# Mitochondrial Haplogroup Assignment for High-Throughput Sequencing Data from Single Individual and Mixed DNA Samples

Fahad Alqahtani[1,2][0000−0002−2498−4871] and Ion I. Măndoiu[1][0000−0002−4818−0237]

[1] Computer Science & Engineering Department, University of Connecticut, Storrs, CT, USA

[2] National Center for Artificial Intelligence and Big Data Technology, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

fahad.alqahtani@uconn.edu,ion.mandoiu@engr.uconn.edu

**Abstract.** The inference of mitochondrial haplogroups is an important step in forensic analysis of DNA samples collected at a crime scene. In this paper we introduced efficient inference algorithms based on Jaccard similarity between variants called from high-throughput sequencing data of such DNA samples and mutations collected in public databases such as PhyloTree. Experimental results on real and simulated datasets show that our mutation analysis methods have accuracy comparable to that of state-of-the-art methods based on haplogroup frequency estimation for both single-individual samples and two-individual mixtures, with a much lower running time.

**Keywords:** Mitochondrial analysis · Haplogroup assignment · High-throughput sequencing · Forensic analysis

## 1 Introduction

Each human cell contains hundreds to thousands of mitochondria, each carrying a copy of the 16,569bp circular mitochondrial genome. Three main reasons have made mitochondrial DNA analysis an important tool for fields ranging from evolutionary anthropology [3] to medical genetics [6, 12] and forensic science [1, 4]. First, the high copy number makes it easier to recover mitochondrial DNA (mtDNA) compared to the nuclear DNA, which is present in only two copies per cell [9, 14]. This is particularly important in applications such as crime scene or mass disaster investigations where only a limited amount of biological material may be available, and where sample degradation may render standard forensic tests based on nuclear DNA analysis unusable [20]. Second, mitochondrial DNA has a mutation rate about 10 times higher than the nuclear DNA, making it an information rich genetic marker. The higher mutation rate is due to the fact that mtDNA is subject to damage from reactive oxygen molecules released in mitochondria as by-product of energy metabolism. Finally, mitochondria are

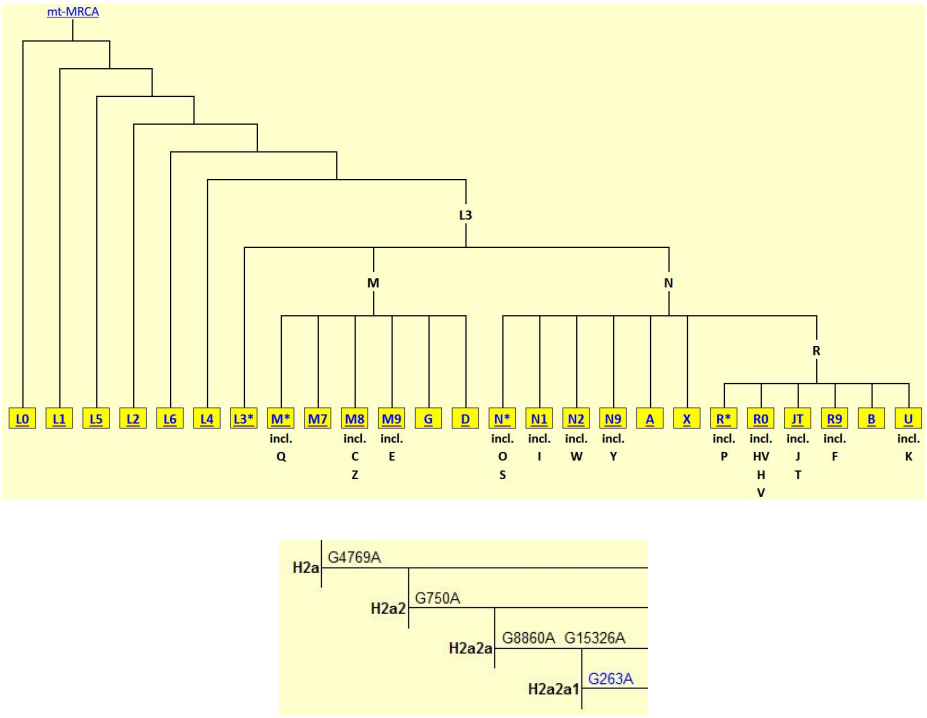


**Fig. 1.** Top level mtDNA haplogroups (top) and sample haplogroups with their mutations (bottom) from Build 17 of PhyloTree [25].

inherited maternally without undergoing recombination like the nuclear genome, which can simplify analysis, particularly for mixed samples [14].

