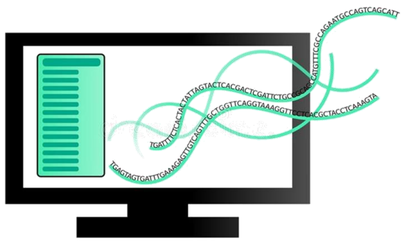
AN ASSESSMENT OF PROBABILISTIC APPROACHES TO MTDNA MIXTURE INTERPRETATION



Jennifer A. McElhoe[1,*], Alyssa Adesso[1], Brian Young[2], Jeff Smith[2], and Mitchell M. Holland[1]

1. The Pennsylvania State University, Forensic Science Program, Department of Biochemistry & Molecular Biology, University Park, PA, United States
2. NicheVision LLC, 526 S. Main St., Akron, OH, United States

I. Background

Mitochondrial (mt) DNA plays an important role in the fields of forensic and clinical genetics, molecular anthropology, and population genetics, with mixture interpretation being of particular interest in medical and forensic genetics. In forensics, mixture deconvolution generally relies on genotyping of STRs, but this approach has limitations when evaluating degraded samples with multiple contributors or samples where the proportion of contributors is similar. The high copy number, haploid state (only a single haplotype contributed per individual), high mutation rate, and well-known phylogeny of mtDNA, makes it an attractive marker for mixture deconvolution in damaged and low quantity samples of all types.

To date there have been a handful of approaches to mtDNA mixture deconvolution in the literature including Mixemt[1], MMDIT[2], and DEploid[3]. Mixemt uses an iterative process to approximate a maximum likelihood function (EM algorithm) and the well-annotated phylogeny of mtGenomes, MMDIT is an open source platform using a Bayesian haplotype estimation, and DEploid uses a Markov Chain Monte Carlo process with a custom bioinformatic pipeline and R. The current options require a reasonable to high level of bioinformatic knowledge and tend to perform well with sequencing methods that produce larger pieces of DNA sequence.

This study presents a cloud-based, user-friendly deconvolution tool, MixtureAce MT Live (NicheVision LLC) that is a combined interface of two previously developed tools, MixtureAce™ (NicheVision) and Mixemt. MixtureAce™ is a plug-in tool for ArmedXpert that was initially designed to allow analysis of sequence-based alleles in STR analysis. Mixemt uses an EM algorithm and heuristic filters to estimate Haplogroup and frequency information from mixed DNA samples. Our evaluation demonstrates the utility of the new tool to deconvolute mtDNA mixtures sequenced using two commercially available MPS kits (the Promega PowerSeq® Whole Mito System and the Thermo-Fisher Scientific Precision ID mtDNA Whole Genome Panel) that target small amplicons (averaging 211 & 172bps) spanning the mtgenome.

II. Dataset Designations

Two- and three-person mixtures were created from individuals of diverse haplotypes and population groups (European, African, Asian, and Latino ancestry) from sole source buccal samples that were previously sequenced using the PowerSeq Whole Mito System on a MiSeq (in-silico dataset) and from mixed DNA buccal extracts sequenced using the Precision ID mtDNA Whole Genome Panel on a MiSeq (biological dataset). Ratios considered were 1:1, 1:3, and 3:1 for 2-person mixtures (n=63 in-silico; n=45 biological) and 1:1:1, 1:3:3, and 3:1:3 for 3-person mixtures (n=63 in-silico; n=15 biological).

