

The author(s) shown below used Federal funds provided by the U.S. Department of Justice and prepared the following final report:

**Document Title: Development of Linkage Phase Analysis
Software for Resolving mtDNA Mixtures**

Author: Phillip B. Danielson, Ph.D.

Document No.: 236536

Date Received: November 2011

Award Number: 2009-DN-BX-K047

This report has not been published by the U.S. Department of Justice. To provide better customer service, NCJRS has made this Federally-funded grant final report available electronically in addition to traditional paper copies.

**Opinions or points of view expressed are those
of the author(s) and do not necessarily reflect
the official position or policies of the U.S.
Department of Justice.**

Final Technical Report

REPORT TITLE: Development of Linkage Phase Analysis Software for Resolving mtDNA Mixtures

AWARD NUMBER: DNA Research and Development Award 2009-DN-BX-K047 (Funded at \$42,646)

AUTHOR: Phillip B. Danielson, Ph.D.

ABSTRACT

Overview – Mitochondrial DNA (mtDNA) sequencing can provide crucial information to forensic investigators when the quantity and quality of DNA would otherwise be limiting. “Situational” mixtures of mtDNA from two or more individuals and naturally occurring heteroplasmy present challenges that typically preclude analysis by direct DNA sequencing. Denaturing High-Performance Liquid Chromatography (DHPLC) is a chromatographic means of fractionating DNA mixtures prior to sequencing. Subsequent linkage phase analysis of direct sequence data from DHPLC fractions makes it possible to reliably deconvolve mtDNA mixtures. Although this approach to mixture deconvolution has been thoroughly validated, the lack of a reliable software application to handle the computational demands of linkage phase analysis represented a major obstacle that discouraged practitioners from evaluating or adopting this otherwise powerful technology for resolving mtDNA mixtures.

Project Objectives - The central goal of the research funded under Award 2009-DN-BX-K047, therefore, was to develop and test a software application to automate the computationally intensive analysis of electrophoretic data necessary to determine the haplotypes of individual contributors to an mtDNA mixture. Developing such an application required completion of three major research objectives:

- (1) **Develop a robust software application and user/friendly graphical interface** to import and deconvolve sequence electropherogram files.
- (2) **Test the accuracy of the software application** on a broad range of mixture ratios and mixed base positions using both reference and casework-type samples.
- (3) **Rigorously analyze the performance and accuracy of the software application** and make appropriate revisions to resolve any anomalies

The successful completion of these objectives now provides forensic practitioners with a reliable means of automating the deconvolution of mtDNA mixtures so as to facilitate the accurate analysis of otherwise challenging samples.

Results and Conclusions - All core objectives have been successfully achieved. The software application (FLiPARS 2.0) efficiently automates the computationally intensive process of mtDNA mixture deconvolution by linkage phase analysis. The current release of FLiPARS 2.0 has been successfully run on a variety of Windows, Mac OS and Linux operating systems. The Graphical User Interface has been designed to be intuitive and user friendly. The base-calling

and alignment algorithms also allow the sequence being analyzed to be edited in accordance with the judgment and experience of a skilled practitioner.

The accuracy of the software application to deconvolve mtDNA mixtures was determined by analyzing a large dataset of mixed sequence electropherograms (up to 11,581 comparisons). Even without manual editing of sequence alignments, FLiPARS 2.0 accurately resolve nearly 70% of all aligned sequences. Of sequences which were not resolved, nearly all involved length variants characterized by stretches of mixed bases at nearly every position – *i.e.*, sequences for which FLiPARS 2.0 was not directly designed to handle. In all cases where mixtures were successfully deconvolved, however, the statistical confidence of the linkage phase determination typically exceeded 99.9%

The performance of the FLiPARS 2.0 application was also tested on a variety of casework-type samples which included mixtures on varied substrates, mixtures subjected to environmental insult and hair and bone samples coated with various body fluids. In all cases, the individual contributor haplotypes identified were in full concordance with sequencing results from individual reference samples. FLiPARS 2.0 produced linkage phase analysis results with a high degree of base resolution confidence - ranging from 99.77% to 99.99%. As a result, the developmental validation of a software application for linkage phase analysis offers practitioners the opportunity to obtain potentially useful information from what might otherwise be uninterpretable samples.

Table of Contents

EXECUTIVE SUMMARY

Introduction and Statement of Problem	4
Core Research Objectives	5
Methods	6
Results and Discussion	7
Implications for Policy, Practice and Future Research	11
Literature Cited in the Executive Summary	12

FINAL TECHNICAL REPORT (MAIN BODY)

Introduction and Statement of the Problem	13
Statement of Core Research Objectives	16
Methods	16
<i>Human Subjects</i>	16
<i>Casework Type mtDNA Mixtures</i>	17
<i>Programming language:</i>	17
<i>Software Forward Compatibility</i>	17
<i>Eliminating Dependencies on Secondary Software</i>	18
<i>Software Performance and User Interface</i>	18
<i>Sequence Alignment Algorithm</i>	19
<i>Validation of the Software Application's Accuracy</i>	19
<i>Version Control and Licensing</i>	19
Results and Discussion	19
<i>FLiPARS 2.0 Platform Interoperability</i>	19
<i>Modularized Plug-Ins</i>	20
<i>Intuitive Graphical User Interface</i>	20
<i>FliPARS 2.0 Performance Testing</i>	24
<i>Casework-Type Samples (Body Fluid Mixtures on Varied Substrates)</i>	26
<i>Casework-Type Samples (Environmental Insult Mixtures)</i>	26
<i>Casework-Type Samples (Hair and Bone Mixtures):</i>	27
<i>Alignment Ambiguities</i>	28
<i>Noisy Sequencing Reactions</i>	30
Implications for Policy and Practice	30
Implications for Further Research	31
Cited References	32
Dissemination of Research Findings	34

EXECUTIVE SUMMARY

Introduction and Statement of Problem

Mitochondrial DNA (mtDNA) sequencing can provide crucial information to forensic investigators when the quantity and quality of DNA would otherwise be limiting. A “situational” mixture of mtDNA from two or more individuals in a single sample, and even naturally occurring heteroplasmy, present challenges that typically preclude successful mtDNA analysis by DNA sequencing^[1-3]. This is because direct sequencing of a mixture of two or more DNA amplicons yields electrophoretic traces characterized by overlapping peaks at sites where the amplicons differ in primary sequence. Because peak height is sequence context dependent, it cannot be used by itself to determine the absolute or even relative quantities of DNA from the individual contributors to the mixture. This can impede the forensic use of mtDNA. A reliable approach to resolving mixtures would substantially increase the power of mtDNA analysis by allowing its use in cases where current approaches yield results of limited or no utility.

Denaturing High-Performance Liquid Chromatography (DHPLC)^[4,5] is a chromatographic means of fractionating natural (heteroplasmic) or situational (multi-contributor) DNA mixtures prior to sequencing^[6,7]. In contrast to alternative approaches that have been proposed for the deconvolution of mtDNA mixtures^[8-10], DHPLC does not require reamplification or excessive sample manipulation to resolve a mixture of different mtDNA haplotypes. Under a prior DNA Research and Development Award (2003-IJCX-K104), this approach was rigorously validated as a means of resolving mtDNA mixtures^[11-13] using both reference and casework type samples. Specifically, it was demonstrated that: (1) DHPLC was sufficiently sensitive to detect and fractionate mixtures involving all classes of polymorphisms; (2) there was a statistically significant correlation between peak height and DNA quantity ratios at mixed-base positions in sequencing electropherograms; (3) linkage phase analysis of DHPLC-fractionated samples was a reliable means of determining individual haplotypes from comparative sequence data.

Through careful quantitative analyses of sequencing electropherograms, changes in the relative heights of overlapping fluorescent peaks at all mixed-base positions can be tracked among two or more DHPLC fractions. The observation of coordinated shifts in relative fluorescence ratios for a given set of nucleotides is consistent with them being in the same linkage phase and thus representing the same amplicon (*i.e.*, contributor). While extremely reliable, this approach requires computationally intensive analyses of enormous datasets - too large of a task for an analyst to manage “in their head”. Accordingly, a prototype software application termed FLiPARS 1.0 (Fractional Linkage Phase Analysis Resource System) was written to automate linkage phase analyses. This early software prototype, was a conglomeration of Perl scripts and dynamically linked Excel files integrated into a Microsoft Windows Visual Basic application. The program demonstrated the potential of the underlying forensic technology it supported. It generated a linkage phase report (Figure 1) listing the mixed-base position analyzed, the amount of fluorescence shift between DHPLC fractions, the linkage groups of the bases for each contributor, the statistical confidence of the individual base assignments, and the number of samples on which the statistical estimates are based.

The FLiPARS 1.0 prototype, however, was built more as a proof of concept than as a production ready application for widespread use by forensic practitioners. Thus it is not surprising, that the prototype was slow, contained a number of bugs and was simply not easy to

use. The graphical user interface was particularly confusing for new users; .scf files exported from ABI sequencers had to be manually converted to Tab delimited .txt files before they could be analyzed and; both incorrect and some seemingly correct files cause FLiPARS 1.0 to hang, inexplicably crash, or to produce anomalous results while providing the end-user with no indication of what went wrong.

Thus, even though DHPLC and linkage phase analysis had been thoroughly validated for the resolution of mtDNA mixtures, the lack of a reliable software application to handle the computational demands of the process represented a major obstacle that discouraged practitioners from evaluating or adopting an otherwise powerful technology for resolving mtDNA mixtures.

Linkage Phase Determination	Position	Comparison	Average	Confidence	Minimum	Samples		
		7 and 11	Confidence	Std Dev	Confidence	Behind Stats	Increasing	Decreasing
	146	78.7%	99.998%	0.007%	99.945%	192	146T	146C
	150	82.8%	99.998%	0.006%	99.950%	190	150T	150C
	185	82.4%	99.998%	0.006%	99.950%	190	185G	185A
	188	81.3%	99.998%	0.006%	99.962%	204	188A	188G
	195	82.7%	99.998%	0.006%	99.950%	190	195C	195T
	198	83.6%	99.998%	0.004%	99.967%	214	198T	198C
	222	82.2%	99.998%	0.006%	99.950%	190	222C	222T
	228	80.8%	99.998%	0.003%	99.979%	184	228G	228A

Executive Summary Figure 1 – A linkage phase determination table (detail from the analysis report) generated using the FLiPARS 1.0 prototype software. Listed are the mixed-base positions, the amount of fluorescence shift between selected DHPLC fractions, statistical confidence parameters of the linkage phase base assignments and the linkage phase results defining each contributor.