Public databases have already amassed tens of thousands of such sequences collected from populations across the globe. Comprehensive phylogenetic analysis of these sequences has been used to infer the progressive accumulation of mutations in the mitochondrial genome during human evolution and track human migrations [31]. Combinations of these mutations, inherited as *haplotypes*, have also been used to trace back our most recent common matrilinear ancestor referred to as the "mitochondrial Eve" [15, 29]. Last but not least, clustering of mitochondrial haplotypes has been used to define standardized *haplogroups* characterized by shared common mutations [29]. Due to lack of recombination, the evolutionary history of these haplogroups can be represented as a tree. The best curated haplogroup tree is PhyloTree [26], which currently catalogues over 5,400 haplogroups defined over some 4,500 different mutations (see Figure 1).

Although many of the available mtDNA sequences have been generated using the classic Sanger sequencing technology, current mtDNA analyses are mainly performed using short reads generated by high-throughput sequencing technologies. Numerous bioinformatics tools have been developed to conduct mtDNA

analysis of such short read data. The majority of these tools – including MitoSuite [11], HaploGrep [15], Haplogrep2 [33], mtDNA-Server [32], MToolBox [5], mtDNAmanager [16], MitoTool [8], Haplofind [29], Mit-o-matic [28], and HiMC [24] – take a reference-based approach, seeking to infer the haplotype (and assign a mitochondrial haplogroup) assuming that the DNA sample originates from a single individual. While these tools can be helpful for conducting population studies [14] or identifying mislabeled samples [18], they are not suitable for mtDNA analysis of mixed forensics samples that contain DNA from more than one individual, e.g., the victim and the crime perpetrator [30]. Even though the mtDNA haplotypes are not unique to the individual, mitochondrial analysis of mixed forensic samples is useful for including/excluding suspects in crime scene investigations since there is a large haplogroup diversity in human populations [10].

To the best of our knowledge, mixemt [30] is the only available bioinformatics tool that can assign haplogroups based on short reads generated from mixed DNA samples. By using expectation maximization (EM), mixemt estimates the relative contribution of each haplogroup in the mixture. To increase assignment accuracy, the EM algorithm of mixemt is combined with two heuristic filters. The first filter removes any haplogroup that has no support from short reads, while the second filter removes haplogroup mutations that are likely to be private or back mutations. Experiments with synthetic mixtures reported in [30] show that mixemt has high haplogroup assignment accuracy. More recently, mixemt has been used to infer mitochondrial haplogroup frequencies from short reads generated from urban sewer samples collected at tens of sites across the globe, and shown to generate estimates consistent with population studies based on sequencing randomly sampled individuals [22].

In this paper we propose new algorithms for haplogroup assignment from short sequencing reads generated from both single individual and mixed DNA samples. There are two types of prior information associated with haplogroups and available from resources such as PhyloTree [26]. First, each haplogroup has one or more complete mtDNA sequences collected from previous studies. These "exemplary" haplotypes can be leveraged to infer the *frequency* of each haplogroup from the short reads. Since many short reads are compatible with more than one of the existing haplotypes, an expectation maximization framework can be used to probabilistically allocate these reads and obtain maximum likelihood estimates for the frequency haplotypes (and hence the haplogroups) in the database. This is the primary approach taken by mixemt – the haplogroups with high estimated frequency are then deemed to be present in the sample, while the haplogroups with low frequency are deemed to be absent.

The second type of information captured by PhyloTree [26] are the mutations associated each branch of the haplogroup tree. Since each haplogroup corresponds to a node in the phylogenetic tree, haplogroups are naturally associated with the set of mutations accumulated on the path from the root to the respective tree node. As an alternative to the frequency estimation approach of mixemt, the short reads can be aligned to the reference mtDNA sequence and

used to call the variants present in the sample. The set of detected variants can then be matched against the sets of mutations associated with each haplogroup, with the best match suggesting the haplogroup composition of the sample.

*A priori* it is unclear which of the two classes of approaches would yield better haplogroup assignment accuracy. The frequency estimation approach critically relies on having a good representation of the haplotype diversity in each haplogroup, and accuracy can be negatively impacted by lack of EM convergence to a global likelihood maximum due to the high similarity between haplogroups. In contrast, the accuracy of the mutation analysis approach depends on the haplogroup tree being annotated with all or nearly all of the shared mutations defining each haplogroup. High frequency of private and back mutations can negatively impact accuracy of this approach.

