A peer-reviewed version of this preprint was published in PeerJ on 25 June 2018.

View the peer-reviewed version (peerj.com/articles/5149), which is the preferred citable publication unless you specifically need to cite this preprint.

Smieszek S, Mitchell SL, Farber-Eger EH, Veatch OJ, Wheeler NR, Goodloe RJ, Wells QS, Murdock DG, Crawford DC. 2018. Hi-MC: a novel method for high-throughput mitochondrial haplogroup classification. PeerJ 6:e5149 https://doi.org/10.7717/peerj.5149

Hi-MC: A novel method for high-throughput mitochondrial haplogroup classification

Sabrina L Mitchell [1] , **Eric H Farber-Eger** [2] , **Olivia J Veatch** [3] , **Robert J Goodloe** [1] , **Quinn S Wells** [4, 5] , **Deborah G Murdock** [6] , **Dana C Crawford** [Corresp. 7, 8]

[1] Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN, United States
[2] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, United States
[3] Department of Neurology, Vanderbilt University Medical Center, Nashville, TN, United States
[4] Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
[5] Department of Pharmacology, Vanderbilt University, Nashville, TN, United States
[6] Center for Mitochondrial and Epigenomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA, United States
[7] Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, United States
[8] Institute for Computational Biology, Case Western Reserve University, Cleveland, OH, United States

Corresponding Author: Dana C Crawford
Email address: dana.crawford@case.edu

Effective approaches for assessing mitochondrial DNA (mtDNA) variation are important to multiple scientific disciplines. Mitochondrial haplogroups characterize branch points in the phylogeny of mtDNA. Several tools exist for mitochondrial haplogroup classification. However, most require full or partial mtDNA sequence which is often cost prohibitive for studies with large sample sizes. The purpose of this study was to develop Hi-MC, a high-throughput method for mitochondrial haplogroup classification that is cost effective and applicable to large sample sizes making mitochondrial analysis more accessible in genetic studies. Using rigorous selection criteria, we defined and validated a custom panel of mtDNA single nucleotide polymorphisms (SNPs) that allows for accurate classification of European, African, and Native American mitochondrial haplogroups at broad resolution with minimal genotyping and cost. We demonstrate that Hi-MC performs well in samples of European, African, and Native American ancestries, and that Hi-MC performs comparably to a commonly used classifier. Implementation as a software package in R enables users to download and run the program locally, grants greater flexibility in the number of samples that can be run, and allows for easy expansion in future revisions. The source code is freely available at https://github.com/vserch/himc .

Hi-MC: A novel method for high-throughput mitochondrial haplogroup classification

Running title: Hi-MC for mitochondrial haplogroup classification

Sabrina L. Mitchell[1], Eric H. Farber-Eger[2], Olivia J. Veatch[3], Robert J. Goodloe[1], Quinn S. Wells[4,5], Deborah G. Murdock[6], Dana C. Crawford[7]

[1]Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[2]Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University, Nashville, TN 37232, USA

[3]Department of Neurology, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[4]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[5]Department of Pharmacology, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[6]Center for Mitochondrial and Epigenomic Medicine, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

[7]Department of Population and Quantitative Health Sciences, Institute for Computational Biology, Case Western Reserve University, Cleveland OH 44106, USA

Corresponding Author

Dana C. Crawford, PhD

2103 Cornell Road, Wolstein Research Building, Suite 2-527

Case Western Reserve University

Cleveland, OH 44106

Telephone: (216) 368-5546

Email: dana.crawford@case.edu

Key words: mitochondrial haplogroups, classification, genetic variation, R

Abstract

Effective approaches for assessing mitochondrial DNA (mtDNA) variation are important to multiple scientific disciplines. Mitochondrial haplogroups characterize branch points in the phylogeny of mtDNA. Several tools exist for mitochondrial haplogroup classification. However, most require full or partial mtDNA sequence which is often cost prohibitive for studies with large sample sizes. The purpose of this study was to develop Hi-MC, a high-throughput method for mitochondrial haplogroup classification that is cost effective and applicable to large sample sizes making mitochondrial analysis more accessible in genetic studies. Using rigorous selection criteria, we defined and validated a custom panel of mtDNA single nucleotide polymorphisms (SNPs) that allows for accurate classification of European, African, and Native American mitochondrial haplogroups at broad resolution with minimal genotyping and cost. We demonstrate that Hi-MC performs well in samples of European, African, and Native American ancestries, and that Hi-MC performs comparably to a commonly used classifier. Implementation as a software package in R enables users to download and run the program locally, grants greater flexibility in the number of samples that can be run, and allows for easy expansion in future revisions. The source code is freely available at https://github.com/vserch/himc.