Core Research Objectives – The central goal of the research funded under Award 2009-DN-BX-K047 was to develop and test a software application to automate the computationally intensive analysis of electrophoretic data that is necessary to determine the linkage phase (*i.e.*, haplotypes) of individual contributors to an mtDNA mixture. This was achieved through the completion of three major research objectives. These were:

- (1) **Develop a robust software application and user/friendly graphical user interface** that could be used to import and deconvolve sequencing electropherogram files from DHPLC fractionated mtDNA mixtures.
- (2) **Test the accuracy of the software application** on multiple electrophoretic data files for mtDNA mixtures consisting of a broad range of mixture ratios, >200 mixed base positions for both reference and casework-type samples.
- (3) **Rigorously analyze the performance and accuracy of the software application** and make appropriate revisions to the software to resolve any anomalies in the linkage phase determination results.

The successful completion of these objectives was designed to provide forensic practitioners with a reliable and easy means of automating the deconvolution of mtDNA mixtures. Ultimately, this will provide practitioners with the opportunity to evaluate and successfully analyze some types of challenging samples.

Methods

Human Subjects – All research employing human-derived mtDNA sequence data under DNA Research and Development Award 2009-DN-BX-K047 was IRB reviewed, approved and conducted in full compliance with U.S. Federal Policy for the Protection of Human Subjects (Basic DHHS Policy for Protection of Human Research Subjects; 56 FR 28003). All projects using human subjects are reviewed no less than annually. All participants signed a statement of informed consent to participate in the research. As no health care associated information was collected, HIPPA authorization was not required.

Casework Type mtDNA Mixtures – Evaluation of the software performance on non-pristine samples employed mtDNA sequence files generated from casework-type material. This included the analysis of mtDNA sequence electropherograms representing a variety of mixed tissue sources (blood, semen, saliva, hair shafts and bone), deposited on a variety of substrates (cotton cloth, nylon carpeting, blue denim, leather, cigarette butts and wall board) and exposed to a range of environmental contaminants/inhibitors (gasoline, used motor oil, soil, laundry detergent, acetic acid and sodium hydroxide).

Programming language – The development of FLiPARS 2.0, employed C# (C-Sharp)^[14], a relative of Visual Basic, as the programming language. This was done to facilitate the stable design and future adaptability of the linkage phase software application. The use of C# allowed FLiPARS 2.0 to be written to take advantage of “strong typing” which improved program stability. The use of C# facilitated the development of an application that was simple to use, reliable and which can grow in response to suggestions and feedback from forensic practitioners.

Software Forward Compatibility – In order to ensure the sustained utility of the FLiPARS 2.0 software application, it was necessary to make it functional on alternative (*i.e.*, non-Windows) and new operating systems. This was achieved by taking advantage Microsoft’s .NET paradigm. The .NET paradigm is platform independent which enabled the FLiPARS 2.0 programs to be run on a number of different operating systems (*e.g.*, Windows XP, Windows Vista, Windows 7, Mac OS X and Linux).

Eliminating Dependencies on Secondary Software – The C# language provided a robust feature set that was employed across the entire application, from the interface to the underlying statistical analyses^[15]. This streamlined approach allowed the elimination of dependencies on secondary applications (*e.g.*, Microsoft Excel) through the use of modularized plug-ins that convert input data into a standard format. This interface has been made freely available; thereby enabling future developers to produce FLiPARS plug-ins as needed. This will ensure the viability of FLiPARS for years to come. The .NET paradigm, also alleviates intellectual property concerns over proprietary data formats from commercial instruments.

Software Performance and User Interface – A more user-friendly graphical user interface (GUI), was designed to enable a practitioner with minimal training to use the software with confidence. In order to achieve this, ambiguous buttons and terms were removed and the intermediate outputs of multistep algorithms were unified. Also, the “Peak Stringency” and “Noise Threshold” dialogs were moved out of the main window and into an “Advanced Settings” tab. Finally, the terminal output of each analysis was limited to a single top-level screen from which the unnecessary machinations of the underlying computer code have been removed.

Sequence Alignment Algorithm – Query sequence were aligned against the revised CRS using a dynamically generated table (Smith-Waterman approach). Alternative algorithms are available, but stray from a strict edit-distance definition of alignment and thus were not used.

Validation of the Software Application's Accuracy – The accuracy of linkage phase determination by FLiPARS 2.0 was evaluated by analyzing a pre-existing dataset of sequence electropherograms generated from mtDNA mixtures. The dataset consisted of two-contributor mixtures at stepped ratios from 1:99 to 99:1. Collectively the mixed samples differed at >200 base positions throughout the HV1 and HV2 regions. A second set of mixed sequences from casework-type samples were also tested. These included mtDNA mixtures involving a variety of tissue sources, substrates, and environmental contaminants/inhibitors. The FLiPARS 2.0 software was used to resolve the mixtures and the results were compared to the known haplotypes for the sample. These analyses were conducted first without user confirmation of peak calls and then after visual inspection of the electrophoretic traces to identify mixed base positions potentially missed or miscalled by the software.

Version Control and Licensing – FLiPARS is licensed under the Apache 2.0 License, providing free access to the source code for virtually any purpose. The code for FLiPARS 2.0 was placed under version control using “Apache Subversion™”. Subversion is an open source version control system that allows for detailed tracking of problems and the resulting fixes for those problems. It has provided a convenient means for other developers to access the source code and to simultaneously work on a project should that need arise.

Results and Discussion

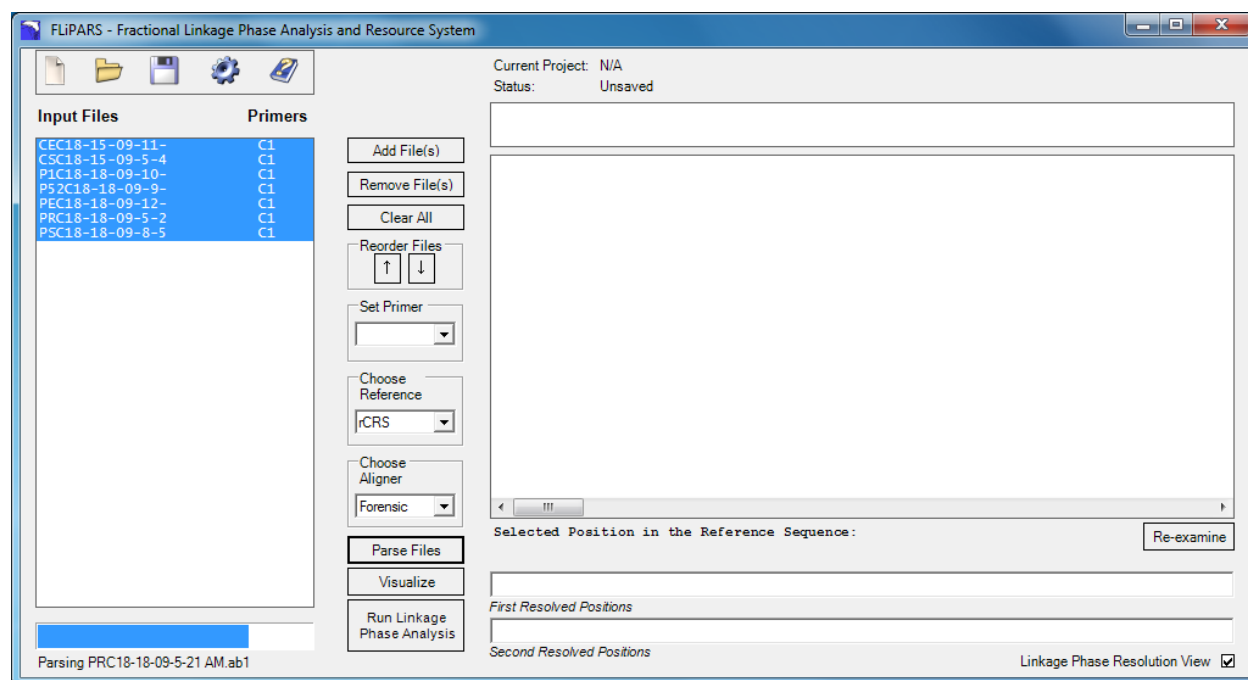
FLiPARS 2.0 Platform Interoperability: A software application (FLiPARS 2.0) which automates the computationally intensive process of mtDNA mixture deconvolution by linkage phase analysis has been written in C#. Unmodified, the current release of FLiPARS 2.0 has been successfully run on the following operating systems:

- Windows XP Professional (32 bit)
- Windows Vista Professional (64 bit)
- Windows 7 Professional (32 bit and 64 bit)
- Slackware Linux 13.0 w/ Gnome Slackbuild and Mono

Modularized Plug-Ins – Dependencies on secondary software applications for input file parsing have been replaced by a set of modularized plug-ins. The plug-ins allow for the parsing of three different file formats which have been commonly used to save sequence data. These include raw tab-delimited text files as well as the more common .ABI and .SCF file formats generated by the Applied Biosystems instruments used by most forensic laboratories that perform mtDNA testing.

Intuitive Graphical User Interface – A completely new FLiPARS 2.0 interface has been implemented. The main screen (Figure 2) was designed to present the user with an input box (on the left side of the window) for their files and the expected “add”, “remove”, “clear” and “reorder functionality” of other typical data input boxes. The associated primers for each input file can be set individually as can the reference sequence that the program uses for alignment of the input sequences. This provided the program with compatibility for potential applications outside

that of the forensic community as well as flexibility in the event that new forensic technologies emerge or different portions of the mtDNA genome are targeted in the future. A status bar has also been implemented to apprise the user of the status of any long running operations that the program may be performing in the background while they wait.