In this paper we show that an efficient implementation of the mutation analysis approach can match the accuracy of the state-of-the-art frequency based mixemt algorithm while running orders of magnitude faster. Specifically, our implementation of mutation-based analysis uses the SNVQ algorithm from [7] to identify from the short sequencing reads the mtDNA variants present in the sample. The SNVQ algorithm, originally developed for variant calling from RNA-Seq data, has been previosly shown to be robust to large variations in sequencing depth (commonly observed in high-throughput mitogenome sequencing [7]) and allelic fraction (as may be expected for a mixed sample with skewed DNA contributions from different individuals). The set of variants called by SNVQ is then matched to the best set of mutations corresponding to single haplogroups or small collections of haplogroups using the classic Jaccard similarity measure. Exhaustively searching the space of small collections of haplogroups was deemed "computationally infeasible" in [30]. We show that for single individual samples finding the haplogroup with highest Jaccard similarity can be found substantially faster than running mixemt. For two individual mixtures, the pair of haplogroups with highest Jaccard similarity can be identified by exhaustive search within time comparable to that required by mixemt, and orders of magnitude faster when using advanced search algorithms [2].

The rest of the paper is organized as follows. In Section 2 we describe our mutation-based haplogroup assignment algorithms. In Section 3 we present experimental results comparing Jaccard similarity algorithms with mixemt on simulated and real sequencing data from single individuals and two-individual mixtures. Finally, in Section 4 we discuss ongoing and future work.

## 2 Methods

### 2.1 Algorithms for single individual samples

In a preprocessing step, we generate the list of mutations for each haplogroup in PhyloTree (MToolBox [5] already includes a file with these lists). For a given sample, we start by mapping the input paired-end reads to the RSRS human mitogenome reference using hisat2 [13]. We next use SNVQ [7] to identify variants from the mapped data. In our brute-force implementation of the algorithm,

referred to as *JaccardBF*, we compute the Jaccard coefficient between the set of SNVQ variants and *each* list of mutations associated with leaf haplogroups in PhyloTree. The Jaccard coefficient of two sets of variants is defined as the size of the intersection divided by the size of the union. The haplogroup with the highest Jaccard coefficient is then assigned as the haplogroup of the input data.

The brute-force algorithm can be substantially speeded up by using advanced indexing techniques. In Section 3 we report results using the "All-Pair-Binary" algorithm of [2], referred to as *JaccardAPB*, as implemented in the SetSimilaritySearch python library.

### 2.2 Algorithms for two-individual mixtures

High-throughput reads are aligned to RSRS using hisat2 and then SNVQ is used to call variants as above. We experimented with several haplogroup assignment algorithms for two-individual mixtures. In the first, referred to as *JaccardBF2*, the Jaccard coefficient is computed using brute-force search for each leaf haplogroup, and the *top 2* haplogroups are assigned to the mixture. Unfortunately this algorithm has relatively low accuracy, mainly since the haplogroup with the second highest Jaccard similarity is most of the time a haplogroup closely related to the haplogroup with the highest similarity rather than the second haplogroup contributing to the mixture. To resolve this issue we experimented with computing the Jaccard coefficient between the set of SNVQ variants and all *pairs* of leaf haplogroups, with the output consisting of the pair with maximum Jaccard similarity. We implemented both brute-force and "All-Pair-Binary" indexing based implementations of this pair search algorithm, referred to as *JaccardBF_pair* and *JaccardAPB_pair*, respectively.

### 2.3 Algorithms for mixtures of unknown size

When only an upper-bound $k$ is known on the mixture size, the Jaccard coefficient can be computed against sets of mutations generated from unions of up to $k$ leaf haplogroups. For mixtures of up to 2 individuals we report results using the "All-Pair-Binary" indexing based implementation, referred to as *JaccardAPB_1or2*.