Introduction

Human mitochondrial DNA (mtDNA) consists of a double-stranded, circular chromosome that spans 16,529 base pairs and encodes 22 transfer RNAs, 2 ribosomal RNAs, and 13 proteins that are part of the oxidative phosphorylation enzyme complexes. Compared with nuclear DNA, unique characteristics of mtDNA include uniparental (i.e. matrilineal) inheritance, lack of recombination, high copy number, and a high mutation rate. These characteristics make mtDNA a powerful tool for investigations in multiple disciplines including population and medical genetics, molecular anthropology, and forensics[1]. Strong evidence exists supporting the involvement of mtDNA variation in human disease phenotypes, underscoring the importance of integrating the mitochondrial genome in genetic association studies. Evidence includes the association of mtDNA single nucleotide polymorphisms (SNPs) and mitochondrial haplogroups with a number of phenotypes encompassing cancer, neurologic, ocular, cardiovascular, and metabolic traits[2-7].

Mitochondrial haplogroups are collections of similar combinations of mtDNA SNPs inherited from a common ancestor. These haplogroups are formed via the sequential accumulation of mutations through the maternal lineage. As a result of population migration, distinct mitochondrial haplogroups are associated with different continental ancestries including African, European, Native American, Asian, and Oceanic[4, 8, 9], allowing for accurate classification of maternal genetic ancestry in large datasets using a small subset of mitochondrial markers.

Currently, several methods are available for mitochondrial haplogroup classification including Haplogrep, HaploFind, MitoTool, HmtDB, MToolBox, and Phy-mer[10-17]. While these methods are powerful tools for mtDNA sequence analysis, including classification of mitochondrial haplogroups, most require full or partial mtDNA sequence, and some are limited in the number of samples that can be processed at once. To address limitations of existing methods we developed a high-throughput method for automated mitochondrial haplogroup classification

that can accommodate large sample sizes with SNP data recorded in the widely used pedigree (PED/MAP) file format.

Using a custom panel of mitochondrial SNPs we constructed a reduced mitochondrial phylogenetic tree, and developed an algorithm (Hi-MC) for broad classification of European, African, and Native American mitochondrial haplogroups. After employing Hi-MC, we determined mitochondrial haplogroup classifications of samples from the International HapMap Project[18-20]. To evaluate the performance of the algorithm we compared Hi-MC mitochondrial haplogroup classifications with those previously reported by HapMap and with classifications generated via Haplogrep, the most widely used web-based application for mitochondrial haplogroup classification. As expected, given the mitochondrial SNPs included in the custom panel, Hi-MC performs well on samples of European, African, and Native American ancestry, but does not perform as well resolving mitochondrial haplogroup in samples of Asian ancestry. Although Hi-MC does not yet resolve mitochondrial haplogroups for all populations, it provides a user-friendly method for high-throughput classification and is provided in an R software package that can be easily expanded in future revisions to capture additional mitochondrial haplogroups.

Materials and methods

Algorithm

The algorithm input is a list of mitochondrial SNP genotypes for each individual DNA sample, and the output is haplogroup classification. The Cambridge reference sequence (rCRS) is used to specify SNP positions. PhyloTree, a comprehensive phylogenetic tree of human mtDNA variation displaying relationships between mitochondrial haplogroups[21], was used as a reference to create a reduced tree of 46 common haplogroups as presented in Mitchell et al[22]. This reduced classification tree was converted into a node-based tree structure. Each haplogroup node has a list of associated SNPs, a parent node, and zero or more child nodes. The SNPs associated with a

node define which SNP genotypes a subject must possess to belong to the corresponding haplogroup. Classification into a haplogroup also requires a subject to recursively meet the definition for the parent haplogroup. Haplogroups that require the reversion to the ancestral genotype (e.g.10398A to 10398G) are accommodated by adding a second hierarchy of required SNP genotypes.

The algorithm determines the appropriate haplogroup in a two-step process (Figure 1). In the first step, the algorithm passes mitochondrial SNP genotype data for each subject into the root node of the tree. The algorithm checks the list of SNP genotypes against those required by the root node. If the array meets the criteria for the parent node, this haplogroup is added to an accumulator. The algorithm then passes the list of SNP genotypes to each of the child nodes connected to that parent node until the tree is exhausted. Next, the algorithm ranks the list of haplogroups in the accumulator according to their distance from the root node. Any haplogroup with a path length less than that of the haplogroup with the longest path length is dropped. The remaining haplogroups, along with their path from the root node to the end node, are returned as a result.

Implementation

The algorithm is implemented as a package in R[23] [https://github.com/vserch/himc]. Data input is standard PED/MAP formatted files. The output is an R dataframe object that includes subject IDs with a corresponding haplogroup classification and the path through the tree from root node to final classification. The output can easily be exported directly to a CSV file or text file. For further details on use of the Hi-MC package in R visit www.icompbio.net.