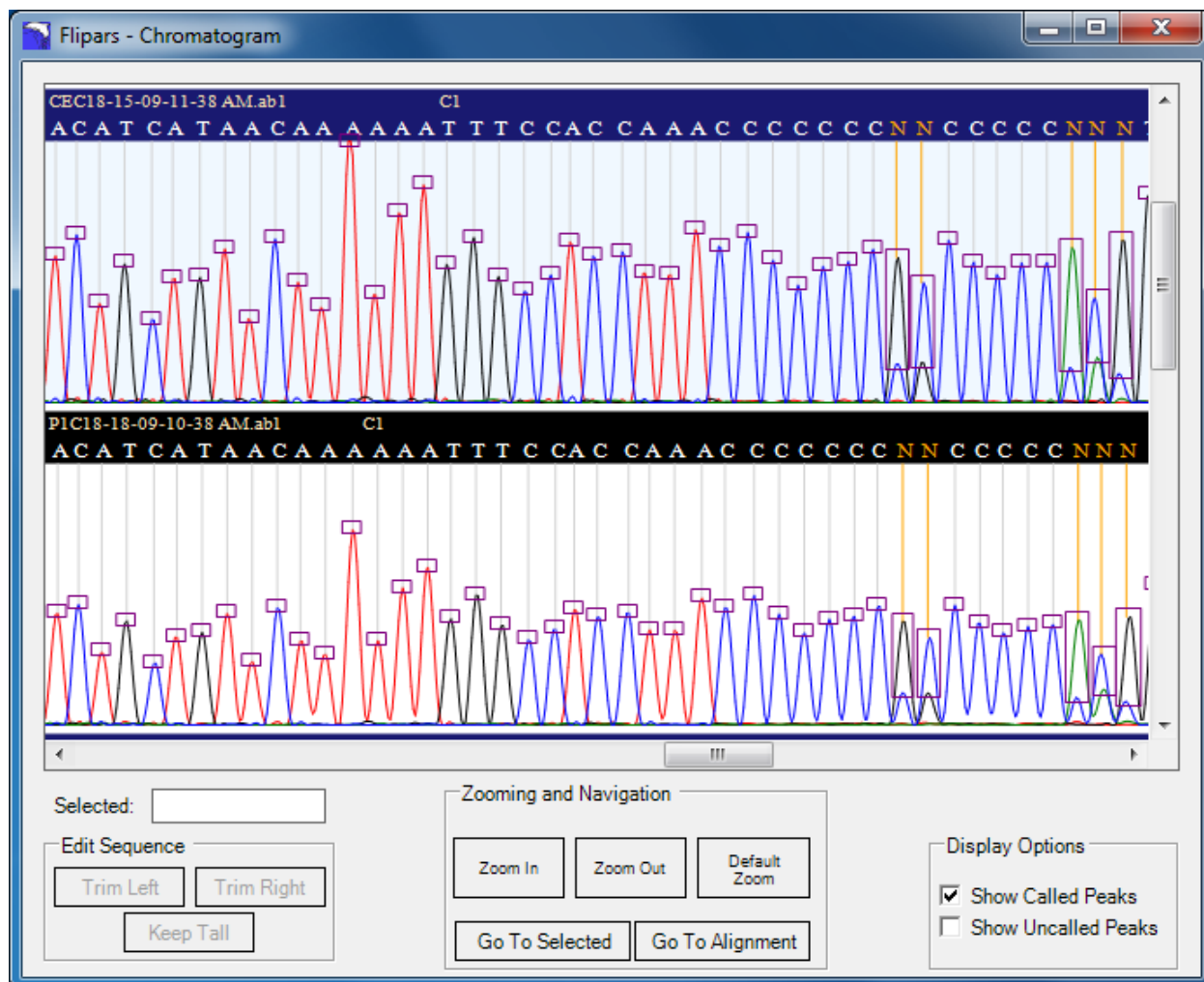


Executive Summary Figure 2 - FLiPARS 2.0 revamped user interface as shown to the user when it is first opened. In this example, a set of files has been added and are being parsed. Note the status bar which shows the progress made in processing a small batch of files.

Once processed by the internal FLiPARS 2.0 base-calling algorithm, the user is able to view the corresponding chromatograms by clicking on the “Visualize” button. This brings up a display (Figure 3) which contains all of the chromatograms input to the software. As needed, the underlying data are (re)processed in accordance with manual edits made by the practitioner. Each chromatogram is offset by different colored bars and backgrounds to ensure that the user does not confuse calls between two adjacent chromatograms.

A separate settings window has been implemented to allow user-specified modifications of the underlying analysis settings (*e.g.*, peak stringencies and noise thresholds) and internal data sets (*e.g.*, primer and reference sequences). Since these are expected to be rarely (if ever) modified by the end user, they have been contained in their own separate window which can be reached from the main screen by clicking on the “Settings” icon. Help boxes (denoted by “?” buttons) have been provided to provide details on each particular setting in the event that the user requires more information. All saved settings are maintained between FLiPARS 2.0 sessions (*i.e.*, after closing and reopening the program).

The base-calling and alignment algorithms also allow practitioners to edit the sequence being analyzed. This is an important aspect of the program as no base-calling algorithm can claim perfect accuracy or substitute for the judgment and experience of a skilled practitioner. All changes are tracked and can be reversed to restore the original peak calls.



Executive Summary Figure 3 - FLiPARS 2.0 chromatogram display at the default zoom view. This display clearly shows mixed base calls highlighted by orange lines at the peak apex and basecall boxes that encompass the peaks of both bases. It also provides the user with the file name and primer used (as specified in the main display window) for each sequence.

FLiPARS 2.0 Performance Testing – The accuracy of the software application to deconvolve mtDNA mixtures was determined by analyzing a large dataset (up to 11,581 comparisons) of mixed sequence electropherograms. These mixtures consisted of two mtDNA amplicons that were combined at stepped ratios from 1:99 to 99:1. Collectively the mixed samples differed in sequence at >200 base positions distributed throughout the HV1 and HV2 regions. FLiPARS 2.0 was used to determine the individual contributor haplotypes for each mixture and the results were compared with the known reference sequences of the respective contributors.

Without manual editing of sequence alignments, FLiPARS 2.0 accurately resolve nearly 70% of all aligned sequences (Table 1). Of sequences which were not resolved, nearly all involved length variants characterized by stretches of mixed bases at nearly every position. The deconvolution of such “out of register” sequence by FLiPARS 2.0 typically required manual editing after which, FLiPARS 2.0 was able to make accurate linkage phase determinations. A second class of alignments that were not amenable to deconvolution by FLiPARS 2.0 were those where only a minimal shift in peak height ratios was present – e.g., comparisons between a 99:1

and a 95:5 mixture. In all cases where mixtures were successfully deconvolved, however, the statistical confidence of the linkage phase determination typically exceeded 99.9%

While the ability of FLiPARS 2.0 to deconvolve many mtDNA mixtures without the assistance of a practitioner is encouraging, it is important to recognize that this software application was not designed to replace forensic practitioners. Rather it is a tool to assist the practitioner in resolving mtDNA mixtures by automating the computationally intensive process of linkage phase analysis. This is evidenced by the fact that successful mixture deconvolution of sequence traces approached 100% for files that were examined by a human analyst and edited based on their experience.

Executive Summary Table 1 – Percentage of successful linkage phase analyses by FLiPARS 2.0 using raw (i.e., not practitioner edited) sequence files.

% First Fraction - % Second Fraction	99 - 1	64.9%	64.6%	62.6%	74.7%	84.6%	85.0%	85.9%	84.2%	81.8%	55.2%	8.000%	1.361%	
	95 - 1	64.9%	65.1%	63.0%	75.0%	83.9%	83.9%	84.5%	83.9%	81.0%	56.4%	8.667%		
	90 - 10	63.3%	63.3%	62.4%	74.5%	84.9%	84.7%	85.4%	84.6%	82.0%	55.4%			
	80 - 20	62.6%	62.8%	62.8%	74.8%	83.8%	83.8%	84.4%	83.7%	79.6%				
	70 - 30	66.4%	66.2%	66.9%	78.2%	87.3%	87.8%	88.7%	87.1%					
	60 - 40	66.9%	67.1%	67.1%	79.1%	87.8%	88.6%	86.4%						
	50 - 50	61.3%	61.1%	61.1%	73.0%	83.4%	85.2%							
	40 - 60	59.3%	58.9%	59.5%	71.3%	80.7%								
	30 - 70	56.0%	55.3%	55.7%	66.4%									
	20 - 80	35.6%	34.7%	35.4%										
	10 - 90	4.0%	3.4%											
	5 - 95	0.0%												
	1 - 99													
			1 - 99	5 - 95	10 - 90	20 - 80	30 - 70	40 - 60	50 - 50	60 - 40	70 - 30	80 - 20	90 - 10	95 - 5

■ = Comparison with identical fraction concentrations - no shift (and thus no linkage phase) would be expected here.
 ■ = Inverse of another comparison on the opposite side of the diagonal (ie a comparison of 99-1 to 50-50 is the same as comparing the 50-50 to 99-1)

Executive Summary Table 2 – Average confidence for linkage phase determination by FLiPARS 2.0

% First Fraction - % Second Fraction	99 - 1	99.99998%	99.99995%	99.99942%	99.84686%	99.92842%	99.96270%	99.76725%	99.90528%	98.84425%	97.34274%	84.72230%	99.82093%	
	95 - 1	99.99982%	99.99983%	99.99936%	99.99783%	99.99640%	99.99577%	99.98365%	99.98722%	99.96453%	99.91154%	99.85614%		
	90 - 10	99.99924%	99.99921%	99.99866%	99.99783%	99.99711%	99.99571%	99.99254%	99.98314%	99.96331%	99.83391%			
	80 - 20	99.99783%	99.99780%	99.99739%	99.99644%	99.99495%	99.99224%	99.98307%	99.92734%	99.53825%				
	70 - 30	99.99706%	99.99697%	99.99304%	99.99486%	99.99150%	99.98350%	99.94751%	99.54715%					
	60 - 40	99.99586%	99.99583%	99.99554%	99.99097%	99.98378%	99.96048%	99.77990%						
	50 - 50	99.99316%	99.99303%	99.99179%	99.98276%	99.95835%	99.66834%							
	40 - 60	99.98540%	99.98524%	99.98239%	99.92625%	99.56661%								
	30 - 70	99.97075%	99.97001%	99.95906%	99.56379%									
	20 - 80	99.95411%	99.95220%	99.85811%										
	10 - 90	99.91946%	99.68254%											
	5 - 95	0.00000%												
	1 - 99													
			1 - 99	5 - 95	10 - 90	20 - 80	30 - 70	40 - 60	50 - 50	60 - 40	70 - 30	80 - 20	90 - 10	95 - 5

Casework-Type Samples: The performance of the FLiPARS 2.0 application was also tested on a variety of casework-type samples which consisted of:

- **Body Fluid Mixtures on Varied Substrates:** Eighteen two-component body fluid mixtures were stained onto a variety of substrates including denim, leather, wood, carpet, nylon and wallboard and aged at room temperature for four weeks.