## 3 Experimental Results

### 3.1 Datasets

**Real datasets.** We downloaded all WGS datasets used in [26]. Specifically, whole-genome sequencing data for 20 different individuals with distinct haplogroups was downloaded from the 1000 Genomes project (1KGP). The 20 individuals come from two populations: British and Yoruba, with the Yoruba individuals sampled from two different locations (the United Kingdom, and Nigeria, respectively). The haplogroups of 14 of the 20 individuals correspond to leaves

**Table 1.** Human WGS datasets for which ground truth haplogroups are Phylotree leaves. Percentage of mtDNA reads was estimated by mapping reads to the published 1KGP sequence, except for the datasets marked with "*" for which there is no 1KGP sequence and mapping was done against the RSRS reference.

| Sample ID | Run ID | #Read pairs | #mtDNA pairs | %mtDNA | Haplogroup |
|---|---|---|---|---|---|
| HG00096 | SRR062634 | 24,476,109 | 43,370 | 0.177 | H16a1 |
| HG00097 | SRR741384 | 68,617,747 | 112,039 | 0.163 | T2f1a1 |
| HG00098* | ERR050087 | 20,892,714 | 37,602 | 0.180 | J1b1a1a |
| HG00100 | ERR156632 | 19,119,986 | 39,169 | 0.204 | X2b8 |
| HG00101 | ERR229776 | 111,486,484 | 169,840 | 0.152 | J1c3g |
| HG00102 | ERR229775 | 109,055,650 | 217,187 | 0.199 | H58a |
| HG00103 | SRR062640 | 24,054,672 | 48,912 | 0.203 | J1c3b2 |
| HG00104* | SRR707166 | 58,982,989 | 94,242 | 0.159 | U5a1b1g |
| NA19093 | ERR229810 | 98,728,262 | 234,170 | 0.237 | L2a1c5 |
| NA19096 | SRR741406 | 55,861,712 | 131,587 | 0.235 | L2a1c3b2 |
| NA19099 | ERR001345 | 7,427,776 | 16,038 | 0.215 | L2a1m1a |
| NA19102 | SRR788622 | 15,134,619 | 28,239 | 0.186 | L2a1a1 |
| NA19107 | ERR239591 | 9,217,863 | 13,297 | 0.144 | L3b2a |
| NA19108 | ERR034534 | 65,721,104 | 3,959 | 0.006 | L2e1a |

**Table 2.** Human WGS datasets for which ground truth haplogroups are Phylotree internal nodes.

| Sample ID | Run ID | #Read pairs | #mtDNA pairs | %mtDNA | Haplogroup |
|---|---|---|---|---|---|
| HG00099 | SRR741412 | 57,222,221 | 102,968 | 0.179 | H1ae |
| HG00106 | ERR162876 | 24,328,397 | 50,635 | 0.208 | J2b1a |
| NA19092 | SRR189830 | 125,888,789 | 337,350 | 0.268 | L3e2a1b |
| NA19095 | SRR741381 | 65,174,483 | 101,118 | 0.155 | L2a1a2 |
| NA19098 | SRR493234 | 40,446,917 | 85,658 | 0.211 | L3b1a |
| NA19113 | SRR768183 | 48,428,152 | 62,412 | 0.128 | L3e2b |

nodes in PhyloTree, while the haplogroups of the other 6 correspond to internal nodes. Accession numbers, basic sequencing statistics, and ground truth haplogroups for the 20 datasets are given in Tables 1 and 2.

**Synthetic datasets.** For the synthetic datasets, we simulated reads using wgsim [17] based on exemplary sequences associated with leaf haplogroups in PhyloTree [27]. Of the 2,897 leaf haplogroups, 423 haplogroups have only one associated sequence, 2,454 haplogroups have two sequences, and 20 haplogroups have three or more sequences. For single individual experiments, we generated two sets of 10,000 simulated read pairs for each haplogroup, using different exemplary sequences as wgsim reference whenever possible, i.e., for all but the 423 haplogroups with a single associated sequence, for which the sole sequence was used to generate both sets of wgsim reads. For mixture experiments we

**Table 3.** Experimental results on human WGS datasets for which the ground truth haplogroups are Phylotree leaves.

| Sample ID | Ground truth | mixemt | | JaccardBF | |
|---|---|---|---|---|---|
| | | Haplogroup | Time | Haplogroup | Time |
| HG00096 | H16a1 | **H16a1** | 8,343 | **H16a1** | 264 |
| HG00097 | T2f1a1 | **T2f1a1** | 897 | **T2f1a1** | 546 |
| HG00098 | J1b1a1a | **J1b1a1a** | 16,423 | **J1b1a1a** | 275 |
| HG00100 | X2b8 | **X2b8** | 12,477 | **X2b8** | 258 |
| HG00101 | J1c3g | **J1c3g** | 61,091 | **J1c3g** | 2,523 |
| HG00102 | H58a | **H58a** | 66,350 | **H58a** | 5,343 |
| HG00103 | J1c3b2 | **J1c3b2** | 29,733 | **J1c3b2** | 1,192 |
| HG00104 | U5a1b1g | **U5a1b1g** | 27,067 | **U5a1b1g** | 5,628 |
| NA19093 | L2a1c5 | **L2a1c5** | 59,107 | **L2a1c5** | 4,345 |
| NA19096 | L2a1c3b2 | **L2a1c3b2** | 44,338 | **L2a1c3b2** | 1,054 |
| NA19099 | L2a1m1a | **L2a1m1a** | 14,515 | **L2a1m1a** | 67 |
| NA19102 | L2a1a1 | **L2a1a1** | 13,642 | **L2a1a1** | 231 |
| NA19107 | L3b2a | **L3b2a** | 8,607 | **L3b2a** | 166 |
| NA19108 | L2e1a | **L2e1a** | 1,423 | **L2e1a** | 1,049 |

