Bioinformatics, 31(8), 2015, 1310–1312
doi: 10.1093/bioinformatics/btu825
Advance Access Publication Date: 12 December 2014
Applications Note

Sequence analysis

Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier

Daniel Navarro-Gomez[1], Jeremy Leipzig[2], Lishuang Shen[1], Marie Lott[3], Alphons P.M. Stassen[4], Douglas C. Wallace[3,5], Janey L. Wiggs[1], Marni J. Falk[5,6], Mannis van Oven[7] and Xiaowu Gai[1,*]

[1]Department of Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear Infirmary, Boston, MA, USA, [2]Center for Biomedical Informaticsand [3]Center for Mitochondrial and Epigenomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, USA, [4]Department of Clinical Genetics, Maastricht University Medical Centre, The Netherlands, [5]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA, [6]Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, PA, USA and [7]Department of Forensic Molecular Biology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on September 14, 2014; revised on December 6, 2014; accepted on December 8, 2014

Abstract

Motivation: All current mitochondrial haplogroup classification tools require variants to be detected from an alignment with the reference sequence and to be properly named according to the canonical nomenclature standards for describing mitochondrial variants, before they can be compared with the haplogroup determining polymorphisms. With the emergence of high-throughput sequencing technologies and hence greater availability of mitochondrial genome sequences, there is a strong need for an automated haplogroup classification tool that is alignment-free and agnostic to reference sequence.

Results: We have developed a novel mitochondrial genome haplogroup-defining algorithm using a k-mer approach namely Phy-Mer. Phy-Mer performs equally well as the leading haplogroup classifier, HaploGrep, while avoiding the errors that may occur when preparing variants to required formats and notations. We have further expanded Phy-Mer functionality such that next-generation sequencing data can be used directly as input.

Availability and implementation: Phy-Mer is publicly available under the GNU Affero General Public License v3.0 on GitHub (https://github.com/danielnavarrogomez/phy-mer).

Contact: Xiaowu_Gai@meei.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Human mitochondrial DNA (mtDNA) sequences can be divided into distinct lineages known as haplogroups, which are important for population genetics and disease studies. Conventional mtDNA haplogroup assignment involves alignment of the mtDNA sequence to a reference sequence and then matching the list of variants with the best-fitting haplogroup in a phylogenetic classification tree, mostly PhyloTree (http://www.phylotree.org) (van Oven and Kayser, 2009). Existing software tools include MitoTool (Fan and Yao, 2013), HaploGrep (Kloss-Brandstätter *et al.*, 2011), EMMA (Röck *et al.*, 2013), H-mito (http://sourceforge.net/projects/h-mito/), mthap (http://dna.jameslick.com/mthap), HAPLOFIND (Vianello

et al., 2013) and MToolBox (Calabrese *et al*., 2014). Most of these tools, except MToolBox, require a list of mtDNA variants properly denoted according to different conventions in use (Bandelt and Parson, 2008; Budowle *et al*., 2010). Moreover, there are currently two mitochondrial reference sequences in use: the traditional revised Cambridge Reference Sequence GenBank NC_012920 (Anderson *et al*., 1981; Andrews *et al*., 1999) and the recently created Reconstructed Sapiens Reference Sequence (Behar *et al*., 2012). The same variant may be named differently depending on the reference sequence used. As a result, haplogroup classification remains cumbersome and error-prone.

Given the need to simplify this process, we developed an alignment-free and reference-independent algorithm that encompasses the entire process from sequence input to haplogroup output without requiring additional efforts from the end user.

2 Methods

2.1 K-mer library of representative haplogroup sequences

PhyloTree Build 16 delineates a total of 4806 haplogroups, each of which is defined by a unique set of haplogroup-defining single nucleotide polymorphisms (SNPs). Representative sequences corresponding to each haplogroup were reconstructed based on these polymorphisms, followed by the construction and optimization of sets of k-mers (DNA subsequence strings of length k) whose combination is uniquely present in only the corresponding representative haplogroup sequence.