- Environmental Insult Mixtures: Seventeen two-component body fluid mixtures were subjected to a variety of environmental insults including gasoline, soil, laundry detergent, used motor oil, sodium hydroxide and acetic acid.
- Hair and Bone Mixtures: Seven casework-type mixed samples were prepared from hair and bone samples mixed with semen or blood to simulate conditions likely to be encountered by forensic mtDNA practitioners working on challenging samples.

All samples were extracted, the HV1 and HV2 regions were amplified by standard forensically validated methods; fractionated by DHPLC and the resulting electropherograms were analyzed by FLiPARS 2.0 to determine the haplotypes of each individual contributor. The individual contributor haplotypes identified were in full concordance with sequencing results obtained from individual non-mixed reference samples. FLiPARS 2.0 yielded linkage phase analysis results with a high degree of average base resolution confidence (99.94% to 99.99% for body fluid mixtures on varied substrates; 99.80% to 99.99% for environmental insult mixtures; 99.77% to 99.99% for hair and bone mixtures).

Alignment Ambiguities: It was discovered during the validation phase that certain mutation patterns may cause alignment ambiguities and that strict adherence to published FBI alignment guidelines^[16] does not always produce an alignment that is biologically reasonable. After consulting with expert forensic practitioners in mtDNA analysis (Terry Melton of Mitotyping Technologies, Suni Edson of AFDIL, and Walther Parson of Institute of Legal Medicine, Innsbruck Medical University), the alignment algorithm of FLiPARS 2.0 was modified through the addition of an alignment algorithm plug-in to better reflect evolutionary processes and to thereby yield a more accurate alignment.

Implications for Policy, Practice and Future Research

A mixture of different mtDNA molecules in a single sample presents an often insurmountable challenge to successful sequence analysis. In a forensic context, mtDNA mixtures are most typically “situational”. In these cases, mtDNA from separate humans is found in association with a single evidentiary sample. Direct sequence analysis of such samples results in ambiguous base calls.

The developmental validation of linkage phase analysis as an accurate means of deconvolving mtDNA mixtures offers practitioners the opportunity to obtain potentially useful information from what might otherwise be uninterpretable samples. The approach has been tested, the underlying reasoning is scientifically valid and the potential error rates and limitations have been evaluated. The FLiPARS 2.0 software application developed and tested under the current DNA Research and Development Award (2009-DN-BX-K047) has automated what had been an extremely laborious and computationally-intensive process. The availability this software application serves to remove a significant and immediate obstacle to the use of this technology by forensic practitioners. As a result, forensic practitioners will now be able to readily ascribe specific haplotypes to the individual contributors to a mixture –and to do so with a reportable degree of statistical confidence. This expands the potential applicability of mtDNA testing to a broader range of evidentiary samples.

The current research program has been successfully completed in accordance with DAB developmental validation standards for sensitivity, reproducibility and accuracy. The future of this research, therefore, will depend on achieving three critical objectives. These are:

- 1) **Rigorous Interlaboratory Validation Studies by Practitioners** of the FLiPARS 2.0 program is essential to the continued development of features that will enable practitioners to add their own expertise in evaluating the validity of peak calls and alignments made by the software.
- 2) **Further Refinement the Software Application** to meet the needs of forensic practitioners is essential. A top priority will be development of improved tools that allow practitioners to use their own expertise in evaluating peak calls and alignments will require a continuing conversation between the developer and the practitioner community.
- 3) **Extend Validation to Additional Amplicons** in recognition of the fact that forensic practitioners often employ a variety of primer pairs in addition to those with which DHPLC was validated. Although these were not evaluated as part of the current research, it is anticipated that FLiPARS 2.0 will perform as well with these primer pairs as it did in the current study.

Literature Cited in the Executive Summary

1. Fournery, R., *Mitochondrial DNA and Forensic Analysis: A Primer for Law Enforcement*. Journal of the Canadian Society of Forensic Science, 1998. **31**(1): p. 45-53.
2. Holland, M.M. and T.J. Parsons, *Mitochondrial DNA Sequence Analysis-Validation and use for forensic casework*. Forensic Science Review, 1999. **11**(1): p. 22-49.
3. Wilson, M.R., *et al.*, *Validation of mitochondrial DNA sequencing for forensic casework analysis*. Int J Legal Med, 1995. **108**(2): p. 68-74.
4. Huber, C.G., P.J. Oefner, and G.K. Bonn, *High-resolution liquid chromatography of oligonucleotides on nonporous alkylated styrene-divinylbenzene copolymers*. Anal Biochem, 1993. **212**(2): p. 351-8.
5. Huber, C.G., P.J. Oefner, and G.K. Bonn, *Rapid and accurate sizing of DNA fragments by ion-pair chromatography on alkylated nonporous poly(styrene-divinylbenzene)*. Analytical Chemistry, 1995. **67**: p. 578-585.
6. Emmerson, P., *et al.*, *Characterizing mutations in samples with low-level mosaicism by collection and analysis of DHPLC fractionated heteroduplexes*. Hum Mutat, 2003. **21**(2): p. 112-5.
7. Etokebe, G.E., *et al.*, *Physical separation of HLA-A alleles by denaturing high-performance liquid chromatography*. Tissue antigens, 2003. **61**(6): p. 443-50.
8. Hanekamp, J.S., W.G. Thilly, and M.A. Chaudhry, *Screening for human mitochondrial DNA polymorphisms with denaturing gradient gel electrophoresis*. Hum Genet, 1996. **98**(2): p. 243-5.
9. Steighner, R.J., *et al.*, *Comparative identity and homogeneity testing of the mtDNA HVI region using denaturing gradient gel electrophoresis*. J Forensic Sci, 1999. **44**(6): p. 1186-98.
10. Barros, F., *et al.*, *Rapid and enhanced detection of mitochondrial DNA variation using single-strand conformation analysis of superposed restriction enzyme fragments from polymerase chain reaction-amplified products*. Electrophoresis, 1997. **18**(1): p. 52-4.
11. Danielson PB, *et al.*, *Resolving mtDNA mixtures by denaturing high-performance liquid chromatography and linkage phase determination*. Forensic Sci Int Gen 2007;1(2):148-53.
12. Danielson, P.B., *et al.*, *Separating human DNA mixtures using denaturing high-performance liquid chromatography*. Expert Rev Mol Diagn, 2005. **5**(1): p. 53-63.
13. Danielson, P.B., *Mitochondrial DNA Analysis by Denaturing Liquid Chromatography for the Separation of Mixtures in Forensic Samples*. Final Technical Report for 2003-IJCX-K104 submitted to the National Institute of Justice, Forensic DNA Research and Development Program. 2008. pp. 105.
14. Troelsen, A., *Pro C# 2008 and the .NET 3.5 Platform (Windows.Net)*. Fourth Edition ed. 2007 Berkely, CA Apress.
15. www.microsoft.com/netframework [cited].
16. Federal Bureau of Investigation "Further Discussion of the Consistent Treatment of Length Variants in the Human Mitochondrial DNA Control Region" *Forensic Science Communications* October 2002. 4 (4).

FINAL TECHNICAL REPORT (MAIN BODY)

Introduction and Statement of the Problem

The quantity and quality of DNA are critical factors in forensic investigations. When the use of short tandem repeat (STR) nuclear loci^[1] fails, however, mitochondrial DNA (mtDNA) often offers investigators the only remaining opportunity to obtain potentially probative genetic information^[2]. Mitochondrial DNA analysis has often been used with especially challenging samples. For example, it has been used to identify severely weathered/degraded remains from the Vietnam War^[3], Czar Nicholas II^[4] and murder victims^[5] as well as to provide useful forensic information on shed head hairs and saliva from robbery caps^[6]. It is also frequently used in cases such as plane crashes where remains may be exposed to conditions that compromise DNA quality. Its exonuclease-resistant nature and the presence of up to several thousand copies of mtDNA per cell facilitates the analyses of degraded and/or low-copy number material^[1-7]. Additionally, the uniparental inheritance of mtDNA, allows reference material to be obtained from maternal relatives^[8].

Analysis of mtDNA is has traditionally been accomplished (almost exclusively) by sequencing the DNA from hypervariable regions 1 and 2 (HV1/HV2) of the control region. This approach has been rigorously validated and has withstood several court challenges (see www.denverda.org for specific case law examples).

Although protocols for sequencing mtDNA are well established, the presence of a mixture of different mtDNA molecules in a single sample can present a significant obstacle to successful mtDNA analyses by standard methods. In fact, such a “situational” mixture of two or more individuals or even naturally occurring heteroplasmic mixtures, typically preclude successful mtDNA analysis^[2, 9, 10]. This roadblock occurs because sequencing a mixture of two or more DNA amplicons yields electropherograms characterized by overlapping peaks at sites where the amplicons differ in sequence. Because peak height is sequence context dependent, it cannot be used by itself to determine the absolute or even relative quantities of DNA from individual contributors to the mixture. This can impede the forensic use of mtDNA.

Underscoring the fact that this is not a minor problem is the observation from extensive casework records that a significant proportion of evidentiary hairs examined were heteroplasmic (11.4%) or displayed a mixed profile (8.7%). Moreover, the occurrence of mixed mtDNA profiles appears to increase with the age of a sample and is usually not ameliorated even following extensive validated cleaning methods^[11]. This likely represents only the “tip of the iceberg” since samples which are suspected as likely to yield mixtures are often not even submitted for analysis. A reliable means of resolving the individual sequences within a mixture could greatly aid investigators by increasing the range of casework samples suitable for mtDNA testing.

