

The Sorenson Molecular Genealogy Foundation Publicly Accessible mtDNA Database for Ancestral and Population Studies

PMS.07 · 0+P03

U.A. Perego,^{1,2} A. Achilli,² N.M. Myres,¹ K.H. Ritchie,¹ N. Angerhofer,¹ R. Hughes,¹ A. Torroni,² S.R. Woodward¹

¹Sorenson Molecular Genealogy Foundation (Salt Lake City, Utah, USA); ²Dipartimento di Genetica e Microbiologia (Pavia, Italy)

ugo@smgf.org; achilli@ipvgen.unipv.it

ABSTRACT

The Sorenson Molecular Genealogy Foundation (<http://www.smgf.org>) is a non-profit research organization with the aim of studying the relationships and histories of modern human populations by creating a publicly accessible database of correlated genetic and genealogical records. In this project, participants from all over the world have the opportunity to submit a copy of their pedigree chart together with a small DNA sample. After seven years, we have collected over 80000 DNA samples and corresponding genealogical records representing 117 countries. One of the first goals is to sequence, for each subject, the mitochondrial DNA (mtDNA) from nucleotide position (np) 15841 to np 720 (1449 bp), which includes the complete control region. To date, 25104 mtDNAs linked to 111991 unique maternal ancestors are available in the SMGF database, providing the general public with the opportunity to find common maternal links and view the temporal and spatial data associated with each haplotype. Moreover, this dataset is a resource for the scientific community in studies concerning the mtDNA phylogeny and its mutational dynamics, and the origin of specific haplogroups.

INTRODUCTION

The human mitochondrial DNA (mtDNA) is a circular molecule of 16569 base pairs (bps) that encodes for 37 genes: 13 for proteins, 22 for transfer RNA (tRNA), and two for ribosomal RNA (rRNA) (Figure 1). This extra-nuclear genome is maternally transmitted and its sequence variation, which cannot be reshuffled by recombination as in autosomal genes, has been generated exclusively by the sequential accumulation of new mutations along radiating maternal lineages. This process of molecular differentiation gave rise to monophyletic units called haplogroups – the branches of the tree in Figure 2 – that are typical of different geographic areas and population groups. As a result the human mtDNA can provide a lot of information about human origin [1,2] and migrations [3] especially using a “phylogeographic” approach [4]. Moreover, these and other peculiar features make the mtDNA a valuable source for different types of research including population, historical [3,5], forensic, medical [6], and even genealogical publications [7]. The vast amount of new knowledge that has become available through these investigations and the availability of consumer-oriented genetic tests offered by a number of laboratories in the USA and in Europe have generated a wave of interest toward DNA testing among non-academic circles. A large array of mtDNA testing services has produced thousands of mtDNA haplotypes shared by individuals over the internet on personal websites or in publicly accessible databases. Moreover, two large projects are currently underway to collect hundreds of thousands of DNA samples representing populations and lineages around the world to increase our understanding of the recent and deep history of mankind [National Geographic’s Genographic Project, Sorenson Molecular Genealogy Foundation Project]. Personal genetic testing and the availability of searchable databases of correlated genetic and ancestral/geographical data complement each other, as they are both needed to analyze the ancient and recent human history. The paradox associated with consumer-based genetic testing is that in order to understand one’s own genetic history, it is required to know something about the genetic past of everyone else. Thus, once a person obtains a copy of his or her mtDNA haplotype, a comparison to other haplotypes, each correlated with additional historical and geographical data will often produce previously unavailable information about the individual’s maternal lineage. Thousands of searchable haplotypes and corresponding genealogical and geographical data are currently available through the non-profit Sorenson Molecular Genealogy Foundation in their online mtDNA database at <http://www.smgf.org>.