Mitochondrial SNP Selection

The SNPs were selected for broad classification of European, African, and Native American mitochondrial haplogroup lineages as previously described[22]. Briefly, SNPs were chosen using Phylotree[21] and an extensive literature search for prior studies related to mitochondrial

haplogroup classification[24-27]. Preference was given to those SNPs that appear only once in Phylotree since such SNPs are specific to a single haplogroup. Sixty-three SNPs were selected, the majority of which are located in the coding region of the mitochondrial genome. Three Sequenom genotyping assay pools including all of these SNPs were designed using the MassARRAY software[22]. As described in Mitchell et al[22], the custom SNP panel was genotyped in the National Health and Nutrition Examination Surveys (NHANES) accessed by the Epidemiologic Architecture for Genes Linked to Environment (EAGLE)[28], a study site of the Population Architecture using Genomics and Epidemiology (PAGE) I study[29]. The Vanderbilt University Institutional Review Board determined that EAGLE was "non-human" subjects research.

Application of Hi-MC

To evaluate the performance of Hi-MC for mitochondrial haplogroup classification we genotyped the custom SNP panel in, and applied the algorithm to, HapMap Phase I and Phase III samples. We selected HapMap samples for the present study as HapMap samples were the preferred reference samples for individual study sites including this study as part of the larger PAGE I study[29]. The populations from HapMap Phase I included: individuals of Northern and Western European ancestry from the Centre d'Etude du Polymorphisme Humain samples collected in Utah, USA (CEU, n=90), Yoruba from Ibadan, Nigeria (YRI, n=90), Japanese in Tokyo, Japan (JPT, n=45), and Han Chinese in Beijing, China (CHB, n=45). The HapMap Phase III samples used in this study included only those of Mexican ancestry from Los Angeles, California (MXL, n=90). The International HapMap Consortium reported mitochondrial haplogroup classifications for the CEU, YRI, CHB, and JPT Phase I HapMap samples[20]; however, mitochondrial haplogroup classifications for the Phase III MXL samples have not been previously reported.

We genotyped the custom SNP panel in the CEU, YRI, and CHB/JPT Phase I HapMap samples and in the MXL samples from Phase III. Briefly, aliquots of DNA from HapMap CEU,

YRI, CHB/JPT, and MXL samples were obtained from the Coriell repository. SNPs were genotyped via the Agena Biosciences (formerly Sequenom) iPLEX® Gold MassArray platform. Multiplex primer extension was performed, and extension products were analyzed by MALDI-TOF mass spectrometry[30].

SNP genotyping efficiency was set to greater than or equal to 0.90. The hypervariable region SNP mt16189 did not meet this threshold and was dropped from the analysis. Additionally, SNP mt9540 was excluded from the analysis due to poor genotyping efficiency. We determined that the primers for SNP mt9540 lacked specificity, consistent with the amplification of nuclear insertions of mitochondrial origin (NumtS) common in the human genome[31]. Therefore, SNP mt9540 is not included in the algorithm for classification. The final list of custom panel SNPs used to classify mitochondrial haplogroups is given in Supplementary Table 1.

Using genotype data from the custom SNP panel we employed Hi-MC and Haplogrep to determine mitochondrial haplogroup classifications in the HapMap samples. Although there are several tools available from which to compare Hi-MC, we selected Haplogrep for comparison given it is the most widely used tool to date with >180 citations in the peer-reviewed literature. We then compared the Hi-MC mitochondrial haplogroup classifications to the HapMap-reported classifications for Phase I samples[20]. We also compared Hi-MC haplogroup classifications to Haplogrep-based haplogroup classifications for both Phase I and Phase III HapMap samples. We calculated percent concordance for each comparison. Classifications were considered concordant if they were in the same haplogroup, even if one classification method resulted in finer resolution. For example, if one method classified a sample as A2 and another method classified the same sample as A2x, such classifications were considered concordant. Differences in the resolution of haplogroup classifications were not unexpected given differences in underlying methodology and the number of SNPs used for classification. The HapMap classifications were generated using more mitochondrial SNP genotypes compared to the reduced number of SNPs necessary to use

Hi-MC. HapMap Phase I sample data includes genotypes for 214 mitochondrial SNPs, 49 of which overlap with the custom SNP panel genotyped in this study (Supplementary Table 2). Additionally, Hi-MC uses a reduced tree for classification while Haplogrep employs all of Phylotree which can result in finer sub-haplogroup resolution.

To resolve discordant classifications, possibly due to missing key SNP genotypes, we used the publicly available Phase I HapMap mitochondrial SNP genotype data to determine the mitochondrial haplogroup classification via Haplogrep. If the classification returned from Haplogrep was concordant with the HapMap-reported classification, then we considered the discordance resolved, as it was likely due to missing SNP genotypes necessary for accurate haplogroup classification by Hi-MC.