There are a number of established molecular strategies that could be employed to separate DNA mixtures into their individual components. These include separation by denaturing gradient gel electrophoresis (DGGE)^[12] or single-strand conformational polymorphism (SSCP) analysis^[13-15] and subcloning into bacterial vectors. These approaches are generally time consuming, necessitate multiple handling steps, require laborious product purification and are not readily adaptable to automation. These factors have all been obstacles to the implementation of these technologies by forensic laboratories. Both DGGE and SSCP require manual recovery of fractionated DNA from polyacrylamide gels and a second round of PCR amplification to

generate enough template DNA for sequencing. Subcloning is an even more time and labor-intensive approach. It would require forensic scientists to screen and sequence DNA from multiple transformed bacterial colonies to ensure that observed sequence differences reflect genuine differences in the starting template rather than artificial variants introduced as a result of nucleotide misincorporation during PCR.

More recently, both pyrosequencing^[16,17] and electrospray ionization-mass spectrometry^[18] have attracted significant interest as means of resolving mtDNA mixtures. Both approaches have the significant advantage of being able to more precisely quantitate heteroplasmic and other mixed samples than earlier nucleic acid analysis methods. In the case of mass-spectrometry, it is possible to take advantage of the fact that the exact mass of each deoxyribonucleotide is a precisely known value. Accordingly, it becomes a relatively straight forward matter to generate a constrained list of possible deoxyribonucleotide combinations that could account for the total mass of a given PCR amplified fragment. Therein, however, is rooted the most significant limitation of a mass-spectrometry based approach. While it is possible to predict the base composition of an assayed fragment, it is not possible to know with absolute certainty the precise base sequence of a fragment. Given the frequency with which new mtDNA variants are reported, it is not inconceivable that fragments having identical base compositions can have different sequences. Furthermore, mass spectrometry approaches based on post amplification restriction digests necessitate additional sample handling while those based on tiled sets of PCR primers^[19] require multiple PCR reactions that in some cases may require the consumption of more of a precious DNA extract than is available.

Pyrosequencing, on the other hand does allow an analyst to determine the actual base sequence of an amplified fragment and it can provide some quantitative information on the composition of a mixture^[20]. This approach, however, is limited by the fact that read lengths are often shorter than by traditional dideoxynucleotide sequencing; homopolymeric sequences are difficult to read and the interpretation of mixtures is complex - typically necessitating manual interpretation to ensure accuracy. While both approaches are certainly promising, dideoxynucleotide sequencing remains the method of choice for mtDNA analysis in most forensic laboratories that process such samples.

Denaturing High-Performance Liquid Chromatography (DHPLC)^[21,22] is a chromatographic means of fractionating natural (heteroplasmic) or situational (multi-contributor) DNA mixtures prior to sequencing^[23, 24]. In contrast to the alternative approaches that have been proposed for the separation of mtDNA mixtures^[12, 13, 19, 20, 25, 26], DHPLC does not require secondary amplification or excessive sample manipulation to resolve a mixture of different mtDNA haplotypes. Furthermore, it is completely consistent with established forensic SOPs for standard fluorescence-based sequencing. Under a previous NIJ DNA Research and Development Award (2003-IJX-K104) this approach to resolving mtDNA mixtures was rigorously validated using both reference and casework type samples^[27-29]. These studies demonstrated: (1) the sensitivity of DHPLC to detect and fractionate mixtures involving all classes of polymorphisms; (2) the reproducibility and statistical correlation between peak height ratios at mixed base positions and DNA quantity ratios in sequencing electropherograms; (3) the reliability of determining individual haplotypes by linkage phase analysis of sequence data from DHPLC fractionated samples. Based on the results of these studies, standard operating procedures and statistically-grounded interpretation guidelines for the use of DHPLC to resolve mtDNA mixtures were developed.

Through careful quantitative analyses of sequencing electropherograms, changes in the relative heights of overlapping fluorescent peaks at all mixed-base positions can be traced across two or more DHPLC fractions. The observation of coordinated shifts in relative fluorescence ratios for a given set of nucleotides was shown to be consistent with them being in the same linkage phase and thus associated with the same amplicon (*i.e.*, contributor). While successful, this approach required computationally intensive analyses of enormous datasets. This was found to be too large of a task for an analyst to manage “in their head”. Accordingly, a prototype software tool termed FLiPARS 1.0 (Fractional Linkage Phase Analysis Resource System) was developed to automate linkage phase analyses. This early “prototype” version was a conglomeration of Perl scripts and dynamically linked Excel files integrated into a Microsoft Windows Visual Basic application. When the prototype worked, it worked well and showed the tremendous potential for the underlying forensic technology it supported. It yielded a linkage phase report spreadsheet which included a listing of the mixed-base position analyzed, the amount of fluorescence shift between DHPLC fractions, the linkage groups of the bases for each contributor, the statistical confidence of the individual base assignments, and the number of samples on which the statistical estimates were based (Figure 1).

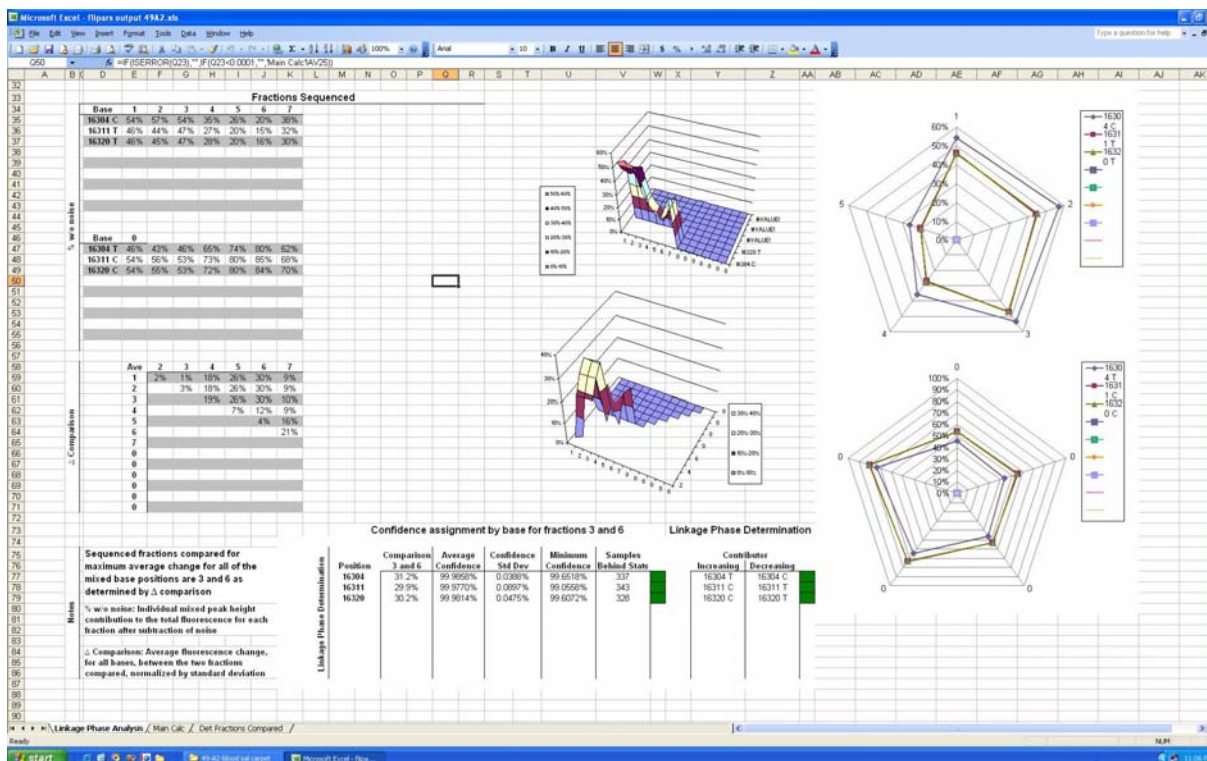


Figure 1: The linkage phase analysis and results table generated using FLiPARS 1.0 software was generated in the context of a Microsoft Excel spreadsheet. Displayed are the tables and their graphical representations of data used to determine the linkage phase of the individual contributors to a mixture. Listed in the linkage phase determination table (bottom center of the figure) are the mixed-base positions, the amount of fluorescence shift between selected DHPLC fractions, statistical confidence parameters of the linkage phase base assignments and the linkage phase results defining each contributor.

FLiPARS 1.0, however, was built more as a proof of concept application than as a production ready software solution for use in forensic laboratories. Thus, it is not surprising, that a number of problems were identified with this prototype. For example, the graphical user

interface was frequently found to be confusing by novice users; it was necessary to convert the .scf files exported from ABI sequencers to tab-delimited .txt files before they could be analyzed and; some seemingly correct files cause FLiPARS 1.0 to hang, inexplicably crash, or to produce anomalous results leaving the user to determine what went wrong.

Thus, even though DHPLC and linkage phase analysis had been thoroughly validated for the resolution of mtDNA mixtures, the lack of a reliable software application to handle the computational end of linkage phase analysis represented a significant obstacle. Without a suitable solution, practitioners were justifiably reluctant to adopt or even evaluate what was an otherwise powerful technology for resolving mtDNA mixtures.

Statement of Core Research Objectives

The central goal of the research funded under DNA Research and Development Award 2009-DN-BX-K047 was to develop and test a software application to automate the computationally intensive analysis of electrophoretic data that is necessary to determine the linkage phase (*i.e.*, haplotypes) of individual contributors to an mtDNA mixture. This was achieved through the completion of three major research objectives. These were:

- (1) **Develop a robust software application and user/friendly graphical interface** that could be used to import and analyze sequencing electropherogram files from fractionated mtDNA mixtures and to then report the linkage phase (*i.e.*, haplotype) of the individual contributors to the mixture.
- (2) **Test the accuracy of the software application** on multiple electrophoretic data sets generated from mtDNA mixtures. The data sets used represented a broad range of stepwise mtDNA mixture ratios and >200 mixed base positions. These data sets were generated from both reference and casework-type samples.
- (3) **Rigorously analyze the performance and accuracy of the software application** and make appropriate revisions to the software so as to effectively resolve any anomalies detected in the linkage phase determination results.

