# Deconvolution of Mixed Mitochondrial Samples

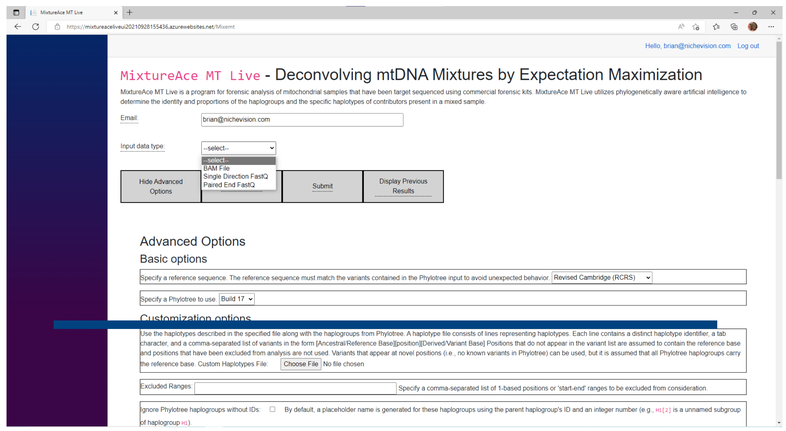
Brian Young PhD
NicheVision Forensics, Syracuse University, Florida International University
brian@nichevision.com

## Abstract

Mitochondrial (mt) DNA holds the potential to address two important current needs in forensic DNA analysis: 1) deriving familial and genetic genealogy evidence from mixed samples, and 2) estimating the number of contributors to DNA mixtures. Both needs can be addressed by exploiting data generated by massively parallel sequencing (MPS) in a phylogenetically aware probabilistic approach. Individual sequencing reads in raw data (FASTQ) are classified by haplogroup based on phylogenetically informative SNPs. Maximum likelihoods (ML) of the haplogroups present in the mixture and their proportions are co-estimated by expectation maximization. Prior to the classification stage, read data are filtered for NUMTs (mt segments transferred to the nuclear genome), sequencing error is corrected, and primer sequences are trimmed (if present). Reads assigned to each contributor are assembled in BAM file format and variant-called to generate (partial or complete) contributor haplotypes. The estimated number of haplogroups provides an empirical estimate of the number of non-maternally related individuals in the mixture. The haplotypes further refine this estimate and can be used in familial investigation.

## MixtureAceMT™ Software Interface

MixtureAceMT™ software uses a web browser interface to receive data and return results. The user can upload single-direction or paired-end FASTQ files, or BAM files. Paired-end FASTQ files are corrected for sequencing error using the FLASH algorithm. Results are returned to the browser interface, and in Excel format via email to the user account.




## References

HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. Nucleic Acids Res. 2016, 44,W58–W63.

Majoc T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genomic assemblies. Bioinformatics 2011 Vol. 27 Issue 21 Pages 2957-63.

McElhoe JA, Addesso A, Young B, Holland MM. A new tool for probabilistic assessment of MPS data associated with mtDNA mixtures. Genes 2024 DOI: 10.3390/genes15020194.

Ramos A, Santos C, Alvarez L, Nogue R, Aluja MP. Human mitochondrial DNA complete amplification and sequencing: A new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification. Electrophoresis 2009, 30, 1587–1593.

Vohr SH, Gordon R, Eizenga JM, Erlich HA, Calloway CD, Green RE. Forensic Sci. Intl. Genetics 30 (2017) 93-105.

## Haplogroup and Mixture Proportion Estimates

MixtureAceMT™ co-estimates the haplogroups and overall proportion ratios via the MIXEMT algorithm. Contributing haplotypes and proportion ratios are accurately called for most mixtures. Minor contributors begin to drop when proportion ratios exceed ~50:1 and when using short reads (e.g., Promega kit). Spurious minor haplogroups are called at all contribution ratios and haplogroup combinations when using short reads due in part to apparent homoplasy in the phylogenetic tree which can generate support for haplogroups not present in the mixture. Spurious haplogroups can be reduced by NUMT filtering and by *post-hoc* heuristic analysis.

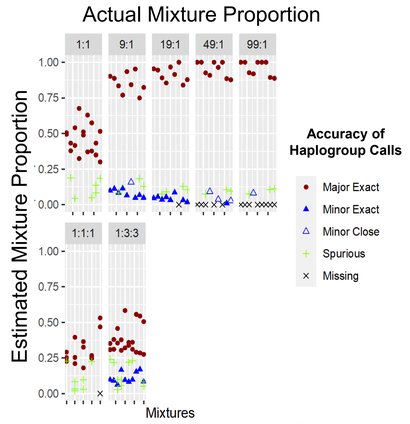



Figure from McElhoe et al. 2024

## Number of Contributor (NOC) Estimates

NOC can be estimated from the count of mtDNA haplogroups or haplotypes deconvoluted from mixed samples. The true NOC can be underestimated due to haplotype sharing. Contributors from the same population group may share a haplogroup. Contributors from the same maternal line may share a haplotype. Haplotype sharing is more likely across the hypervariable region (HVR) than across the whole genome (WG). Recovery of contributor haplotypes is partially dependent upon read length, with longer reads recovering more complete haplotypes.