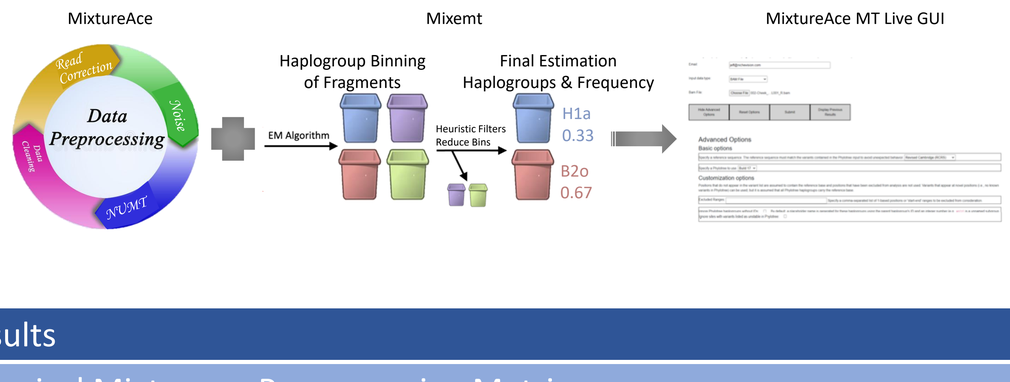
In-silico mixtures were created using a combination of Bedtools and Samtools to subset a total of 720K fragments (2-person) and 840K fragments (3-person) of DNA sequence that were aligned to the mtgenome using GeneMarker HTS (SoftGenetics).

Biological mixtures were created from sole source buccal extracts that were quantified using a custom mtDNA qPCR assay[4] to determine the copies of mtDNA present in the individual extracts, diluted, and then combined to form 2- or 3-person mixtures. Mixed DNA samples were amplified using the Precision ID kit and a 600-cycle-kit with 275x275 paired end sequencing on a MiSeq.

III. Mixture Deconvolution

MixtureAce MT Live - a Cloud-based Deconvolution Tool

MixtureAce MT Live combines MixtureAce preprocessing for read correction, noise reduction, NUMT reduction, and primmer trimming with the Mixemt EM algorithm & heuristic filters in a user-friendly, cloud-based graphic user interface.

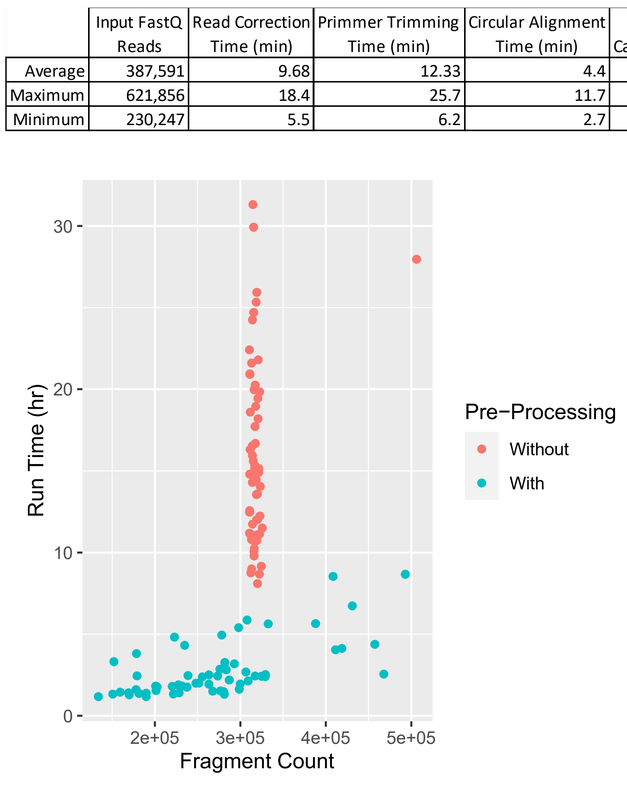


IV. Results

1. Biological Mixtures – Preprocessing Metrics

MixtureAce MT Live preprocessing summary for biological mixtures. Metrics include input FastQ reads from the MiSeq, read correction run time, primmer trimming run time, alignment run time, number of possible NUMTs identified, the run times for NUMT identification & removal, the number of NUMTs ultimately removed, and the total time required for preprocessing.

