MixtureAceMT™ Technical Note

Identifying Haplogroups and Their Proportions in Mixed mtDNA

November 2025

1. Overview

MixtureAceMT estimates the number and identity of haplogroups and their proportions in mixed mtDNA samples using a probabilistic method called expectation maximization (1-3). Here we describe the general principles of the method and illustrate two aspects using a step-by-step walk-through using toy data. First, we illustrate how individual reads are scored for their support of haplogroups using a simple one read and two haplogroup scenario. Second, we illustrate how haplogroup proportion ratios are calculated using a two haplogroup and four read scenario.

1.1. Scoring Read Support to Haplogroups

MixtureAceMT calculates a log likelihood that each read supports each possible haplogroup. This is accomplished by scoring how many and which haplogroup-diagnostic variants (if any) are present in each read.

Each haplogroup (***h***) has a set of expected diagnostic variants as specified in PhyloTree (4):

$$V_h = \{diagnostic\ variants\}$$

Each read (***r***) has a set of observed variants:

$$O_r = \{Variants\ in\ \boldsymbol{r}\}$$

Reads containing diagnostic variants can provide support for the presence of one or more haplogroups in the sample.

$$\log Pr(O_r|h)$$

Table 1. Q score derived strengths of evidence for several cases.

Case	Evidence	Calculation
The read contains a variant that is expected in ***h***	Positive	log(1-ε)
An expected variant is missing	Weakly negative	log(0.5+ε)
Read contains a diagnostic variant not expected in ***h***	Negative	log(ε)
Read contains a private variant	Neutral	log(0.1)

Individual reads can be considered to represent sub-haplotypes of the entire mtDNA haplotype. MixtureAceMT assigns likelihoods (probabilities) that express the partial support individual sub-haplotypes provide for a given full haplotype.

1.2. Scoring Haplogroup Probabilities

Per-read likelihoods are combined into a haplogroup probability.

$$\log Pr_h = \sum_r \mathrm{Z}_{rh} \cdot \log Pr(O_r|h)$$

Where:

$$\mathrm{Z}_{rh} = membership\ weightof\ read\ \boldsymbol{r}\ in\ mixture\ component\ \boldsymbol{h}$$

Weights are assigned by expectation maximization (EM). Several aspects of the EM process ensure that read-weighted results can be haplotype likelihoods are built from proper probabilities.

1.3. Complete vs Incomplete Data

The data available to the examiner are individual reads arising from the various contributors to a mixture. If reads somehow had some easily observable feature such as color or shape corresponding to the contributor from which they arose, then a maximum likelihood solution to the deconvolution would be straight forward. The maximum likelihood estimate of the true proportion of reads from >2 haplogroups (categories) is simply the ratio of the number of reads from the ith category divided by the total number of reads across all categories:

$$\hat{p}_i = \frac{k_i}{n}$$

This represents a complete data case in that the values of all relevant variables are known (i.e., the contributor origin of each read, and their counts).

The data arising from mtDNA mixtures represents an incomplete data set in the sense that the contributor haplogroups present in the mixture and the contributor origin of individual reads are both unknown. MixtureAceMT jointly solves the identity of the contributing haplogroups and their proportion ratios using EM. The EM process "completes" the data in each iteration so that incrementally better estimates of latent parameter values can be generated.

2. Calculating Haplogroup Likelihoods

Here a simple example is given with numeric values to illustrate how haplogroup likelihoods are calculated. The example uses only three mtDNA variant positions from PhyloTree (4), all assumed to be covered by the same read.

2.1. Assumptions

- Three positions are considered: 73A, 146C, and 263A (reference bases shown).
- Base quality at all positions is Q = 20; translating to a base-call error rate ε = 0.01
- Match/mismatch strengths of evidence are derived from Q scores:
  - Pr(match) = 1 – ε = 0.99; ln(0.99) ≈ -0.0101
  - Pr(mismatch) = ε = 0.01; ln(0.01) ≈ -4.6052
- Two candidate haplogroups are assumed, with each containing a characteristic set of expected variants:
  - Candidate haplogroup A expected variants V_{hapA}:
    - 73 A>G
    - 146C (ref)
    - 263 A>G
  - Candidate haplogroup B expected variants V_{hapB}:
    - 73A (ref)
    - 146C (ref)
    - 263A>G
- A read is observed to cover all three positions with observed bases:
  - 73G, 146C, 263G

We will now calculate the support this hypothetical read provides for each of the two candidate haplogroups.

