–1331.
10. Bandelt HJ, Lahermo P, Richards M, Macaulay V (2001) Detecting errors in mtDNA data by phylogenetic analysis. Int J Legal Med 115: 64–69.
11. Bandelt HJ, Kivisild T (2006) Quality assessment of DNA sequence data: Autopsy of a mis-sequenced mtDNA population sample. Ann Hum Genet 70: 314–326.
12. Bandelt HJ, Quintana-Murci L, Salas A, Macaulay V (2002) The fingerprint of phantom mutations in mitochondrial DNA data. Am J Hum Genet 71: 1150–1160.
13. Bandelt HJ, Salas A, Lutz-Bonengel S (2004) Artificial recombination in forensic mtDNA population databases. Int J Legal Med 118: 267–273.
14. Forster P (2003) To err is human. Annals Hum Genet 67: 2–4.
15. Salas A, Prieto L, Montesino M, Albarran C, Arroyo E, et al. (2005) Mitochondrial DNA error prophylaxis: assessing the causes of errors in the GEP'02–03 proficiency testing trial. Forensic Sci Int 148: 191–198.
16. Brandstatter A, Niederstatter H, Parson W (2004) Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region. Int J Legal Med 118: 47–54.
17. 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.
18. Loogvali EL, Roostalu U, Malyarchuk BA, Derenko MV, Kivisild T, et al. (2004) Disuniting uniformity: A pied cladistic canvas of mtDNA haplogroup H in Eurasia. Mol Biol Evol 21: 2012–2021.
19. Behar DM, Metspalu E, Kivisild T, Achilli A, Hadid Y, et al. (2006) The matrilineal ancestry of Ashkenazi Jewry: Portrait of a recent founder event. Am J Hum Genet 78: 487–497.
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. 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.
22. Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, 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.
23. Knight A (2003) The phylogenetic relationship of Neandertal and modern human mitochondrial DNAs based on informative nucleotide sites. J Hum Evol 44: 627–632.
24. Bandelt HJ, Kivisild T, Parik J, Villems R, Bravi CM, et al. (2006) Lab-Specific Mutation Process. In: Bandelt HJ, Macaulay V, Richards M, editors. Human Mitochondrial DNA and the Evolution of Homo sapiens. 1st ed. Berlin: Springer. pp. 117–146.
25. Chinnery PF (2006) Mitochondrial DNA in Homo Sapiens. In: Bandelt HJ, Macaulay V, Richards M, editors. Human Mitochondrial DNA and the Evolution of Homo sapiens. 1st ed. Berlin: Springer. pp. 3–15.
26. 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.
27. Richards MB, Macaulay VA, Bandelt HJ, Sykes BC (1998) Phylogeography of mitochondrial DNA in western Europe. Ann Hum Genet 62: 241–260.
28. 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.
29. YCC (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res 12: 339–348.
30. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186–194.
31. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8: 175–185.
32. Petkov PM, Ding Y, Cassell MA, Zhang W, Wagner G, et al. (2004) An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res 14: 1806–1811.
33. Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, et al. (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res 35: D823–D828.
34. Forster P, Harding R, Torroni A, Bandelt HJ (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet 59: 935–945.
35. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, et al. (2004) Ethiopian mitochondrial DNA heritage: Tracking gene flow across and around the gate of tears. Am J Hum Genet 75: 752–770.
36. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, et al. (2004) The African diaspora: Mitochondrial DNA and the Atlantic slave trade. Am J Hum Genet 74: 454–465.
37. Finnila S, Majamaa K (2001) Phylogenetic analysis of mtDNA haplogroup TJ in a Finnish population. J Hum Genet 46: 64–69.
38. Bandelt H-J, Kong Q-P, Richards M, Macaulay V (2006) Estmation of mutation rates and coalescence times: Some caveats. In: Bandelt H-J, Macaulay V, Richards M, editors. Human Mitochondrial DNA and the Evolution of Homo sapiens. 1st ed. Berlin: Springer. pp. 47–90.