similarly generated two groups of 2,897 two-individual mixtures by pairing each haplogroup with a second haplogroup selected uniformly at random from the remaining ones. Within each group, the reads were generated using wgsim and the first and the second PhyloTree sequence, respectively, except for haplogroups with a single PhyloTree sequence in which the sole sequence was used to generate both sets of wgsim reads. For each pair of haplogroups we generated 10,000 read pairs, with an equal number of read pairs from each haplogroup. We used default wgsim parameters for simulating reads, in particular the sequencing error rate was 1% and the mutation rate 0.001.

### 3.2 Results on real datasets

Tables 3 and 4 give the results obtained by mixemt and JaccardBF on the real datasets consisting of PhyloTree leaf and internal haplogroups, respectively. Both algorithms infer the expected haplogroup when the ground truth is a leaf PhyloTree node. For the six datasets in which the ground truth is an internal node of PhyloTree mixemt always infers the haplogroup correctly, while JaccardBF always infers a leaf haplogroup in the subtree rooted at the ground truth haplogroup. Despite using brute-force search to identify the best matching haplogroup, JacardBF is substantially faster (one order of magnitude or more) than mixemt.

### 3.3 Accuracy results for single individual synthetic datasets

The above results on real datasets already suggest that the mitochondrial haplogroup can be accurately inferred from WGS data. For a more comprehen-

**Table 4.** Experimental results on human WGS datasets for which the ground truth haplogroups are Phylotree internal nodes.

| Sample ID | Ground truth | mixemt | | JaccardBF | |
|---|---|---|---|---|---|
| | | Haplogroup | Time | Haplogroup | Time |
| HG00099 | H1ae | **H1ae** | 40,733 | **H1ae1** | 1,820 |
| HG00106 | J2b1a | **J2b1a** | 24,614 | **J2b1a5** | 1,040 |
| NA19092 | L3e2a1b | **L3e2a1b** | 137,218 | **L3e2a1b1** | 6,610 |
| NA19095 | L2a1a2 | **L2a1a2** | 94,921 | **L2a1a2b** | 1,529 |
| NA19098 | L3b1a | **L3b1a** | 46,110 | **L3b1a11** | 650 |
| NA19113 | L3e2b | **L3e2b** | 62,643 | **L3e2b3** | 822 |

**Table 5.** Experimental results on synthetic single individual datasets generated from the 2,897 leaf haplogroups in Phylotree.

| | mixemt | | JaccardBF | | JaccardAPB | |
|---|---|---|---|---|---|---|
| | Acc. | Avg. time | Acc. | Avg. time | Acc. | Avg. time |
| Group1 | 99.275 | 7,251.490 | 99.379 | 83.780 | 99.413 | 0.041 |
| Group2 | 99.448 | 7,185.373 | 99.517 | 81.428 | 99.620 | 0.043 |
| Mean | 99.361 | 7,218.432 | 99.448 | 82.604 | 99.517 | 0.042 |
| Std. Dev. | 0.122 | 46.752 | 0.098 | 1.663 | 0.146 | 0.001 |

sive evaluation we simulated reads using exemplary sequences from all leaf haplogroups in PhyloTree. Table 5 gives the results of this comparison. Both mixem and Jaccard algorithms achieve over 99% accuracy on simulated datasets. As for real datasets, JaccardBF is more than one order of magnitude faster than mixemt. The indexing approach implemented in JaccardABP further reduces the running time needed to find the best matching haplogroup with no loss in accuracy.

