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\(^7\), Lluís Quintana-Murci\(^8\)\(^9\), Chris Tyler-Smith\(^1\)\(^0\), 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.

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 offer 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
Copyright: © 2007 Behar et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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.

Table 1. Genotyping Parameters of the Reference Database

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

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.

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

<table>
<thead>
<tr>
<th>Hg by SNPs</th>
<th>Final Hg</th>
<th>Entire Database</th>
<th>Reference Database</th>
<th>Consented Data</th>
<th>Consented Data (22 SNPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L0/1</td>
<td>L0*</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>L0a*</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>L0a1</td>
<td>149</td>
<td>49</td>
<td>49</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>L0a2</td>
<td>46</td>
<td>9</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>L0d</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>L0f</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>L1*</td>
<td>62</td>
<td>18</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>L1b</td>
<td>343</td>
<td>73</td>
<td>107</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>L1c*</td>
<td>41</td>
<td>3</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>L1c1</td>
<td>130</td>
<td>23</td>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>L1c2</td>
<td>118</td>
<td>29</td>
<td>35</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>L1c3</td>
<td>44</td>
<td>12</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>L0/L1 total</td>
<td></td>
<td>952</td>
<td>221</td>
<td>279</td>
<td>101</td>
</tr>
<tr>
<td>L2</td>
<td>L2*</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>L2a</td>
<td>780</td>
<td>194</td>
<td>212</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>L2b</td>
<td>133</td>
<td>39</td>
<td>43</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>L2c</td>
<td>185</td>
<td>39</td>
<td>51</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>L2d</td>
<td>47</td>
<td>6</td>
<td>14</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>L5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L2 total</td>
<td></td>
<td>1,151</td>
<td>279</td>
<td>321</td>
<td>122</td>
</tr>
<tr>
<td>L3*</td>
<td>L3*</td>
<td>141</td>
<td>28</td>
<td>46</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>L3b</td>
<td>183</td>
<td>35</td>
<td>46</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>L3d</td>
<td>241</td>
<td>47</td>
<td>66</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>L3e1*</td>
<td>117</td>
<td>34</td>
<td>34</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>L3e1a</td>
<td>40</td>
<td>8</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>L3e1b</td>
<td>12</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>L3e2*</td>
<td>139</td>
<td>23</td>
<td>41</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>L3e2b</td>
<td>172</td>
<td>30</td>
<td>45</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>L3e3</td>
<td>100</td>
<td>26</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>L3e4</td>
<td>30</td>
<td>6</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>L3f</td>
<td>178</td>
<td>33</td>
<td>43</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>L3g</td>
<td>20</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>L3* total</td>
<td></td>
<td>1,373</td>
<td>279</td>
<td>370</td>
<td>102</td>
</tr>
<tr>
<td>M*</td>
<td>M*</td>
<td>1,025</td>
<td>255</td>
<td>290</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>M1</td>
<td>98</td>
<td>21</td>
<td>39</td>
<td>13</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>24</td>
<td>6</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>M* total</td>
<td></td>
<td>1,147</td>
<td>282</td>
<td>336</td>
<td>127</td>
</tr>
<tr>
<td>D</td>
<td>D</td>
<td>825</td>
<td>229</td>
<td>228</td>
<td>101</td>
</tr>
<tr>
<td>N*</td>
<td>N*</td>
<td>651</td>
<td>147</td>
<td>193</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>N9a</td>
<td>225</td>
<td>18</td>
<td>40</td>
<td>5</td>
</tr>
<tr>
<td>N* total</td>
<td></td>
<td>302</td>
<td>36</td>
<td>63</td>
<td>11</td>
</tr>
<tr>
<td>N1*</td>
<td>N1*</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>N1a</td>
<td>160</td>
<td>29</td>
<td>59</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>N1b</td>
<td>442</td>
<td>106</td>
<td>119</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>N1c</td>
<td>57</td>
<td>12</td>
<td>18</td>
<td>7</td>
</tr>
<tr>
<td>N1* total</td>
<td></td>
<td>665</td>
<td>149</td>
<td>197</td>
<td>69</td>
</tr>
<tr>
<td>I</td>
<td>I</td>
<td>1,769</td>
<td>421</td>
<td>475</td>
<td>170</td>
</tr>
<tr>
<td>W</td>
<td>W</td>
<td>1,252</td>
<td>282</td>
<td>407</td>
<td>144</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>1,080</td>
<td>237</td>
<td>338</td>
<td>111</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>1,413</td>
<td>361</td>
<td>373</td>
<td>130</td>
</tr>
<tr>
<td>R*</td>
<td>R*</td>
<td>186</td>
<td>26</td>
<td>53</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>R1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>R2</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>R5</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>R6</td>
<td>11</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>R* total</td>
<td></td>
<td>214</td>
<td>38</td>
<td>60</td>
<td>17</td>
</tr>
<tr>
<td>U*</td>
<td>U*</td>
<td>958</td>
<td>186</td>
<td>262</td>
<td>89</td>
</tr>
<tr>
<td></td>
<td>U1</td>
<td>22</td>
<td>5</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>U1a</td>
<td>131</td>
<td>22</td>
<td>43</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>U1b</td>
<td>71</td>
<td>17</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>U2</td>
<td>885</td>
<td>195</td>
<td>264</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>U3</td>
<td>437</td>
<td>102</td>
<td>143</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>U4</td>
<td>1,669</td>
<td>348</td>
<td>460</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>U5</td>
<td>480</td>
<td>110</td>
<td>150</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>U5a</td>
<td>907</td>
<td>196</td>
<td>249</td>
<td>83</td>
</tr>
<tr>
<td></td>
<td>U5a1</td>
<td>667</td>
<td>130</td>
<td>183</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>U5a1a</td>
<td>2,168</td>
<td>483</td>
<td>600</td>
<td>206</td>
</tr>
</tbody>
</table>
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, transversions, 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 transversions 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 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
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 haplotype 16030 by state 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

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 16030 by descent. Samples A and D or B and D share HVS-I haplotype 16030 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.