2.2. The Likelihood that The Read Supports Haplogroup A

The term Pr(read | hapA) refers to the probability of observing a given read under the assumption that it arose from a given haplogroup (haplogroup A in this case). An individual read can be considered to cover a sub-haplotype of the entire mtDNA haplotype.

MixtureAceMT assigns likelihoods (probabilities) that individual sub-haplotypes support or are consistent with a given full haplotype. The probability that the hypothetical read arose from a given haplogroup is calculated by considering the joint probability of all the variants observed in a read. This probability is calculated as the product of the individual variant probabilities.

Table 2. Q-score derived probabilities of match and mismatch relative to expected base identities expected from haplogroup A.

Position	Case	Raw Probability	Log-Likelihood
73	Match	0.99	-0.0101
146	Match	0.99	-0.0101
263	Match	0.99	-0.0101

$$\Pr(read|hapA) = 0.99 \times 0.99 \times 0.99 = 0.9703$$

$$\ln L_{hapA} = (-0.0101) + (-0.0101) + (-0.0101) = -0.0302$$

2.3. The Likelihood that The Read Supports Haplogroup A

The equivalent calculation is performed assuming the read derived from haplogroup B.

Table 3. Q-score derived probabilities of match and mismatch relative to expected base identities expected from haplogroup B.

Position	Case	Raw Probability	Log-Likelihood
73	Mismatch	0.01	-4.6052
146	Match	0.99	-0.0101
263	Match	0.99	-0.0101

$$\Pr(read|hapB) = 0.01 \times 0.99 \times 0.99 = 0.009801$$

$$\ln L_{hapA} = (-4.6052) + (-0.0101) + (-0.0101) = -4.6253$$

2.4. Compare Candidate Haplogroup Probabilities and Likelihoods

$$\frac{\Pr(read \mid hapA)}{\Pr(read \mid hapB)} = \frac{0.9703}{0.009801} \approx 99$$

$$\ln L_{hapA} - \ln L_{hapB} = -0.0302 - (-4.6523) \approx 4.5951$$

Converting back to non-log scale: $e^{4.5951} \approx 99$

By either probability or log-likelihood it can be said that the read is about 99x more likely under the hypothesis of haplogroup A as compared to haplogroup B.

The general case of an arbitrary number of haplogroup candidates is managed in a matrix with ***h*** columns representing PhyloTree haplogroups, and ***j*** rows representing reads from the sample. Each cell of the **h x j** matrix contains the support value from that read for that haplogroup:

$$E_{jh} = \log \Pr\left(read_j \,\middle|\, hap_h\right)$$

In this single read example, $E_{read,\ hapA} \approx -0.0302$ and $E_{read,\ hapB} \approx -4.6253$.

3. Expectation Maximization (EM)

Maximum likelihood estimates of parameter values in the incomplete data case can be performed using EM which is an iterative two-step process that incrementally estimates the maximum likelihood of missing parameter values. Here, the proportion ratios of the contributor haplogroups are estimated from the read data. At the same time, the identity and number of haplogroups present are jointly estimated.

In each iteration the data set is completed by estimating the expected read counts given the current estimates of the latent proportion parameter values. In the initial round, the parameter values are simply guessed. These values are incrementally refined in subsequent rounds. In the second iteration, weighted probabilities for each hypothesis (haplogroup) are created from the likelihoods of the first step. This process is repeated until convergence around a set of parameter values is reached. That is, a stopping rule is implemented such that iteration is terminated when the incremental change in estimated parameter values falls below a threshold.

3.1. Illustration Setup

An illustration follows using four reads that cover two diagnostic variant positions. Only two haplogroups are under consideration: haplogroup A with expected variants (GG) and haplogroup B with expected variants (AG) (**Table *4***).