Results

CEU and YRI populations

Overall, concordance between Hi-MC and both HapMap and Haplogrep was high for the CEU and YRI populations. Among the CEU samples mitochondrial haplogroup classifications were 100% concordant between Hi-MC and HapMap, as well as between Hi-MC and Haplogrep (Table 1). In the YRI samples, concordance between Hi-MC and HapMap was 96.3% (Table 1). Among the YRI samples, three classifications were discordant between Hi-MC and HapMap, one classification was discordant between Hi-MC and Haplogrep, and four classifications were discordant between Haplogrep and HapMap. The three samples that were discordant between Hi-MC and HapMap were also discordant between Haplogrep and HapMap.

Among the eleven YRI samples that were either discordant or unclassified seven were resolved. These samples were missing many SNP genotypes and/or crucial haplogroup-defining SNPs in our genotype data which likely accounts for the discordance. The four YRI samples for which discordance could not be resolved (Y024-NA18861, Y024-NA18663, Y043-NA19137, and Y043-NA19139) were classified as ‘L1’ by HapMap, but were classified as ‘L0a’ by

Haplogrep using HapMap-generated genotype data. The 'L0' classification is consistent with the classification obtained via Hi-MC and Haplogrep when using genotypes from our custom SNP panel. In the HapMap genotype data, all of these samples have eight of the ten SNP genotypes that define haplogroup 'L0', suggesting that 'L0' is the correct classification.

CHB/JPT populations

Compared with the CEU and YRI populations, we observed less concordance among the CHB/JPT samples. Between Hi-MC and HapMap-reported classifications, 37 (41.6%) were concordant at the haplogroup level and 31 (34.8%) were considered concordant at the macro-haplogroup level. Concordance at the macro-haplogroup level is defined as appropriate macro-haplogroup classification in the absence of sub-haplogroup defining SNP genotype data. For example, consider that haplogroup E is a sub-haplogroup of the macro-haplogroup M. Genotypes for SNPs that define haplogroup E were not included in the custom SNP panel; therefore, individuals classified as haplogroup E by HapMap, but classified as haplogroup M by Hi-MC were considered concordant at the macro-haplogroup level. There were 21 (23.6%) discordant classifications among the CHB/JPT samples. These results were not unexpected given that the SNPs included on the custom panel do not capture all Asian-specific haplogroup lineages. Among the 21 CHB/JPT samples that were discordant, two samples were resolved at the haplogroup level and five samples were resolved at the macro-haplogroup level. The remaining samples with discordant classifications could not be resolved.

Determination of mitochondrial haplogroups in HapMap Phase III samples of Mexican ancestry

The mitochondrial haplogroups for the samples of Mexican ancestry from HapMap Phase III have not been previously reported. Samples in this data set include 30 trios of Mexican ancestry from Los Angeles, CA. We applied Hi-MC to determine mitochondrial haplogroups in these samples and characterized the distribution of mitochondrial haplogroups among the MXL. Due to

matrilineal inheritance of mtDNA, offspring have the same mitochondrial haplogroup as their mother; therefore, offspring were excluded when calculating the frequency distribution of mitochondrial haplogroups. One additional sample was excluded from frequency calculations due to poor genotyping efficiency. Overall in the MXL samples, 84.8% of mitochondrial haplogroups identified were of Native American ancestry and 15.3% were of European ancestry (Table 2). The distribution of haplogroups in the HapMap MXL samples is similar to the distribution of haplogroups observed in Mexican Americans ascertained for the National Health and Nutrition Examination Surveys (NHANES)[22].

To further evaluate the performance of Hi-MC, we compared the Hi-MC mitochondrial haplogroup classifications of MXL samples to Haplogrep-based classifications. Percent concordance between Hi-MC and Haplogrep for classification of the MXL samples was 98.9%. There was one sample out of 89 with a discordant mitochondrial haplogroup classification. This sample was missing the haplogroup H-defining SNP genotype therefore Hi-MC was unable to classify the sample beyond haplogroup 'HV.' Haplogrep classified this sample as H1c1b. For this individual the classifications differ between Hi-MC and Haplogrep due to differences in methodology.

Discussion

Using a custom panel of mitochondrial SNPs that we previously applied to participants in the NHANES data sets[22], we developed Hi-MC, a method for high-throughput classification of European, African, and Native American mitochondrial haplogroup lineages. We evaluated the performance of Hi-MC, and with genotype data from the custom SNP panel, demonstrate that Hi-MC performs comparably to the widely-used tool Haplogrep. While Haplogrep is an excellent tool for mitochondrial haplogroup classification that accepts either sequence or SNP genotype data, it was developed primarily for sequence level data. The ability to alternatively genotype a

relatively small number of SNPs (n=63) allows for rapid haplogroup classification in a large number of genetic samples.