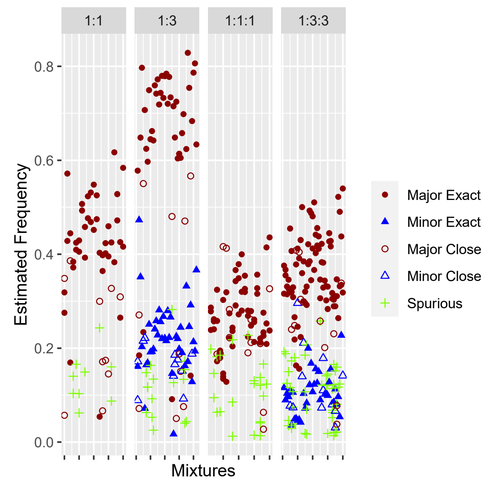
	Input FastQ Reads	Read Correction Time (min)	Primmer Trimming Time (min)	Circular Alignment Time (min)	NUMT Candidates	NUMT ID Time (min)	NUMT's Removed	NUMT Removal Time (min)	Total Run Time (min)
Average	387,591	9.68	12.33	4.4	60,195	2.84	3,915	2.42	31.66
Maximum	621,856	18.4	25.7	11.7	100,402	5.7	8,114	5.3	64.20
Minimum	230,247	5.5	6.2	2.7	28,445	1.8	885	1.1	17.60



- *One of the major drawbacks of the previously available software for mtDNA mixture deconvolution has been that implementation can be time-consuming.*
- *MixtureAce MT Live reduces confounding features such as potential NUMTs and noise which reduces the time required for the EM algorithm and heuristic filters to produce haplogroup and frequency estimates.*

Mixemt run time based on the sequencing file size represented as the number of sequencing fragments used for haplogroup and frequency estimation. The comparison highlights the difference in run time for biological samples that were preprocessed with MixtureAce (biological dataset) and samples without preprocessing (in-silico dataset).

2. In-Silico Mixtures - Deconvolution

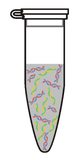


Frequency estimates for in-silico mixtures of 2- and 3- individuals. Filled shapes are where Mixemt identified the exact haplogroup for the contributor and outlined shapes are a closely related haplogroup. Contributors not detected (x) and spurious haplogroups (+).

Summary tables for 2- and 3-person in-silico mixtures of 720K and 840K fragments respectively.

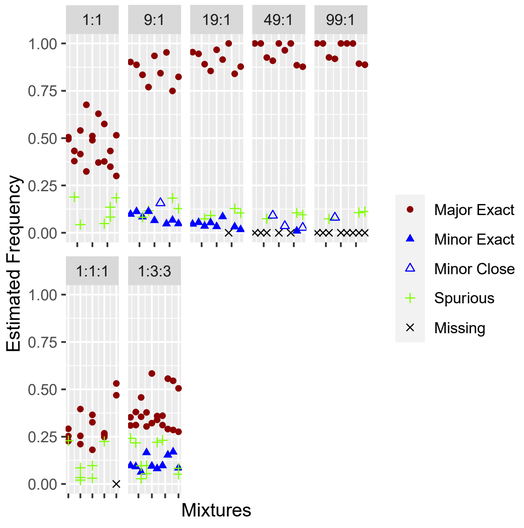
In-Silico Mixture	Two Exact Matches Only	Major Haplogroup		Minor Haplogroup		Additional Haplogroups
		Exact Match	Close Match	Exact Match	Close Match	
1:1	8/21	40	2	-	-	12/21
1:3	19/42	41	1	39	3	22/42

Mixture	Three Exact Matches Only	Major Haplogroup		Minor Haplogroup		Additional Haplogroups
		Exact Match	Close Match	Exact Match	Close Match	
1:1:1	5/20	56	4	-	-	14/20
1:3:3	10/40	79	1	35	5	27/40



- *Mixemt was able to detect the correct haplogroups for 95% of contributors.*
- *High incidence of spurious haplogroups.*
- *Spurious haplogroups most likely due to small amplicon size and true contributor haplotype differences from Phylotree (private mutations).*
- *Inclusion of known haplotypes resolves spurious haplogroups.*

3. Biological Mixtures - Deconvolution



Frequency estimates for biological mixtures of 2- and 3-individuals. Filled shapes are where MixtureAce MT Live identified the exact haplogroup for the contributor and outlined shapes are a closely related haplogroup. Contributors not detected (x) and spurious haplogroups (+).