### 3.4 Accuracy results for two-individual synthetic mixtures

Table 6 gives experimental results on two-individual synthetic mixtures generated as described in Section 3.1. In these experiments we assume that it is *a priori* known that the mixture consists of two different haplogroups. Consistent with this assumption, the mixemt prediction is taken to be the two haplogroups with highest estimated frequencies (regardless of the magnitude of the estimated frequencies). Under this model, the accuracy of mixemt remains high but is slightly lower for mixtures than for single haplogroup samples, with an overall mean accuracy of 98.792% compared to 99.361%. JaccardBF2, which returns the two haplogroups with highest Jaccard similarity to the set of mutations called by SNVQ, performs quite poorly, with a mean accuracy of only 22.765%. The JaccardBF_pair algorithm, which returns the pair of haplogroups whose union has the highest Jaccard similarity to the set of mutations called by SNVQ, nearly matches the accuracy of mixemt (with a mean accuracy of 98.398%) with a lower running time. The running time is drastically reduced by indexing the

**Table 6.** Experimental results on synthetic two-individual mixtures generated from the 2,897 leaf haplogroups in Phylotree.

| | mixemt | | JaccardBF2 | | JaccardBF_pair | | JaccardAPB_pair | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | Avg. time | Acc. | Avg. time | Acc. | Avg. time | Acc. | Avg. time |
| Group1 | 98.619 | 4,890.769 | 22.540 | 83.116 | 98.343 | 1,224.589 | 97.480 | 2.101 |
| Group2 | 98.964 | 5,273.326 | 22.989 | 80.440 | 98.452 | 1,484.743 | 98.171 | 2.315 |
| Mean | 98.792 | 5,082.048 | 22.765 | 81.778 | 98.398 | 1,354.666 | 97.825 | 2.208 |
| Std. Dev. | 0.244 | 270.509 | 0.317 | 1.893 | 0.077 | 183.957 | 0.488 | 0.151 |

haplogroups for Jaccard similarity searches, although the predefined threshold required for indexing (0.8 in our experiments) does lead to a small additional loss of accuracy (mean overall accuracy of 97.825% for JaccardAPB_pair).

### 3.5 Accuracy results for unknown mixture size

In practical forensics applications there are scenarios in which the number of individuals contributing to a DNA mixture is not *a priori* known. In this case, joint inference of the number of individuals *and* their haplogroups is required. Although mitochondrial haplogroup inference with unknown number of contributors remains a direction of future research, in this section we report experimental results for the most restricted (but still practically relevant) such scenario, in which a mixture is *a priori* known to contain *at most two* haplogroups. Specifically, the 2,897 single individual synthetic datasets analyzed in Section 3.3 and the 2,897 two-individual synthetic datasets analyzed in Section 3.4 were reanalyzed using several joint inference algorithms. For mixemt, the joint inference was performed by using a 5% cutoff on the estimated haplogroup frequencies, while for JaccardAPB_1or2 the joint inference was performed by matching the set of SNVQ variants to the set of one or two haplogroups that has the highest Jaccard similarity. Table 7 reports the accuracy and runtime of the two methods. Overall, mixemt achieves a mean accuracy of 93.398%, with most of the errors due to the incorrect estimate of the number of individuals in the two-individual mixtures. In contrast, most of the JaccardAPB_1or2 errors are due to mis-classification of single individual samples as mixtures. Overall, JaccardAPB_1or2 achieves a mean accuracy of 96.538%.

## 4 Conclusions

In this paper we introduced efficient algorithms for mitochondrial haplogroup inference based on Jaccard similarity between variants called from high-throughput sequencing data and mutations annotated in public databases such as PhyloTree. Experimental results on real and simulated datasets show an accuracy comparable to that of previous state-of-the-art methods based on haplogroup frequency estimation for both single-individual samples and two-individual mixtures, with a much lower running time.

**Table 7.** Experimental results for joint inference of mixture size and haplogroup composition.

| | mixemt | | JaccardAPB_1or2 | |
|---|---|---|---|---|
| | Acc. | Avg. time | Acc. | Avg. time |
| Group1 Singles | 99.275 | 7,251.490 | 94.028 | 1.4794 |
| Group2 Singles | 99.448 | 7,185.373 | 96.548 | 2.098 |
| Group1 Pairs | 83.914 | 4,890.769 | 97.376 | 1.468 |
| Group2 Pairs | 90.956 | 5,273.326 | 98.205 | 2.244 |
| Mean | 93.398 | 6,150.240 | 96.539 | 1.822 |
| Std. Dev. | 7.462 | 1,243.583 | 1.806 | 0.407 |