**MixtureAceMT™ analysis of in-silico long reads duplicating the molecular extents of 6-amplicon sequencing by ONT. Discrimination within haplogroups is possible when considering hotspot and private mutations not considered in haplotyping.**

| Sample ID | Subpopulation | Haplogroup | Pairwise Discriminating Variants | |
|---|---|---|---|---|
| | | | HVR | WG |
| DQ304908 | African American | L1b1a3 | 0 | 1 |
| JX303892 | African American | L1b1a3 | | |
| GU296641 | Caucasian | U5b2b1b | 0 | 2 |
| JQ702927 | Caucasian | U5b2b1b | | |
| JQ704040 | SW Hispanic | C1c6 | 0 | 5 |
| JQ705574 | SW Hispanic | C1c6 | | |

## NUMT Filtering Reduces Spurious Haplogroups

NUMTs (nuclear mitochondrial DNA) are fragments of mitochondrial DNA (mtDNA) that have been integrated into the nuclear genome. When NUMTs include primer target sequences, they may co-amplify with mtDNA potentially yielding false SNP signals due to divergent evolution. NUMTs appear to introduce low level spurious haplogroups, which are suppressed by NUMT filtering. NUMT filtering does not affect recovery of true minor contributor haplogroups (i.e., ~50:1).

**1:1:1 Mixture of Contributors from Dissimilar Haplogroups**

| Sample Haplogroup | Before Filtering | | After Filtering | |
|---|---|---|---|---|
| | Contribution | Ratio | Contribution | Ratio |
| **D4e2** | 0.27 | 1.9 | 0.32 | 1.1 |
| **A2ac1** | 0.21 | 1.5 | 0.30 | 1.0 |
| **U5a1a1** | 0.14 | 1.0 | 0.38 | 1.3 |
| U5a1a | 0.19 | 1.4 | 0 | na |
| HV[1] | 0.18 | 1.2 | 0 | na |
| A2 | 0.01 | <1 | 0 | na |

*PSU Sample: 04D-05U-09A*

**3:3:1 Mixture of Contributors from Close Haplogroups**

| Sample Haplogroup | Before Filtering | | After Filtering | |
|---|---|---|---|---|
| | Contribution | Ratio | Contribution | Ratio |
| **J1c2c2** | 0.35 | 2.5 | 0.36 | 2.0 |
| **J1c8a** | 0.33 | 2.3 | 0.46 | 2.6 |
| **J1c4** | 0.14 | 1.0 | 0.18 | 1.0 |
| J | 0.18 | 1.3 | 0.00 | na |

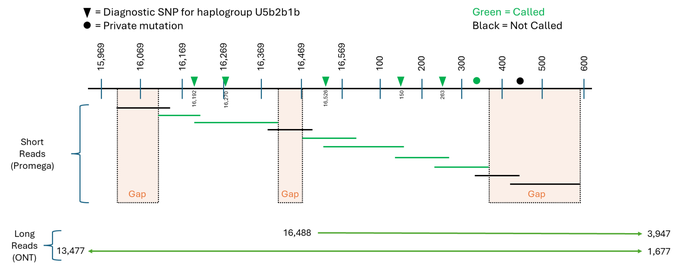
*PSU Sample: 06J-01J-08J*

**49:1 Mixture of Contributors**

| Haplogroup | Before Filtering | | After Filtering | |
|---|---|---|---|---|
| | Contribution | Ratio | Contribution | Ratio |
| **U5a1a2a1** | 0.95 | 19 | 0.97 | 29 |
| **A2ac1** | 0 | na | 0 | na |
| H26 | 0.05 | 1 | 0.03 | 1 |

*PSU Sample: 10U-09A*

**Illustration of short read haplotype gap filling when using long reads.**




## Materials and Methods

Short read biological mixtures used in NUMT filtering were amplified using the PowerSeq™ Whole Mito System (Promega) as described in McElhoe et al. (2024). Long read biological mixtures avoiding NUMT amplification were amplified as described in Ramos et al. (2009) and sequenced using the Oxford Nanopore Technologies GridION with R10.4 flow cells. ONT sequencing was performed at the University of Vermont Integrative Genomics Facility. Long range in-silico mixtures were generated by trimming 1000 Genomes FASTA sequences to match biological amplicon extents. FASTQ files were generated using custom Python scripts. All biological and in-silico mixtures were analyzed using MixtureAceMT™ (NicheVision). Haplogroup confirmation was performed using HaploGrep2 (short reads) or HaploGrep 3 (long reads).