Mitochondrial SNPs captured by standard genotyping arrays vary widely, and often the SNPs on these arrays are not informative for haplogroup determination. Hi-MC uses a defined panel of mitochondrial SNPs for classification of mitochondrial haplogroups. This defined panel of SNPs eliminates the need for investigators to spend time identifying appropriate SNPs for mitochondrial haplogroup classification. Additionally, the relatively small number of SNPs in the custom panel makes Hi-MC particularly useful for large data sets where full mitochondrial genome sequencing is not practical. As examples, approaches like Hi-MC promise to be of use to large biobank and cohort efforts such as Million Veteran Program[32] and the UK Biobank[33], both of which continue to rely on cost-effective array-based assays rather than cost-prohibitive sequencing to generate genome-wide and mitochondrial data on hundreds of thousands to a million participants.

Hi-MC employs the commonly used PED/MAP file format as the input. There are a number of software programs that make use of the PED/MAP format, including PLINK[34] which is widely used for analyzing genotypic data. Thus, in contrast to Haplogrep, many Hi-MC users will not have to reformat data prior to use. Additionally, Hi-MC is an R-based software package that can be downloaded and run locally allowing for memory limits that are dependent on the machine where R is being run, thus granting greater flexibility in the number of samples that can be processed at once. Once samples have been classified using Hi-MC, figures or tables displaying haplogroup frequencies can be easily generated via other R packages such as ggplot2[35].

We determined that Hi-MC performs well with samples of European, African, and Native American descent. However, because many Asian-specific haplogroups are not captured by the custom SNP panel it does not perform as well on samples of Asian maternal lineage. While

progress has been made in characterizing the phylogeny of Asian mtDNA[36, 37], in general, the Asian branches of the mitochondrial phylogenetic tree are not as well-defined as other parts of the tree. Thus, compared to other ancestries, classifying Asian lineage haplogroups continues to be more challenging. As more mtDNA sequences are obtained from individuals of Asian descent the phylogeny of mitochondrial genetic variation will be better understood. Future versions of Hi-MC will be updated to incorporate additional knowledge regarding subjects of Asian descent.

We applied Hi-MC to the HapMap Phase III MXL samples as the mitochondrial haplogroups for these participants have not been previously reported. The haplogroup distribution observed in the HapMap Phase III MXL samples is somewhat similar to the recently reported Haplogrep2-generated distribution for the MXL samples sequenced as part of the 1000 Genomes Project[38]. In this newer reference dataset, the most common reported haplogroup is A (25%) followed by B (15%) and C (9%)[38] compared with a higher A (A2) frequency in the present study (39%; Table 2). Overall, the distribution of Native American and European haplogroups in the MXL samples from HapMap Phase III is similar to the distribution observed in the NHANES Mexican American samples[22]. No African lineage mitochondrial haplogroups were identified among the HapMap MXL samples. This differs from the NHANES Mexican Americans in which 4.4% had mitochondrial haplogroups of African ancestry[22]. The lack of African haplogroups in the HapMap MXL samples is likely due to the small sample size and the regional ascertainment of these samples. While the NHANES samples were collected from across the United States, the HapMap Phase III MXL samples were ascertained solely from Los Angeles, CA, therefore are likely to be more homogeneous.

While there are several benefits to Hi-MC, there are some limitations. Currently, Hi-MC employs a reduced mitochondrial phylogenetic tree for classification. As a result, it is currently limited to classification of the major haplogroups of European, African, and Native American lineages, and requires that SNPs from the described custom panel be genotyped. While this panel

was customized for populations expected for the PAGE I study, it is notable that several SNPs in this panel (MT1736, MT2092, MT3552, MT4883, MT10400, MT11177, MT11251, MT11719, MT12007, MT12308, MT12705, MT13368, MT14766) overlap with previously published panels[1, 39], suggesting the potential for both greater resolution and generalizability in future extensions of Hi-MC. Additionally, because the method relies on a limited number of SNPs, it is not very robust to missing genotype data and it has the ability to classify mitochondrial haplogroups at a broad level, but currently cannot capture sub-haplogroups at finer resolution. As such, in instances where sequence level data is available another method for mitochondrial haplogroup classification, such as Haplogrep, would be more appropriate.

Despite these limitations, Hi-MC offers several advantages including a defined panel of mitochondrial SNPs that is used in conjunction with the software for mitochondrial haplogroup classification. Hi-MC utilizes PED/MAP files for a user-friendly input file format, saving time and reducing opportunities for errors to be incorporated into the data. Also, Hi-MC is implemented in the commonly used statistical software environment R allowing for classification of relatively large sample sizes, as well as the ability to easily utilize other available R packages for visualization of results.