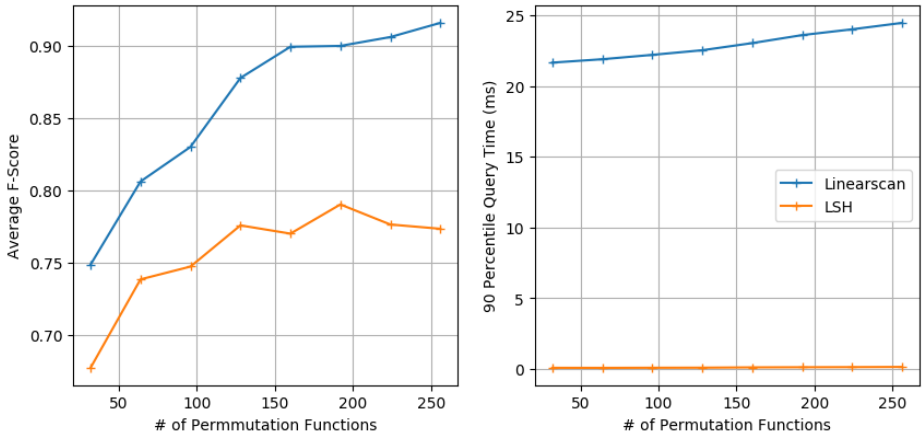


**Fig. 2.** Comparison of accuracy and running time needed to compute all sets with a Jaccard coefficient greater than 0.9 using MinHash sketches with varying number of hash functions from 2,897 randomly generated sets of average size 44.

In ongoing work we are exploring methods for haplogroup inference of more complex DNA mixtures. Specifically, we are seeking to scale the mutation analysis approach to larger haplogroup mixtures by employing probabilistic techniques such as MinHash sketches and indexing for locality sensitive hashing (LSH) [23]. Implementations such as MinHashLSH [34] can be used to generate all haplogroups with a Jaccard similarity exceeding a given user threshold in sublinear time, resulting in dramatic speed-ups. However, MinHashLSH is an approximate algorithm, which may miss some of the haplogroups with high Jaccard similarity and may also generate false positives. The accuracy and runtime of MinHashLSH depend among other parameters on the number of hash functions, and the user can generally achieve higher precision and recall at the cost of increased running time (Figure 2). Finally, we are exploring hybrid methods that combine mutation analysis with highly scalable frequency estimation algorithms such as IsoEM [21, 19].

## References


1. Amorim, A., Fernandes, T., Taveira, N.: Mitochondrial DNA in human identification: a review. PeerJ Preprints **7**, e27500v1 (2019)
2. Bayardo, R.J., Ma, Y., Srikant, R.: Scaling up all pairs similarity search. In: Proceedings of the 16th international conference on World Wide Web. pp. 131–140 (2007)
3. Blau, S., Catelli, L., Garrone, F., Hartman, D., Romanini, C., Romero, M., Vullo, C.: The contributions of anthropology and mitochondrial DNA analysis to the identification of the human skeletal remains of the australian outlaw Edward 'Ned' Kelly. Forensic science international **240**, e11–e21 (2014)
4. Budowle, B., Allard, M.W., Wilson, M.R., Chakraborty, R.: Forensics and mitochondrial DNA: applications, debates, and foundations. Annual review of genomics and human genetics **4**(1), 119–141 (2003)
5. Calabrese, C., Simone, D., Diroma, M.A., Santorsola, M., Gutta, C., Gasparre, G., Picardi, E., Pesole, G., Attimonelli, M.: Mtoolbox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics **30**(21), 3115–3117 (2014)
6. Chinnery, P.F., Howell, N., Andrews, R.M., Turnbull, D.M.: Clinical mitochondrial genetics. Journal of medical genetics **36**(6), 425–436 (1999)
7. Duitama, J., Srivastava, P.K., Măndoiu, I.I.: Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data. BMC genomics **13**(2), S6 (2012)
8. Fan, L., Yao, Y.G.: MitoTool: a web server for the analysis and retrieval of human mitochondrial DNA sequence variations. Mitochondrion **11**(2), 351–356 (2011)
9. Hahn, C., Bachmann, L., Chevreux, B.: Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic acids research **41**(13), e129–e129 (2013)
10. Hu, N., Cong, B., Li, S., Ma, C., Fu, L., Zhang, X.: Current developments in forensic interpretation of mixed DNA samples. Biomedical reports **2**(3), 309–316 (2014)
11. Ishiya, K., Ueda, S.: Mitosuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. PeerJ **5**, e3406 (2017)
12. Johns, D.R.: Mitochondrial DNA and disease. New England journal of medicine **333**(10), 638–644 (1995)
13. Kim, D., Langmead, B., Salzberg, S.: HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes (2017)
14. Kivisild, T.: Maternal ancestry and population history from whole mitochondrial genomes. Investigative genetics **6**(1), 3 (2015)
15. Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., Kronenberg, F.: HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Human mutation **32**(1), 25–32 (2011)
16. Lee, H.Y., Song, I., Ha, E., Cho, S.B., Yang, W.I., Shin, K.J.: mtDNAmanager: a web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. BMC bioinformatics **9**(1), 483 (2008)
17. Li, H.: wgsim-read simulator for next generation sequencing. Github Repository (2011)
18. Luo, S., Valencia, C.A., Zhang, J., Lee, N.C., Slone, J., Gui, B., Wang, X., Li, Z., Dell, S., Brown, J., et al.: Biparental inheritance of mitochondrial DNA in humans. Proceedings of the National Academy of Sciences **115**(51), 13039–13044 (2018)