The successful completion of these objectives was designed to aid forensic practitioners by providing them with a reliable and easy to use means of automating the process of linkage phase-based determination of individual haplotypes in mtDNA mixtures. Ultimately, this would provide practitioners with the opportunity to evaluate and adopt an efficient means of successfully analyzing some types of challenging samples.

Methods

Human Subjects – The University of Denver (DU) Institution review Board for Research Involving Human Subjects (IRB) reviews all research involving human subjects, regardless of funding source, to ascertain that the rights and welfare of subjects are being protected. The IRB is responsible for assuring that recruitment advertising is not misleading or coercive to the research subject. All projects using human subjects are reviewed no less than annually.

All research employing human-derived mtDNA sequence data under DNA Research and Development Award 2009-DN-BX-K047 was IRB reviewed, approved and conducted in full compliance with U.S. Federal Policy for the Protection of Human Subjects (Basic DHHS Policy

for Protection of Human Research Subjects; 56 FR 28003). Electrophoretic sequence data representing the mtDNA haplotypes of 96 adult (>18 y.o.) human volunteers were employed for this study. All volunteers had been previously recruited from within the DU student population as part of a previous NIJ DNA Research and Development Award (2003-IJX-K104). Recruitment notices were posted in campus science buildings to attract interested volunteers. The student traffic in these buildings consisted primarily of science-oriented graduate and undergraduate students. The purpose and significance of the research and the methods that were used to collect mtDNA samples was thoroughly explained to each volunteer. All participants signed a statement of informed consent to participate in the research. As no health care associated information was collected, HIPPA authorization was not required.

Casework Type mtDNA Mixtures – Evaluations of the software performance on non-pristine samples employed mtDNA sequence files generated from casework-type material. This included the analysis of mtDNA sequence electropherograms representing a variety of mixed tissue sources (blood, semen, saliva, hair shafts and bone), deposited on a variety of substrates (cotton cloth, nylon carpeting, blue denim, leather, cigarette butts and wall board) and exposed to a range of environmental contaminants/inhibitors (gasoline, used motor oil, soil, laundry detergent, acetic acid and sodium hydroxide).

Programming language – The prototype program, FLiPARS 1.0, was written in Visual Basic and was partially based on Microsoft’s .NET framework, which featured a large library of built-in functionality. FLiPARS 1.0, however, also employed older Microsoft technologies that have been phased out in favor of the new .NET paradigm. While the prototype application was designed to reduce programmer error and to speed product development, the internal code was not strongly typed meaning that the program was often expected to ‘guess’ with regard to the types of data that were expected. This resulted in significant overhead to its execution which slowed its overall performance. For the development of FLiPARS 2.0, C# (C-Sharp)^[30], a more modern relative of Visual Basic was used as the programming language. This was done to facilitate the stable design and future adaptability of the linkage phase software application. The use of C# allowed FLiPARS 2.0 to be written to take advantage of “strong typing” which improved program stability by requiring only a specified data type and format to be used. Strong typing also improved program security by minimizing the potential for malicious activity (*i.e.*, hacking) that might result from attempts to input data that weren’t explicitly accounted by the code. The use of the C# programming language facilitated the development of an application that was simple to use, reliable and will be malleable for growth in response to suggestions and feedback from forensic practitioners.

Software Forward Compatibility – FLiPARS 1.0 was written using outdated Windows development constructs (development software) that required extensive code modifications to enable it to work under the non-Windows XP operating systems. In order to ensure the sustained utility of the FLiPARS 2.0 software application, it was necessary to make it functional on both alternative (*i.e.*, non-Windows) and new operating systems. This was achieved by taking advantage of the Microsoft’s .NET paradigm – a powerful and robust set of tools that have been used to enable software developers to work across a number of programming languages. The .NET paradigm is platform independent and this has enabled the FLiPARS 2.0 program to be run on a number of different operating systems (*e.g.*, Windows XP, Windows Vista, Windows 7, Mac OS X and Linux). This helped to ensure that the software maintenance requirements were be extremely minimal.

Eliminating Dependencies on Secondary Software – The FLiPARS 1.0 prototype, made it very difficult and cumbersome to modify the program code to handle various file input types. In fact, its file input was tied directly into the program’s underlying algorithms, making it difficult to modify without breaking the actual functionality of the software. The C# language provided a robust feature set that was employed across the entire application, from the interface to the underlying statistical analyses^[31]. This streamlined approach eliminated foreign dependencies, such as the use of Microsoft Excel from the FLiPARS 2.0 version of the software. Specifically, secondary software-associated dependencies were eliminated through the use of modularized plug-ins that were used to convert data into a standard format that FLiPARS 2.0 could “understand”. This standard interface format has been made freely available thereby enabling any future developer to produce a plug-in for FLiPARS. Any plug-in that implements this interface can be used without impacting the underlying operation of the software. This has not only made it easier to store and retrieve sequence information from many sources, including sequence databases and the internet, but it has also ensured that new version of FLiPARS will remain viable for years to come. The .NET paradigm, was also used to alleviate the potential intellectual property concerns of sequencer manufacturers regarding proprietary data formats ported from their instruments. *In toto*, this has enabled the development of a faster, simpler and easier to maintain application^[32, 33].

Software Performance and User Interface – The FLiPARS 1.0 prototype was extremely slow and processing functions were not accompanied by any indication of the actual progress being made. These problems were resolved through the use of “strong typing” and a single programming language for all of the program’s functionalities.

The original graphical user interface (GUI) for FLiPARS 1.0 was found to be unnecessarily confusing and provided the practitioner with few useful guides to walk them through the process of linkage phase analysis. It assumed, instead, that the user was already familiar with the software and underlying jargon used for linkage phase analysis. This was acceptable when the only users of the application were those in the lab where the software was originally created. In order to create a more user-friendly GUI, however, the prototype FLiPARS 1.0 GUI was repackaged into a more intuitive form. This was designed to enable a practitioner with minimal training to use the software with confidence. In order to achieve this, ambiguous buttons and terms were removed and the intermediate outputs of multistep algorithms were unified so that only the final output was displayed. Also, the “Peak Stringency” and “Noise Threshold” dialogs were moved out of the main window and into an “Advanced Settings” tab. Finally, the terminal output of each analysis was limited to a single top-level screen from which the unnecessary machinations of the underlying computer code had been removed.

Sequence Alignment Algorithm – Each sequence was aligned against the entire revised CRS using a dynamically generated table (Smith-Waterman approach). The technique was found to work reasonably well although time and resources are necessary for the generation and traversal of the alignment table. Alternative algorithms are available, but stray away from the strict edit-distance definitions of alignments, and as such the decision was made to not use these approaches.

Validation of the Software Application’s Accuracy – The accuracy of the linkage phase analysis software application developed under DNA Research and Development Award 2009-DN-BX-K047 was evaluated by analyzing a pre-existing dataset of sequencing electropherograms generated from mtDNA mixtures. These were the same sequences generated from the mtDNA

mixtures that were used to validate the accuracy and reliability of the linkage phase analysis itself as a means of resolving mtDNA mixtures. The dataset consisted of a series of pristine two-contributor mixtures of mtDNA amplicons. These were combined at stepped ratios from 1:99 to 99:1. Collectively the mixed samples differed in sequence at >200 base positions distributed throughout the HV1 and HV2 regions. The applicability of FLiPARS 2.0 for use by forensic practitioners, however, necessitated the use of a second more rigorous set of validation studies. These included the analysis of casework-type mtDNA mixtures involving a variety of tissue sources, substrates, and environmental contaminants/inhibitors. For each mixture (pristine and casework-type) the haplotypes of HV1 and HV2 were known for each contributor.

The FLiPARS 2.0 software was used to resolve the mixed sequence electropherograms and the results obtained were compared to the known haplotypes for the sample. These analyses were conducted first without user confirmation of peak calls and then with the inspection of the electrophoretic traces to identify those mixed base positions that might have been missed or miscalled by the software under the default peak detection parameters.

Version Control and Licensing – The code for FLiPARS 2.0 was placed under version control using “Apache Subversion™”. Subversion is an open source version control system that allows for detailed tracking of problems and fixes for those problems. It also provides a convenient way for other developers to access the source code and to simultaneously work on the project should that need arise. Additionally, it provides access to the FLiPARS 2.0 code wherever there is an internet connection.

FLiPARS 2.0 is licensed under the Apache 2.0 License, providing free access to the source code for virtually any purpose. Developers can modify and distribute the software as they see fit, so long as the appropriate licensing matters are followed. Modifications or additions to FLiPARS may be published under any license - again assuming that the original portions of the software retain and follow the Apache 2.0 Licensing guidelines.