METHODS

The Sorenson Molecular Genealogy Foundation (SMGF) is a non-profit research organization with the objective of building large databases of correlated genetic and genealogical data to be used for ancestral and population studies. Data for this project is collected from volunteer participants worldwide who have read and signed a written consent form in their native tongue. Each participant in the study has donated a small DNA sample collected either as a blood draw, a cheek swab, or a mouthwash rinse together with a copy of their pedigree data (names, places and dates of birth for as many ancestors as possible). The biological specimens are coded to protect the participants’ confidentiality and the DNA extraction, pre-PCR, PCR, and sequencing procedures for each sample is outsourced to the Sorenson Genomics Laboratories (<http://www.sorensongenomics.com>). Genealogical data are screened, linked, and extended by a team of professional genealogists and the genetic and genealogical data thus produced are correlated by the SMGF bioinformatics team and posted in searchable databases online. To date, two such databases are available on the project website at <http://www.smgf.org>: the Sorenson Y Chromosome Database, comprising 19113 Y chromosome haplotypes representing 95213 unique paternal ancestors and 11293 surnames, and the Sorenson Mitochondrial DNA Database, comprising 25104 haplotypes representing 111991 unique maternal ancestors and 20385 unique surnames. Participants in the Y chromosome and mitochondrial DNA databases descend from ancestors from 117 countries (Figure 3). The two databases are updated quarterly.



Figure 3. Sample collection map



Figure 4. MtDNA search results from the SMGF db

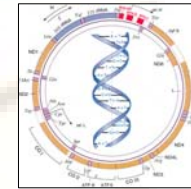


Figure 1. Human mtDNA

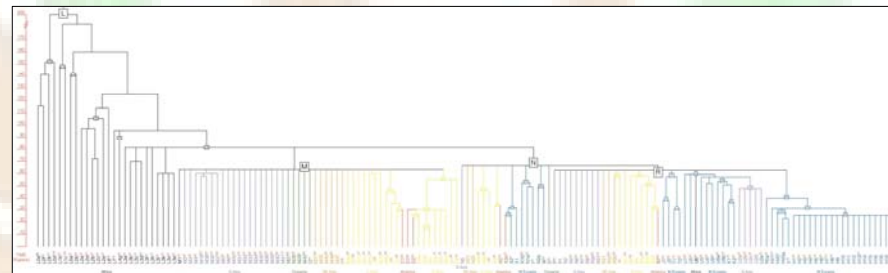


Figure 2. Schematic representation of the worldwide phylogeny of human mtDNA

THE SORENSON MITOCHONDRIAL DNA DATABASE

The Sorenson Mitochondrial DNA Database (Sorenson mtDNA db) was launched on June 2006 and at that time it contained a little over 5000 haplotypes and corresponding genealogical data. In a little over a year, the database has experienced a fivefold growth with an estimated 80000 haplotypes by the beginning of 2008. The objective of SMGF is to eventually sequence each DNA sample from np 15841 to np 720 (1449 bp), which includes the three hypervariable segments (HVS-I: np16024-16383, HVS-II: np657-372, and HVS-III: np438-576) of the mtDNA control region (Fig. 1). Only mutational differences relative to the reference sequence (rCRS) [8] are reported in the database and data can be searched either by haplotype or by surname. A typical user that has received his or her mtDNA test results from a commercial laboratory or that is interested in a particular set of mutations may enter a given haplotype in the search page and select to view mtDNA haplotypes that are a perfect match, or are one or two mutations off from it (Figure 4). Each displayed haplotype in the result page has a corresponding viewable pedigree chart with genealogical data from the participant in the SMGF database. Genealogical data from the last 100 years is withheld to protect participants’ confidentiality. While mtDNA haplotypes alone are not a sufficient tool to confidently ascertain common maternal ancestry between two individuals sharing an exact control region sequence, it can still provide initial clues about one’s personal ancestry. This is particularly true when the haplotype is particularly rare in the general population and when the genealogical data of the two individuals sharing a common mtDNA haplotype also contains clues of a possible common ancestry. The combined genetic and genealogical data available in the Sorenson mtDNA db has already provided valuable leads to individuals who were researching their maternal ancestral roots [7,9].