19. Mandric, I., Temate-Tiagueu, Y., Shcheglova, T., Al Seesi, S., Zelikovsky, A., Măndoiu, I.I.: Fast bootstrapping-based estimation of confidence intervals of expression levels and differential expression from rna-seq data. Bioinformatics **33**(20), 3302–3304 (2017)
20. Melton, T.W., Holland, C.W., Holland, M.D.: Forensic mitochondrial DNA analysis: Current practice and future potential. Forensic science review **24 2**, 101–22 (2012)
21. Nicolae, M., Mangul, S., Măndoiu, I.I., Zelikovsky, A.: Estimation of alternative splicing isoform frequencies from rna-seq data. Algorithms for molecular biology **6**(1), 9 (2011)
22. Pipek, O.A., Medgyes-Horváth, A., Dobos, L., Stéger, J., Szalai-Gindl, J., Visontai, D., Kaas, R.S., Koopmans, M., Hendriksen, R.S., Aarestrup, F.M., et al.: Worldwide human mitochondrial haplogroup distribution from urban sewage. Scientific reports **9**(1), 1–9 (2019)
23. Rajaraman, A., Ullman, J.D.: Mining of massive datasets. Cambridge University Press (2011)
24. Smieszek, S., Mitchell, S.L., Farber-Eger, E.H., Veatch, O.J., Wheeler, N.R., Goodloe, R.J., Wells, Q.S., Murdock, D.G., Crawford, D.C.: Hi-mc: a novel method for high-throughput mitochondrial haplogroup classification. PeerJ **6**, e5149 (2018)
25. Van Oven, M.: Phylotree. https://www.phylotree.org/, accessed January 7, 2020
26. Van Oven, M.: PhyloTree Build 17: Growing the human mitochondrial DNA tree. Forensic Science International: Genetics Supplement Series **5**, e392–e394 (2015)
27. Van Oven, M., Kayser, M.: Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Human mutation **30**(2), E386–E394 (2009)
28. Vellarikkal, S.K., Dhiman, H., Joshi, K., Hasija, Y., Sivasubbu, S., Scaria, V.: mit-o-matic: A comprehensive computational pipeline for clinical evaluation of mitochondrial variations from next-generation sequencing datasets. Human mutation **36**(4), 419–424 (2015)
29. Vianello, D., Sevini, F., Castellani, G., Lomartire, L., Capri, M., Franceschi, C.: HAPLOFIND: A new method for high-throughput mtDNA haplogroup assignment. Human mutation **34**(9), 1189–1194 (2013)
30. Vohr, S.H., Gordon, R., Eizenga, J.M., Erlich, H.A., Calloway, C.D., Green, R.E.: A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. Forensic Science International: Genetics **30**, 93–105 (2017)
31. Wallace, D.C., Chalkia, D.: Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harbor perspectives in biology **5**(11), a021220 (2013)
32. Weissensteiner, H., Forer, L., Fuchsberger, C., Schöpf, B., Kloss-Brandstätter, A., Specht, G., Kronenberg, F., Schönherr, S.: mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic acids research **44**(W1), W64–W69 (2016)
33. Weissensteiner, H., Pacher, D., Kloss-Brandstätter, A., Forer, L., Specht, G., Bandelt, H.J., Kronenberg, F., Salas, A., Schönherr, S.: Haplogrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic acids research **44**(W1), W58–W63 (2016)
34. Zhu, E.E.: Minhash lsh. available online at http://ekzhu.com/datasketch/index.html (2019)