Conclusions

We have developed a custom SNP panel and algorithm for mitochondrial haplogroup classification. The algorithm, Hi-MC is implemented in R and makes use of PED/MAP file format for data input. We evaluated the performance of Hi-MC and demonstrate that classifications are comparable to the widely-used tool Haplogrep. Hi-MC offers an algorithm that leverages a validated mtDNA SNP panel for mitochondrial haplogroup classification and is particularly valuable for studies in which sequencing is not feasible.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

Special thanks to Paxton Baker, MS, Melissa Allen, Ping Mayo, MS, and Nathalie Schnetz-Boutaud, PhD for their work in genotyping these samples. This work was supported in part by National Institutes of Health grant [U01 HG004798] and associated American Recovery and Reinvestment Act (ARRA) supplements.

Figure 1: Hi-MC algorithm structure

Input for the algorithm is a list of sample IDs and corresponding SNP genotype data in pedigree (PED/MAP) format. These genotypes are recursively analyzed through a node-based tree structure. Each successive genotype classification is passed on to the Accumulator. They are then ranked according to specificity [longer path through the tree -> more SNPs checked -> more specific], with the most specific haplogroup as the final output. MRCA = most recent common ancestor

References

1. Chaitanya L, van Oven M, Weiler N, Harteveld J, Wirken L, Sijen T, et al. Developmental validation of mitochondrial DNA genotyping assays for adept matrilineal inference of biogeographic ancestry at a continental level. *Forensic Science International: Genetics*. 2014;11(Supplement C):39-51.

2. van der Walt JM, Dementieva YA, Martin ER, Scott WK, Nicodemus KK, Kroner CC, et al. Analysis of European mitochondrial haplogroups with Alzheimer disease risk. *Neuroscience Letters*. 2004;365(1):28-32.

3. K S, Jalali S, Scaria V, Bhardwaj A. MitoLSDB: A Comprehensive Resource to Study Genotype to Phenotype Correlations in Human Mitochondrial DNA Variations. *PLOS ONE*. 2013;8(4):e60066.

4. Wallace DC. Bioenergetics in human evolution and disease: implications for the origins of biological complexity and the missing genetic variation of common diseases. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2013;368(1622).

5. Mitchell S, Hall J, Goodloe R, Boston J, Farber-Eger E, Pendergrass S, et al. Investigating the relationship between mitochondrial genetic variation and cardiovascular-related traits to develop a framework for mitochondrial phenome-wide association studies. *BioData Mining*. 2014;7(1):6.

6. Hudson G, Gomez-Duran A, Wilson IJ, Chinnery PF. Recent Mitochondrial DNA Mutations Increase the Risk of Developing Common Late-Onset Human Diseases. *PLOS Genetics*. 2014;10(5):e1004369.

7. Fetterman Jessica L, Zelickson Blake R, Johnson Larry W, Moellering Douglas R, Westbrook David G, Pompilius M, et al. Mitochondrial genetic background modulates bioenergetics and susceptibility to acute cardiac volume overload. *Biochemical Journal*. 2013;455(2):157-67.

8. Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet*. 2001;2:13.

9. Forster P. Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*. 2004;359(1442):255-64.

10. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, et al. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutation*. 2011;32(1):25-32.

11. Weissensteiner H, Pacher D, Kloss-Brandstätter A, Forer L, Specht G, Bandelt HJ, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res*. 2016;44(W1):W58-W63.

12. Fan L, Yao Y-G. MitoTool: A web server for the analysis and retrieval of human mitochondrial DNA sequence variations. *Mitochondrion*. 2011;11(2):351-6.

13. Rubino F, Piredda R, Calabrese FM, Simone D, Lang M, Calabrese C, et al. HmtDB, a genomic resource for mitochondrion-based human variability studies. *Nucleic Acids Res*. 2012;40(Database issue):D1150-D9.

14. Fan L, Yao Y-G. An update to MitoTool: Using a new scoring system for faster mtDNA haplogroup determination. *Mitochondrion*. 2013;13(4):360-3.

15. Vianello D, Sevini F, Castellani G, Lomartire L, Capri M, Franceschi C. HAPLOFIND: A New Method for High-Throughput mtDNA Haplogroup Assignment. *Human Mutation*. 2013;34(9):1189-94.

16. Navarro-Gomez D, Leipzig J, Shen L, Lott M, Stassen AP, Wallace DC, et al. Phy-Mer: a novel alignment-free and reference-independent mitochondrial haplogroup classifier. *Bioinformatics*. 2015;31(8):1310-2.

17. Calabrese C, Simone D, Diroma MA, Santorsola M, Gutta C, Gasparre G, et al. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritizatoin analysis of human mitochondrial viarnats in high-throughput sequencing. *Bioinformatics*. 2014;30(21):3115-7.

18. The International HapMap Project. *Nature*. 2003;426(6968):789-96.

19. Consortium IH. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52-8.

20. Consortium TIH. A haplotype map of the human genome. *Nature*. 2005;437(7063):1299-320.

21. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*. 2009;30(2):E386-E94.