Results and Discussion

FLiPARS 2.0 Platform Interoperability: A software application (FLiPARS 2.0) which automates the computationally intensive process of mtDNA mixture deconvolution by linkage phase analysis has been written in C#. The use of this programming language in combination with the .NET paradigm has produced an application with a high degree of compatibility across multiple operating systems without the need for modifications to the source code. Demonstrating the solid compatibility of the application was the observation that it could be run unmodified on Linux systems (Figure 2). This is impressive given the fact that the underlying machinations of the operating system vary to a great degree between Windows and UNIX-like systems such as Linux and Mac OS X. Unmodified, the current release of FLiPARS 2.0 has been successfully run on the following operating systems:

- Windows XP Professional (32 bit)
- Windows Vista Professional (64 bit)
- Windows 7 Professional (32 bit and 64 bit)
- Slackware Linux 13.0 w/ Gnome Slackbuild and Mono

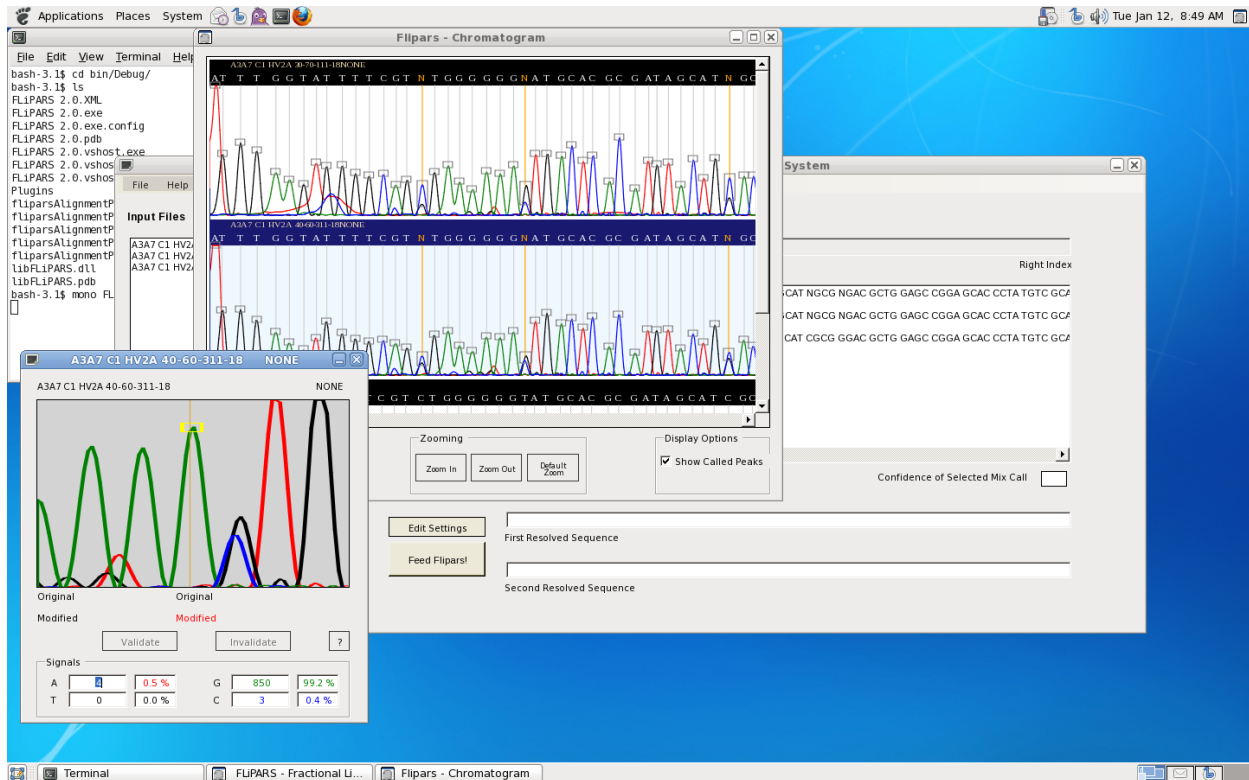


Figure 2 - Screenshot of FLIPARS 2.0 running in an unmodified state on Slackware 13.0 Linux with Mono installed

Modularized Plug-Ins – Dependencies on secondary software applications that were a feature of the prototype software application for linkage phase analysis have been replaced by a set of modularized plug-ins. The plug-ins themselves are restricted to modifying how the program obtains data from outside sources. They include support for different file formats and determine how the input sequences are parsed and then aligned. The plug-ins that have been implemented allow for the parsing of three different file formats which are commonly used to save sequence data. These plug-ins provide support for raw tab-delimited text files as well as the more common .ABI and .SCF file formats used by the Applied Biosystems instruments which are found in nearly all forensic laboratories that perform mtDNA testing.

Intuitive Graphical User Interface – A new FLIPARS 2.0 interface has been implemented. The main screen was designed to present the user with a file input box on the left for their files and the expected “add”, “remove”, “clear” and “reorder functionality” of other typical data input boxes (Figure 3). The associated primers for each input file can be set individually as can the reference sequence that the program uses to align the input sequences. This provided the program with compatibility for non-forensic applications as well as flexibility in the event that new technologies emerge or different portions of the mtDNA genome are targeted in the future.

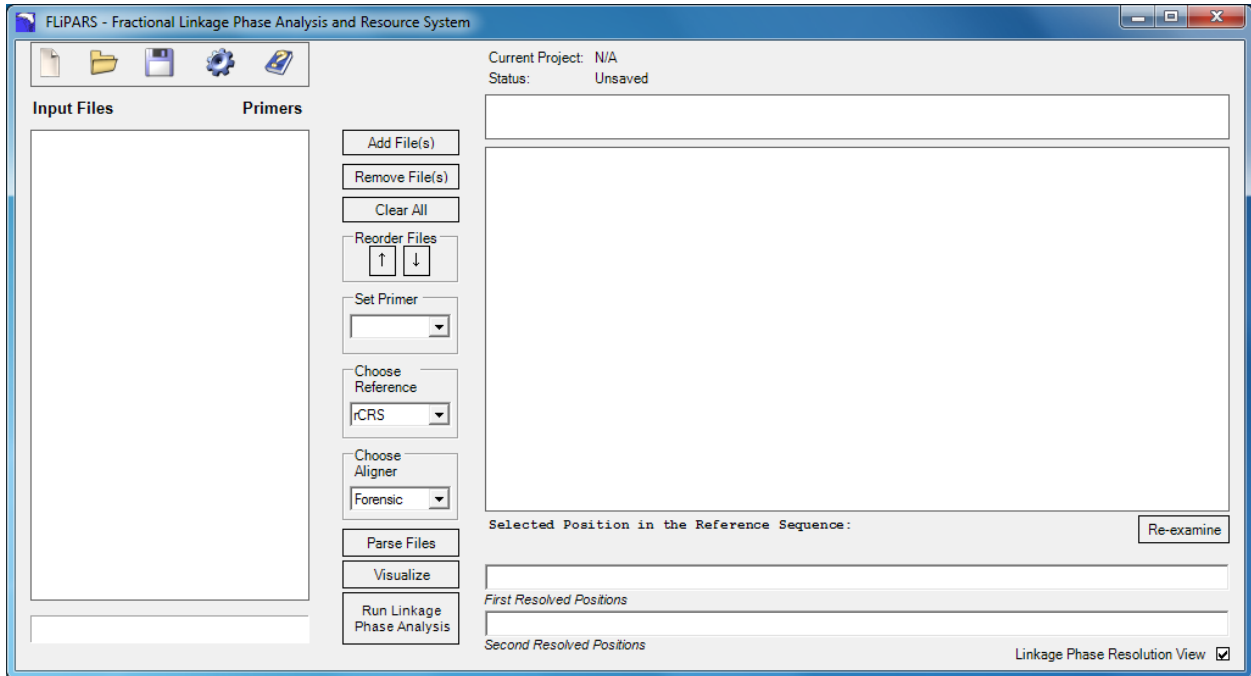


Figure 3 - FLiPARS 2.0 revamped user interface as shown to the user when it is first opened.

A status bar has also been implemented (Figure 4) to apprise the user of the status of any long running operations that the program may be performing in the background while they wait. Additionally, the file being processed and/or task being performed are listed below the status bar.

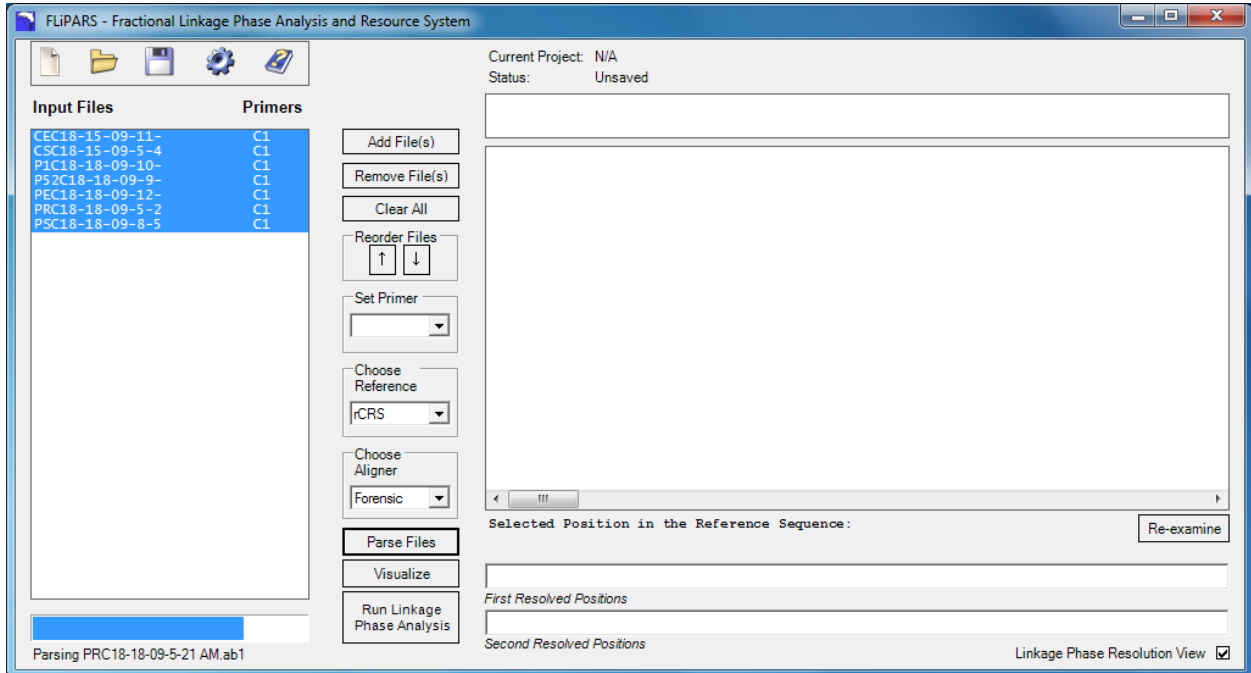


Figure 4 - FLiPARS status bar showing the progress made in processing a small batch of files.

Once processed by the internal FLiPARS 2.0 base calling algorithm, the user is able to view the corresponding chromatograms and base calls by clicking on the "Visualize" button. This

brings up a display containing all of the chromatograms input to the software. If needed, the software will also (re)process the underlying data. Each chromatogram is offset by different colored bars and backgrounds (Figure 5) to ensure that the user does not confuse calls between adjacent chromatograms.