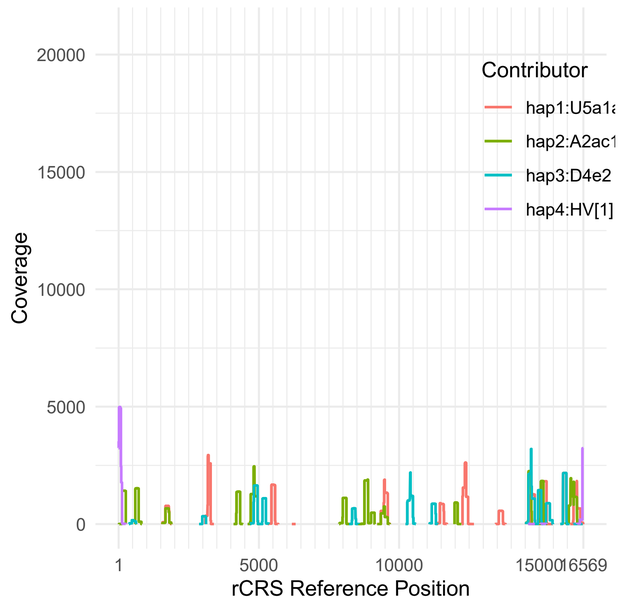
Summary tables for 2- and 3-person biological mixtures.

Mixture	Two Exact Matches Only	Major Haplogroup		Minor Haplogroup			Additional Haplogroups
		Exact Match	Close Match	Exact Match	Close Match	Not Detected	
1:1	4/9	9	0	-	-	-	4/9
9:1	4/9	9	0	8	1	-	4/9
19:1	4/9	9	0	8	0	1	4/9
49:1	0/9	9	0	0	4	5	3/9
99:1	0/9	9	0	0	1	8	3/9

Mixture	Three Exact Matches Only	Major Haplogroup			Minor Haplogroup		Additional Haplogroups
		Exact Match	Close Match	Not Detected	Exact Match	Close Match	
1:1:1	0/5	14	0	1	-	-	4/5
1:3:3	3/10	20	0	0	10	0	7/10

- *MixtureAce MT Live was able to detect the correct haplogroups for 95% of contributors in samples with minor contribution ≥ 5%.*
- *68% of minor contributors were not detected with minor frequencies of 1- 2%*
- *High incidence of spurious haplogroups*
- *Inclusion of known haplotypes resolves spurious haplogroups and helps to identify minor contributors present in trace quantities.*

4. Biological Mixture – Example Genome Coverage



Coverage of mtDNA reference sequence from fragments assigned to detected mixture contributors for a 3-person, 1:1:1 mixture sample.

Regions surrounding known variant positions from Phylotree are recovered using information from assigned reads including private variants.

The amount of the underlying haplotypes that can be recovered depends on the number of differences that are known to exist between the haplogroups and the size of the fragments in the sample.

Fragments that cannot be assigned to a contributor are represented by the gray area while colored lines represent coverage for read fragments assigned to a detected contributor.

V. Conclusions

- **MixtureAce MT Live effectively detects the correct contributing haplogroups in sequencing data generated from small amplicons.**
- **MixtureAce MT Live effectively reduces confounding features such as NUMTs & noise, reducing traditionally prohibitive processing times.**
- **Spurious haplogroups resulting from small amplicon sequencing and private mutations that differ from Phylotree can be resolved by including known haplotype/s in the evaluation.**
- **Trace contributors remain challenging to resolve even with deep sequencing.**

VI. Citations

1. Vohr et al. 2017. FSIG 30:93-105. DOI:https://doi.org/10.1016/j.fsigen.2017.05.007
2. Mandape et al. 2021. FSIG 55. DOI:https://doi.org/10.1016/j.fsigen.2021.102568
3. Smart et al. 2021. Genes 12(2)28. DOI:https://doi.org/10.3390/genes12020128
4. Galimore et al. 2018. FSIG 32:7-17. DOI:https://doi.org/10.1016/j.fsigen.2017.09.013

VII. Acknowledgements

Funding: National Institute of Justice 2020-DQ-BX-0004.