22. Mitchell SL, Goodloe R, Brown-Gentry K, Pendergrass SA, Murdock DG, Crawford DC. Characterization of mitochondrial haplogroups in a large population-based sample from the United States. *Hum Genet*. 2014;133(7):861-8.

23. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013 2013.

24. Herrnstadt C, Elson JL, Fahy E, Preston G, Turnbull DM, Anderson C, et al. Reduced-Median-Network Analysis of Complete Mitochondrial DNA Coding-Region Sequences for the Major African, Asian, and European Haplogroups. *The American Journal of Human Genetics*. 2002;70(5):1152-71.

25. van der Walt JM, Nicodemus KK, Martin ER, Scott WK, Nance MA, Watts RL, et al. Mitochondrial Polymorphisms Significantly Reduce the Risk of Parkinson Disease. *The American Journal of Human Genetics*. 2003;72(4):804-11.

398 26. Poole Jason C, Procaccio V, Brandon Martin C, Merrick G, Wallace Douglas C. Multiplex
399 analysis of mitochondrial DNA pathogenic and polymorphic sequence variants. *Biol Chem*.
400 2010;391(10):1115-30.
401 27. Paneto GG, Köhnemann S, Martins JA, Cicarelli RMB, Pfeiffer H. A single multiplex
402 PCR and SNaPshot minisequencing reaction of 42 SNPs to classify admixture populations into
403 mitochondrial DNA haplogroups. *Mitochondrion*. 2011;11(2):296-302.
404 28. Crawford DC, Goodloe R, Farber-Eger E, Boston J, Pendergrass SA, Haines JL, et al.
405 Leveraging epidemiologic and clinical collections for genomic studies of complex traits. *Human*
406 *Heredity*. 2015;79(3-4):137-46.
407 29. Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, et al. The Next
408 PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture
409 Using Genetics and Epidemiology (PAGE) Study. *American Journal of Epidemiology*.
410 2011;174(7):849-59.
411 30. Tang K, Oeth P, Kammerer S, Denissenko MF, Ekblom J, Jurinke C, et al. Mining disease
412 susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass
413 spectrometry. *J Proteome Res*. 2004;3(2):218-27.
414 31. Hazkani-Covo E, Zeller RM, Martin W. Molecular Poltergeists: Mitochondrial DNA
415 Copies (numts) in Sequenced Nuclear Genomes. *PLOS Genetics*. 2010;6(2):e1000834.
416 32. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran
417 Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical*
418 *Epidemiology*. 2016.
419 33. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open
420 Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and
421 Old Age. *PLoS Med*. 2015;12(3):e1001779.

422 34. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a
423 tool set for whole-genome association and population-based linkage analysis. *Am J Hum Genet*.
424 2007;81(3):559-75.
425 35. Wickham H. ggpolot2: Elegant graphics for data analysis. New York: Springer-Verlag;
426 2009.
427 36. Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, et al. The emerging limbs
428 and twigs of the East Asian mtDNA tree. *Mol Biol Evol*. 2002;19(10):1737-51.
429 37. Kong QP, Bandelt HJ, Sun C, Yao YG, Salas A, Achilli A, et al. Updating the East Asian
430 mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. *Hum Mol Genet*.
431 2006;15(13):2076-86.
432 38. Rishishwar L, Jordan IK. Implications of human evolution and admixture for
433 mitochondrial replacement therapy. *BMC Genomics*. 2017;18(1):140.
434 39. van Oven M, Vermeulen M, Kayser M. Multiplex genotyping system for efficient
435 inference of matrilineal genetic ancestry with continental resolution. *Investigative Genetics*.
436 2011;2(1):6.

Table 1(on next page)

Table 1.

Percent concordance in CEU and YRI populations for pair-wise comparisons of mitochondrial haplogroup classifications

Table 1: Percent concordance in CEU and YRI populations for pair-wise comparisons of mitochondrial haplogroup classifications

	CEU (n=86*)	YRI (n=82*)
Hi-MC vs HapMap	100%	96.3%
Hi-MC vs Haplogrep	100%	98.8%
Haplogrep vs HapMap	100%	95.1%

*Due to missing genotypes at key haplogroup-defining SNPs four CEU and eight YRI samples were excluded from the percent concordance calculations.