A virtual lineage map using Google maps technology displays the spatial distribution of the mtDNA haplotypes in intervals of twenty years based on the geographic data provided on the pedigree charts of the donors. Also, based on 86309 mother/child pairs, the database mean generation interval is 27.9 years, with a median of 27 years. The database’s most common mutations are also displayed in a separate table. In the database we have identified more substitutions (188343) than insertions (47375) or deletions (11061) with an average number of mutations per sequence of 9.9. To date 1337 are unique mutations.

In addition to the online searching capabilities and statistical data, the vast amount of mtDNA haplotypes produced by SMGF provide a valuable tool to the scientific community which is interested in studying specific haplotypes, haplogroups, and populations. Figure 5 is just an example of the geographical distribution of three mitochondrial haplogroups (B2, H, and L3) from data contained in the Sorenson mtDNA db. Further analysis of the data contained in the Sorenson mtDNA db will assist in identifying new subclades and better define old ones in an attempt to continually improve the world mtDNA phylogeny (Figure 2). Indeed, after evaluating each study proposal, an internal query of the database can be performed by the SMGF bioinformatics team and the results can be analyzed to determine if the samples of interest are available in the SMGF database. While no personal information is made available to third parties so as to protect the participants’ confidentiality, genetic data may be employed for scientific studies and publications.

ACKNOWLEDGMENTS

We gratefully acknowledge Amanda Pollock’s contribution in creating the mtDNA distribution maps. We also wish to thank all the other SMGF employees for their work in collecting and processing the data and all the volunteer participants who donated DNA samples and genealogical records for the SMGF databases.

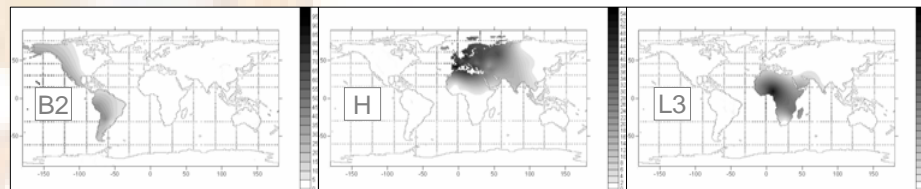


Figure 4. Geographical distribution of mtDNA haplogroups B2, H and L3 from data contained in the Sorenson mtDNA database

REFERENCES

1. Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Moreau W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308: 1034-6 (2005).
2. Achilli A, Olivieri A, Pala M, Metspalu E, Fornarino S, Battaglia V, Accetturo M, Kucuev I, Khushnudinova E, Pennarun E, Cerutti N, Di Gaetano C, Crobu F, Falli D, Matullo G, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Semino O, Villems R, Bandelt HJ, Piazza A, Torroni A. Mitochondrial DNA variation of modern Tuscans supports the Near East origin of Etruscans. *Am J Hum Genet* 80: 759-768 (2007).
3. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugonjon JM, Coudray C, Santachiara-Benerecetti AS, Semino O, Bandelt HJ, Torroni A. The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science* 314: 1767-70 (2006).
4. Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ. Harvesting the fruit of the human mtDNA tree. *Trends Genet* 22: 339-45 (2006).
5. Gill P, Ivanov PL, Kimpton C, Percy R, Benson N, Tully G, Even I, Hagelberg E, Sullivan K. Identification of the Remains of the Romanov Family by DNA Analysis. *Mol Biol Evol* 16: 130-5 (1999).
6. Carelli V, Achilli A, Valentini ML, Rengo C, Semino O, Pala M, Olivieri A, Mattarzi M, Palchetti F, Carrara F, Zeviani M, Leuzzi V, Carlucci C, Valle G, Simonian B, Mendetti L, Salomao S, Belfort R Jr, Sadun AA, Torroni A. Haplogroup effects and recombination of mtDNA: novel clues from the analysis of LHON pedigrees. *Am J Hum Genet* 78: 564-74 (2006).
7. Perego UA, Woodward SR. Mountain Meadows Survivor? A Mitochondrial DNA Examination. *JMH* 32:45-53 (2006).
8. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23:147
9. Seattle woman has surprising success researching her ancestral grandmothers by using new Sorenson mtDNA-genealogy database – available at http://www.smgf.org/press_release_spx?pr=15