Table 4. Scenario used for illustration.

Position	Ref Allele	HapA	HapB	Q
73	A	G	A	20
263	A	G	G	20

As before, if the observed base matches or mismatches the haplogroup's expected base the probability is derived from Q scores which are taken to be equal to 20:

- Pr(match) = 1 – ε = 0.99; ln(0.99) ≈ -0.0101
- Pr(mismatch) = ε = 0.01; ln(0.01) ≈ -4.6052

Table 5. Four reads covering positions 73 and 263.

Read No.	Position		Read Probabilities and Likelihoods	
	73	263	Pr(read | hapA)	Pr(read | hapB)
1	G	G	Pr(read = GG | hapA) = 0.9801 Log L_A(GG) = -0.0201	Pr(read = GG | hapB) = 0.0099 Log L_B(GG) = -4.6052
2	G	G	Pr(read = GG | hapA) = 0.9801 Log L_A(GG) = -0.0201	Pr(read = GG | hapB) = 0.0099 Log L_B(GG) = -4.6052
3	A	G	Pr(read = AG | hapA) = 0.0099 Log L_A(AG) = -4.6052	Pr(read = AG | hapB) = 0.9801 Log L_B(AG) = -0.0201
4	G	G	Pr(read = GG | hapA) = 0.9801 Log L_A(GG) = -0.0201	Pr(read = GG | hapB) = 0.0099 Log L_B(GG) = -4.6152

Note that GG reads are 99x more likely under a hypothesis that they arose from haplogroup A than haplogroup B, and AG reads are 99x more likely under haplogroup B than haplogroup A.

3.2. EM First Iteration

The unknown mixture proportions are estimated by iteration of the EM algorithm. The proportion estimates are incrementally improved over the iterations. Once the process is started, each iteration updates the mixture proportions from the previous round. However, at the initiation, data-driven mixture proportion values are not available. This dilemma is solved by using a (usually uninformative) guess for the initial values. For this illustration, only two haplogroups are considered so an uninformative distribution of guesses is a uniform distribution:

$$\theta_A^0 = \theta_B^0 = 0.5$$

Where Θ represents unknown proportion values, and the superscript "0" indicates the initial value, and the subscript indicates the haplogroup.

We are interested in the conditional probability of each haplogroup, after having observed the read data. This is a posterior probability calculated using Bayes formula weighted by proportion ratio:

$$\mathrm{Z}_{j,A} = \Pr(Hap\ A \mid read_j) = \frac{\theta_A^0 \cdot \Pr\,(read_j|hap\ A)}{\theta_A^0 \cdot \Pr\left(read_j\middle|hap\ A\right) + \theta_B^0 \cdot \Pr\,(read_j|hap\ B)}$$

$$\mathrm{Z}_{j,B} = \Pr(Hap\ B \mid read_j) = \frac{\theta_B^0 \cdot \Pr\,(read_j|hap\ B)}{\theta_A^0 \cdot \Pr\left(read_j\middle|hap\ A\right) + \theta_B^0 \cdot \Pr\,(read_j|hap\ B)}$$

As proper proportions, the following relationship holds:

$$\mathrm{Z}_{jB} = 1 - \mathrm{Z}_{jA}$$

Applying numbers to variables for a GG read:

$$\mathrm{Z}_{(GG),A} = \frac{0.5 \cdot 0.9801}{0.5 \cdot 0.9801 + 0.5 \cdot 0.0099} = 0.99$$

and

$$\mathrm{Z}_{(GG),B} = 0.01$$

Applying numbers to variables for an AG read:

$$\mathrm{Z}_{(AG),A} = \frac{0.5 \cdot 0.0099}{0.5 \cdot 0.0099 + 0.5 \cdot 0.9801} = 0.01$$

and

$$\mathrm{Z}_{(AG),B} = 0.99$$

Where:

$\mathrm{Z}_{j,g} = conditional\ probabilty\ of\ haplogroup\ \boldsymbol{g}\ given\ observation\ of\ read\ j$

$\theta_g^n = the\ latent\ proportion\ parameter\ value\ of\ haplogroup\ g\ in\ iteration\ n$

Table 6. Summary of calculations for first iteration of the four read, two haplogroup scenario.