2.2 Haplogroup classification

When a mitochondrial sequence is provided, Phy-Mer first decomposes it into a set of all possible k-mers, which are then compared against each of the k-mer sets of all haplogroups. IUPAC codes in the sequence can be properly handled. A score is derived for each haplogroup. Details about Phy-Mer score are provided in the Supplementary Information. The haplogroup with the top score is assigned as the query sequence's classification. When multiple haplogroups share the same top score, all top haplogroups are returned. When SNP calls are given for an mtDNA sequence, a representative sequence can be generated with a utility script, which is then used as input. For next-generation sequencing (NGS) data, Phy-Mer takes input in the standard FASTQ or BAM format. K-mers are derived from individual reads, which are then used as a set for haplogroup classification in the same way as if using single mtDNA sequences.

2.3 Implementation, installation and computational performance

Phy-Mer is written in Python v2.7.3, which supports multiple platforms including Windows, Linux and OSX. Detailed installation instructions, tutorials, performance metrics and usage examples are provided at Phy-Mer website on GitHub. They are also provided in the Supplementary Methods, along with detailed explanations of each aforedescribed method.

3 Results

3.1 Haplogroup determination using full-length and partial mtDNA sequences

For performance evaluation, we ran Phy-Mer, as well as HaploGrep [via MitoMaster (Lott *et al*., 2013)], on 6924 full-length mtDNA genome sequences that were manually assigned to haplogroups when PhyloTree Build 16 was constructed. HaploGrep is the most popular and currently one of the most accurate tools for mitochondrial haplogroup assignment (Bandelt *et al*., 2012). Results are summarized in Figure 1. In total, both Phy-Mer and HaploGrep calls had perfect agreements with the original PhyloTree assignment in 6780 cases (97.9%). Specifically, Phy-Mer agreed with the original PhyloTree assignments of 6870 sequences (99.2%) while HaploGrep agreed with the original PhyloTree assignments of 6826 sequences (98.6%). This shows comparable performance between Phy-Mer and HaploGrep.

Phy-Mer made slightly different haplogroup calls for 54 mtDNA genome sequences compared with their original PhyloTree assignments (Supplementary Table S1). In 90 cases, Phy-Mer calls agreed with the original PhyloTree assignments but not with HaploGrep calls. Among them, 74 discrepancies (Supplementary Table S2) could be traced back to different notations and difficulty in calling indels in repetitive, low-complexity regions ('jumping alignments') (Den Hartog *et al*., 2009).

The unique algorithm that we developed is capable of haplogroup classification using partial mtDNA sequences. To evaluate this, we use the combined first 600 bp and last 600 bp (roughly corresponding to the control region or D-loop) of each of the 6924 full-length mtDNA sequences as inputs. Phy-Mer provided identical calls in 2607 (37.7%) cases and overlapping calls in 2372 (34.3%) cases comparing to full-length calls. For the 1945 (28.1%) cases that different calls were given, the top-level of haplogroup assignments were identical for 1849 (95.1%). While the reduced precision is expected, the results support the robustness of the Phy-Mer algorithm, which is relevant given that investigators still frequently use these regions of mtDNA sequences for forensic purposes. In comparison, Phy-Mer analyses of randomly selected 1200 bp subsequences of the 6924 sequences showed only 3986 (57.6%) top-level haplogroup matches, reflecting the known enrichment of the control region for (haplogroup-defining) polymorphisms.

3.2 Haplogroup determination using NGS data

To assess the performance of Phy-Mer on NGS data sets, we ran Phy-Mer on 58 data sets that have 99.99% of the mitochondrial genome sequenced at a minimum depth of 10X using capture probes that we developed in collaboration with Agilent Inc. (Falk *et al*., 2012). We did the same analysis with MToolBox, which was the only other tool capable of haplogroup assignment using NGS data. The two fully agreed in 53 out of 58 cases (91.4%) (Supplementary Table S3). In the three discordant cases, the Phy-Mer calls overlapped with MToolBox calls. In the two other cases, the calls differed but agreed in the broader haplogroup.