Table 2(on next page)

Table 2. Distribution of mitochondrial haplogroups in the HapMap Phase III samples of Mexican ancestry in Los Angeles, CA

Table 2: Distribution of mitochondrial haplogroups in the HapMap Phase III samples of Mexican ancestry in Los Angeles, CA

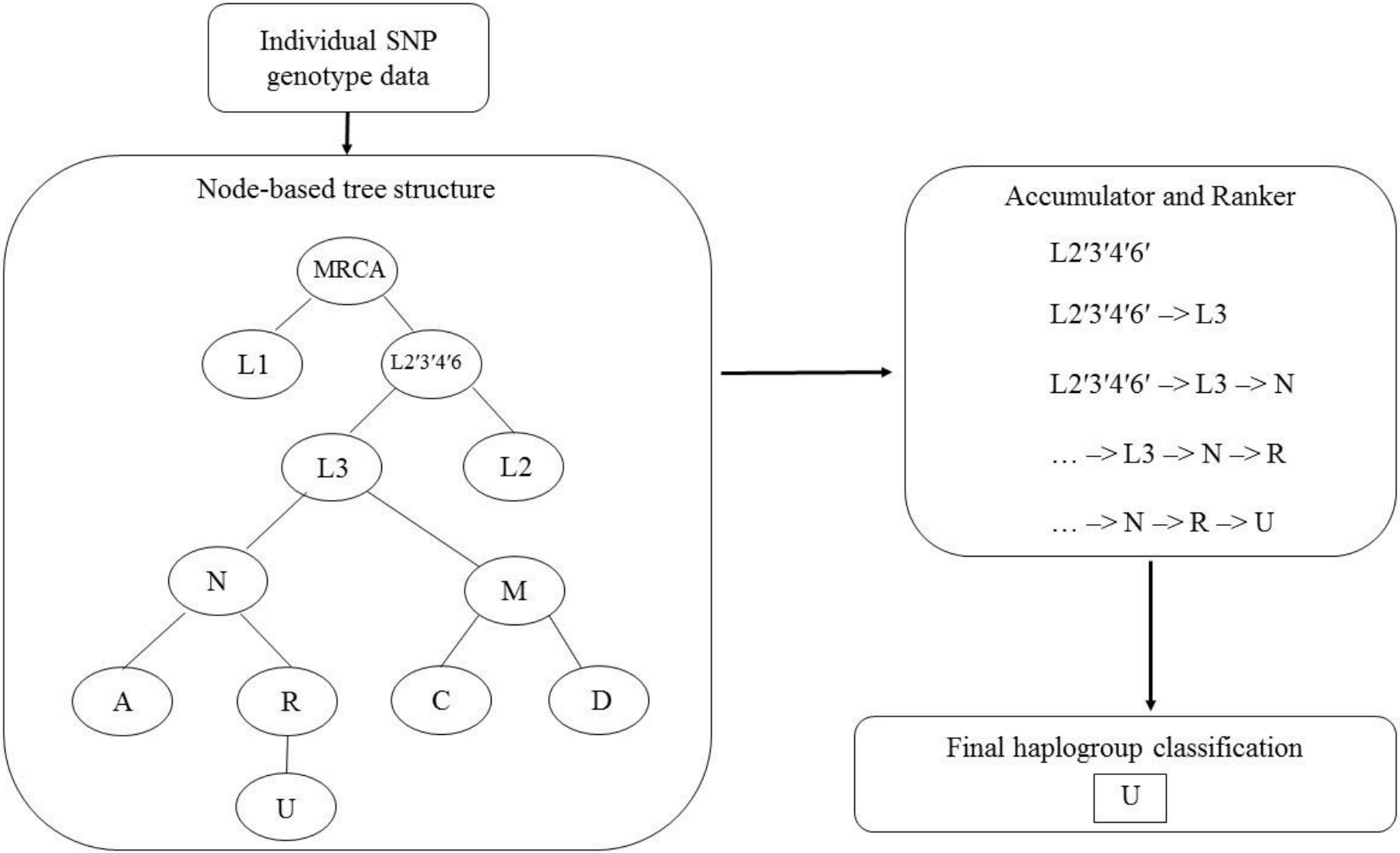
Mitochondrial Haplogroup	Number (%)
Native American	
A2	23 (39.0%)
B2	11 (18.6%)
C	9 (15.3%)
D1	7 (11.9%)
European	
H	3 (5.1%)
H/V	2 (3.4%)
U	2 (3.4%)
V	1 (1.7%)
W	1 (1.7%)

Given that the mitochondrial haplogroup of the offspring is the same as that of the mother, offspring were excluded when determining the frequency distribution of haplogroups. One sample was excluded from frequency calculations due to missing genotype data (n=59).

Figure 1(on next page)

Hi-MC algorithm structure

Input for the algorithm is a list of sample IDs and corresponding SNP genotype data in pedigree (PED/MAP) format. These genotypes are recursively analyzed through a node-based tree structure. Each successive genotype classification is passed on to the Accumulator. They are then ranked according to specificity [longer path through the tree -> more SNPs checked -> more specific], with the most specific haplogroup as the final output. MRCA = most recent common ancestor



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3420v1 | CC BY 4.0 Open Access | rec: 20 Nov 2017, publ: 20 Nov 2017