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–16270T–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 16232T–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, T17050C 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 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.
expected to harbor transitions A11719G, T14766C, and T7028C. However, 83 of the total 6292 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 16,608 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
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

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
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 NNw-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
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 depurate 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 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., 16059A). When a doublet has been inferred by 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 (16185C). 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 be replaced by other SNPs. Hg M is one of a large number of subclades 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 fall in this or any different Hg M sub-branch. Hgs such as M are directly inspected by an operator [30,31]. (3) All positions that differ from the rCRS are recorded electronically. (4) Forward and reverse amplifications, all samples are subject to bidirectional sequencing (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 confirmed for each sample for each SNP. (2) Sequences (Genetic 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 observer. (7) All sequences are double-verified by an independent observer. (8) All samples are subject to bidirectional sequencing 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 base substitutions (i.e., A, T, C, G, and not 16126G followed by 16093C). (10) All reported variant positions 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 dataset 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, 6571, 8994, 10054, 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 7920 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 12413 and 3970. Therefore, the current panel includes all coding-region SNPs and the 9-bp deletion. Among the currently populations in which the project is advertised [2,34,35]. For example, the R0 clade within macroHgs R and N is not mentioned. (3) Hgs by which the HVS-I projects is named. It 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 part of the mtDNA phylogeny, the "inconsistent" 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 3% Hg of 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: L23* '5 '6 '7, L3* '7 'M, D, C, N, I, A, W, X, R, R9, B, J, T, U, K, R0, HV, V, and H [21]. Therefore, a total of 23 Hg clusters are inferred from the SNP panel. The coding-region SNPs by including the following ten SNPs: 4248, 6571, 8994, 10054, 10238, 10550, 12612, 13263, 13368, and 13928. The panel was further augmented to a total of 22 coding-region SNPs by including the following additional ten SNPs: 4248, 6571, 8994, 10054, 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 7920 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 12413 and 3970. Therefore, the current panel includes all coding-region SNPs and the 9-bp deletion. Among the current populations in which the project is advertised [2,34,35]. For example, the R0 clade within macroHgs R and N is not mentioned. (3) Hgs by which the HVS-I projects is named. It 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 part of the mtDNA phylogeny, the "inconsistent" 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 3% Hg of labeling, based on NN methodology, is developed herein.
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
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 by applying the principle of parsimony, and by applying the principle of the most similar. The final Hg can be further supported by the HVS-I information.

**Statistical analysis. General indices.** Success rates of each of the genotyping protocols (frequency distributions including the frequency of transitions, transversions, 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 phenotypic 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 be able to 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 $s_1, ... , s_n$ and a new HVS-I sample $t$, we define the pair-wise distance as $d(s_i, t) = \sum_j w_j I(\hat{t}_j \neq \hat{s}_j)$, where $f$ is the set of HVS-I loci (defined as 16092–16569 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-weight the loci with a higher mutation rate (such as 16311).

Denote these mutation rates (in units of substitutions per year) as $p_1, ... , p_j$. Then a weighting of $w_j = \log(20,000 / p_j)$ 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 ($20,000 \times p_j$) distribution, under sufficiently simple substitution models. If we assume that 20,000 $\times p_j$ is still very small, as would be the case for practically all sites, then we can approximate the Poisson by a Bernoulli ($20,000 \times (p_j$ where 1 if two alleles 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 $d(s_i, t) = \sum_j w_j I(\hat{t}_j \neq \hat{s}_j)$.

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 probabilities 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 situations where 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.s0001 (8.1 MB XLS).

**Table S1.** Amplification and Sequencing Primers

Found at doi:10.1371/journal.pgen.0030104.s0001 (4 KB XLS).

**Table S2.** Hg-Predicting Motifs

Found at doi:10.1371/journal.pgen.0030104.s0002 (9 KB XLS).

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

Found at doi:10.1371/journal.pgen.0030104.s0003 (457 KB XLS).

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

Found at doi:10.1371/journal.pgen.0030104.s0004 (189 KB XLS).

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

Found at doi:10.1371/journal.pgen.0030104.s0005 (48 KB XLS).

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

Found at doi:10.1371/journal.pgen.0030104.s0006 (17 KB XLS).

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

Found at doi:10.1371/journal.pgen.0030104.s0007 (745 KB XLS).

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

Found at doi:10.1371/journal.pgen.0030104.s0008 (16 KB XLS).

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

Found at doi:10.1371/journal.pgen.0030104.s0009 (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.s0010 (10 KB XLS).

**Table S11.** Divergent Measures of Neanderthals (Entire Database)

Found at doi:10.1371/journal.pgen.0030104.s0011 (4 KB XLS).

**Video S1.** Introductory Video of the Genographic Project

Found at doi:10.1371/journal.pgen.0030104.s0001 (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; Fabrizio R. Santos, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; Lluís Quinlan-Murcet, Institut Pasteur, Institut Pasteur, Paris, France; Jaime 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; Ajan K. Royyuru, IBM, Yorktown Heights, New York, United States of America; Saharon Rosset, IBM, 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 Mitochondrial DNA Database
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