Read	Sub-haplotype	$\mathbf{Z}_{j,A}$	$\mathbf{Z}_{j,B}$	Expected Number of Reads (Expectation Step)		Updated Parameters (Maximization Step)	
				$\boldsymbol{N_A^1}$	$\boldsymbol{N_B^1}$	$\boldsymbol{\theta_A^1}$	$\boldsymbol{\theta_B^1}$
1	GG	0.99	0.01	$\sum \mathrm{Z}_{j,A} =$ 2.98	$\sum \mathrm{Z}_{j,B} =$ 1.02	$\frac{N_A^1}{4} =$ 0.745	$\frac{N_B^1}{4} =$ 0.225
2	GG	0.99	0.01				
3	AG	0.01	0.99				
4	GG	0.99	0.01				

At the end of the initial EM round the estimated proportion parameter values are updated from the initial guesses:

$$\theta_A^0 = 0.5 \; updated \; to \; \theta_A^1 = 0.745$$

$$\theta_B^0 = 0.5 \; updated \; to \; \theta_B^1 = 0.225$$

3.3. EM Second Iteration

The second iteration begins with the maximum likelihoods of the parameter estimates from the first iteration:

$$\theta_A^1 = 0.745$$

$$\theta_B^1 = 0.225$$

Table 7. Summary of calculations for second iteration of the four read, two haplogroup scenario.

Read	Sub-haplotype	$\mathbf{Z}_{j,A}$	$\mathbf{Z}_{j,B}$	Expected Number of Reads		Updated Parameters (Maximization)	
				$\boldsymbol{N_A^2}$	$\boldsymbol{N_B^2}$	$\boldsymbol{\theta_A^2}$	$\boldsymbol{\theta_B^2}$
1	GG	0.9966	0.0034	$\sum \mathrm{Z}_{j,A} =$ 3.0185	$\sum \mathrm{Z}_{j,B} =$ 0.9815	$\frac{N_A^1}{4} =$ 0.755	$\frac{N_B^1}{4} =$ 0.245
2	GG	0.9966	0.0034				
3	AG	0.0287	0.9713				
4	GG	0.9966	0.0034				

At the end of the second EM round the estimated proportion parameter values are updated from the first iteration:

$$\theta_A^1 = 0.745 \; updated \; to \; \theta_A^2 = 0.755$$

$$\theta_B^1 = 0.225 \; updated \; to \; \theta_B^2 = 0.245$$

4. Conclusion

The simple illustration presented here with only two pre-specified haplogroups avoids the issue of haplogroup discovery. In the more general case, the haplogroups present in the mixture are discovered. Because of a lack of prior information, all the haplogroups in PhyloTree (>5,000) are initially considered as potential contributors to the mixture. The starting proportion values are drawn from the Dirichlet distribution, thereby creating for each read a vector of proportion values with dimension equal to the number of haplogroups. MixtureAceMT is applied to this more complex scenario. As the proportion ratios are refined, some haplogroups emerge with substantial support while most haplogroups receive little to no support. The denominator in the weighted Bayes formula is summed over all haplogroups to render the expression Z_{jg} a proper probability.

$$\mathrm{Z}_{jh} = \frac{\theta_h \cdot \Pr\,(read_j|hap\ h)}{\sum_{h' \in H} \sum_{j \in N} \theta_{h'} \cdot \Pr\,(read_j|hap\ h')}$$

References

1. Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat Biotechnol. 2008;26(8):897-9.

2. Kessner D, Turner TL, Novembre J. Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. Mol Biol Evol. 2013;30(5):1145-58.

3. Vohr SH, Gordon R, Eizenga JM, Erlich HA, Calloway CD, Green RE. A phylogenetic approach for haplotype analysis of sequence data from complex mitochondrial mixtures. Forensic Sci Int Genet. 2017;30:93-105.

4. van Oven M, Kayser M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 2009;30(2):E386-94.