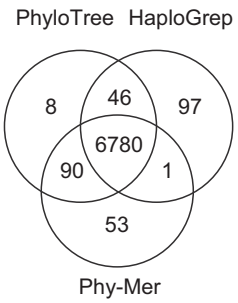

Fig. 1. Haplogroup calls by Phy-Mer and HaploGrep are largely in agreement with PhyloTree Build 16

4 Discussion

Phy-Mer is a novel algorithm in that it avoids the alignment and variant-calling step, unlike all other haplogroup classifiers, yet performs equally well with a single command as the leading haplogroup classifiers. Phy-Mer's k-mer library can be easily updated when there is a new release of PhyloTree. Since it is implemented as a Python package with few dependencies, Phy-Mer is platform-independent. It can be run like a UNIX command, offered as a Web service, or incorporated into any NGS analysis pipeline. We anticipate wide applicability of Phy-Mer given the ever-increasing adoption of NGS technologies in both research and clinical care settings.

Funding

This work was largely supported by institutional fund provided by Massachusetts Eye and Ear Infirmary (MEEI) for MEEI Bioinformatics Center (X.G.). It was also supported in part by the National Institutes of Health (U54-NS078059-North American Mitochondrial Disease Consortium pilot award #NAMDC7407 to M.J.F. and X.G.; and U41-HG006834) and the United Mitochondrial Disease Foundation (UMDF) support for MSeqDR (https://mseqdr.org). J.L., M.L., and D.C.W. were supported in part by the National Institutes of Health grant NS021328 (D.C.W.). MvO was supported in part by a grant from the Netherlands Genomic Initiative (NGI)/ Netherlands Organization for Scientific Research (NWO) within the framework of the Forensic Genomics Consortium Netherlands (FGCN).

Conflict of Interest: none declared.

References

Anderson,S. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457–465.

Andrews,R.M. *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**, 147.

Behar,D.M. *et al.* (2012) A "Copernican" reassessment of the human mitochondrial DNA tree from its root. *Am. J. Hum. Genet.*, **90**, 675–684.

Bandelt,H.J. and Parson,W. (2008) Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int. J. Legal Med.*, **122**, 11–21.

Bandelt,H.J. *et al.* (2012) Haplogrouping mitochondrial DNA sequences in Legal Medicine/Forensic Genetics. *Int. J. Legal Med.*, **126**, 901–916.

Budowle,B. *et al.* (2010) Automated alignment and nomenclature for consistent treatment of polymorphisms in the human mitochondrial DNA control region. *J. Forensic Sci.*, **55**:1190–1195.

Calabrese,C. *et al.* (2014) MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*, **30**:3115–3117.

Den Hartog,B.K. *et al.* (2009) The impact of jumping alignments on mtDNA population analysis and database searching. *Forensic Sci. Int.: Genet. Suppl. Ser.*, **2**, 315–316.

Falk,M.J. *et al.* (2012) Mitochondrial disease genetic diagnostics: optimized whole-exome analysis for all MitoCarta nuclear genes and the mitochondrial genome. *Discov. Med.*, **14**, 389–399.

Fan,L. and Yao,Y.G. (2013) An update to MitoTool: using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion*, **13**, 360–363.

Kloss-Brandstätter,A. *et al.* (2011) HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.*, **32**, 25–32.

Lott,M.T., *et al.* (2013) mtDNA variation and analysis using MITOMAP and MITOMASTER. *Curr. Protoc. Bioinformatics,* **44**: 1.23.1–1.23.26.

Röck,A.W. *et al.* (2013) Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). *Forensic Sci. Int. Genet.*, **7**, 601–609.

van Oven,M. and Kayser,M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, **30**, E386–E394.

Vianello,D., *et al.* (2013) HAPLOFIND: a new method for high-throughput mtDNA haplogroup assignment. *Hum. Mutat.,* **34**, 1189–1194.