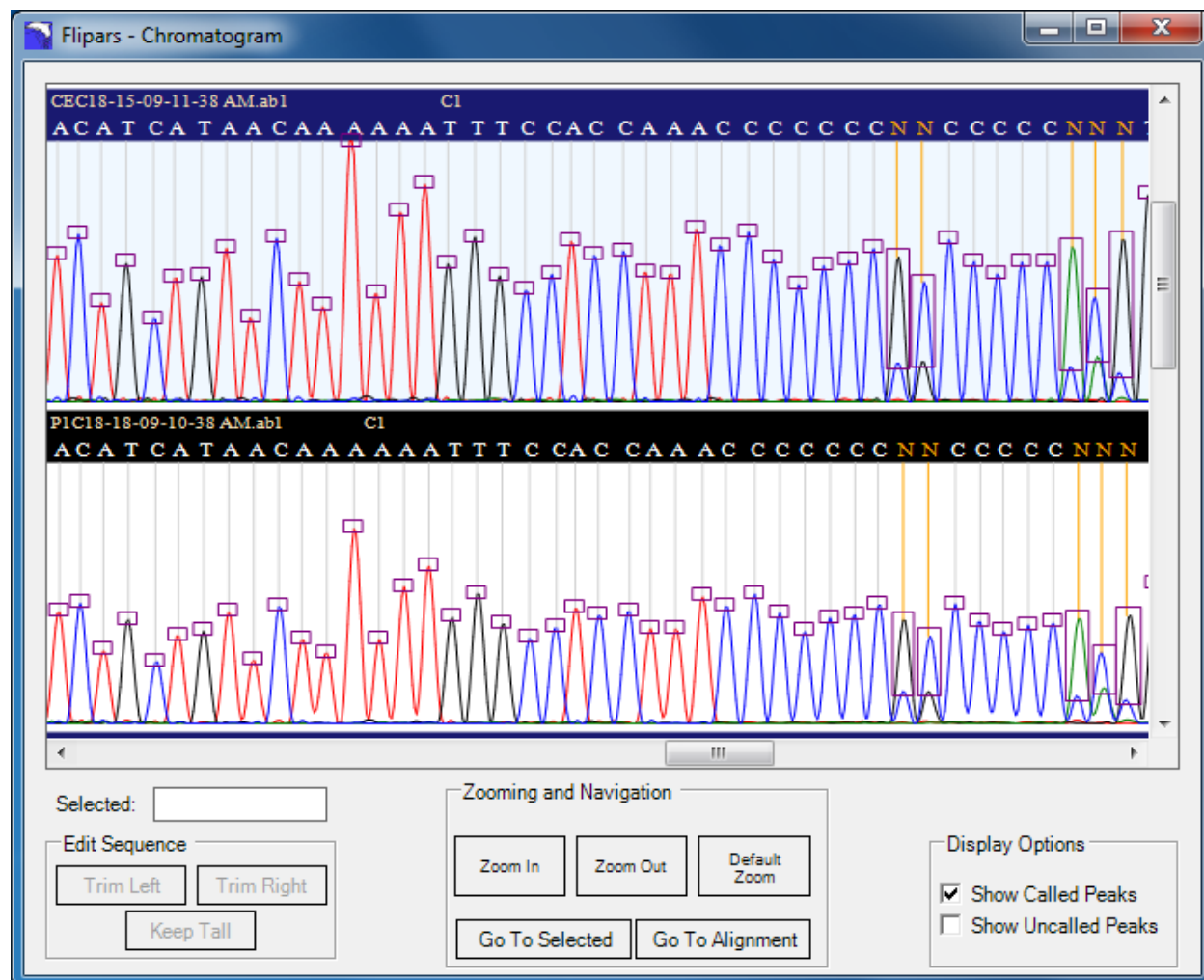


Figure 5 - FLiPARS 2.0 chromatogram display at the default zoom view. This display clearly shows mixed base calls highlighted by orange lines at the peak apex and basecall boxes that encompass the peaks of both bases. It also provides the user with the file name and primer used (as specified in the main display window) for each sequence.

A settings window has been implemented to allow user-specified modification of the underlying analysis settings and internal data sets (Figure 6). Since these are expected to be rarely (if ever) modified by the end user, they have been contained in their own separate window which can be reached from the main screen by clicking on the “Settings” icon. Help boxes (denoted by buttons containing “?”) have been provided to provide details on each particular setting should the user require more information. The primer modification box also displays graphically whenever a primer has been modified and indicates whether or not any changes have been saved. New plug-ins may also be easily added to the program through the FLiPARS 2.0 plug-in installation dialog window (Figure 7). All saved settings are maintained between FLiPARS 2.0 sessions (*i.e.*, after closing and reopening the program), so a lab using a custom set

of primers will only need to enter those primers into the program once. Peak stringencies and noise thresholds (which relate to the base-calling algorithm internal to FLiPARS) can also be changed. Finally, reference sequences may be added (but not modified) to the program (Figure 8).

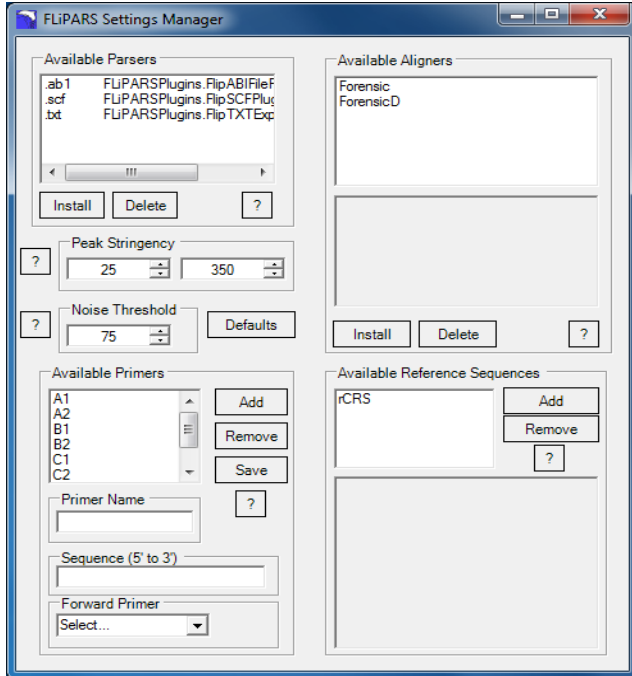


Figure 6 - FLiPARS 2.0 Edit Settings dialog. Options to modify nearly all relevant underlying information are accessible here. The ability to add or remove any needed primers or reference sequences gives FLiPARS a large degree of compatibility for future standards. Changes made to primers are displayed as “Unsaved” to the user.

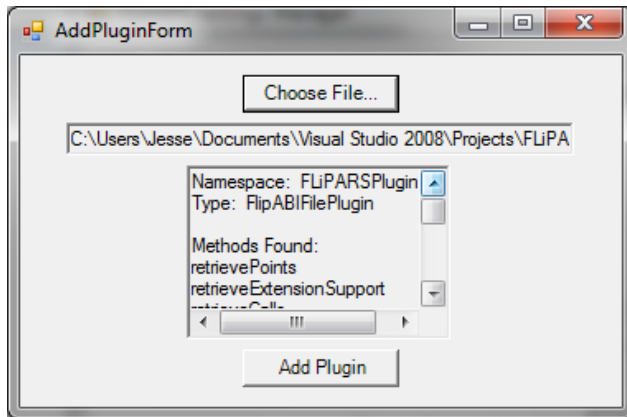


Figure 7 - FLiPARS 2.0 plug-in installation dialog. The user points the program to the *.dll file to be installed and the program checks it for validity before making any permanent changes to the system. All that is required is a working plug-in for FLiPARS to automatically make use of the appropriate tools to open and parse the respective file types based upon their file extension.

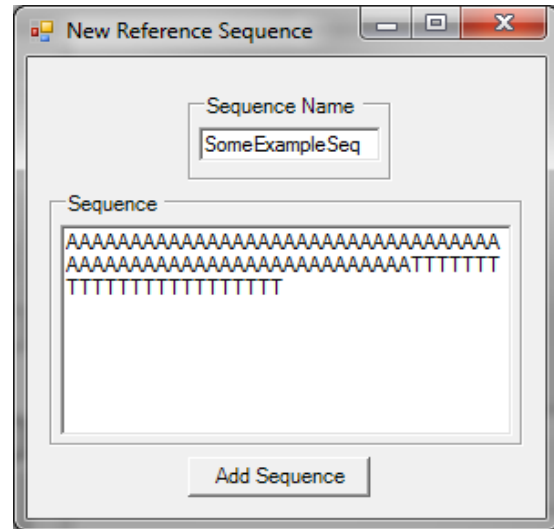


Figure 8 - FLiPARS 2.0 new reference sequence dialog prompts the user to enter a distinguishing name for the sequence as well as the plain-text (ATGC) nucleotide sequence. Invalid entries (letters other than A,T,G or C) are disallowed, helping to minimize potential user error.


The base calling and sequence alignment algorithms have been designed to enable the forensic practitioner to manually modify the sequence being analyzed. The ability of a practitioner to be able to modify the sequence after examining the raw data is an important aspect of the program as no base calling algorithm can claim perfect accuracy or substitute for the measured judgment and experience of a skilled practitioner. All changes are tracked by the application and the original unedited peak calls can be readily restored.


FLiPARS 2.0 Performance Testing – The accuracy of the software application to deconvolve mtDNA mixtures was determined by analyzing a large dataset of mixed sequence electropherograms (up to 11,581 comparisons of mixed sequence electropherograms). These mixtures each consisted of two mtDNA amplicons that were combined at stepped ratios from 1:99 to 99:1. Collectively the mixed samples differed in sequence at >200 base positions distributed throughout the HV1 and HV2 regions. Sequences were grouped into 155 separate sets of stepped mixture ratios. Each set consisted of triplicate assays of two contributors at known concentrations of each contributor’s DNA. The molar ratios of DNA within each set were 1:99, 5:95, 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, 90:10, 95:5 and 99:1. These were used for pairwise linkage phase assays (Table 1). The sequencing files (.AB1 format) associated with each set of mixtures were added to unique FLiPARS 2.0 projects, parsed, and the haplotypes of each contributor were determined. For first pass automated linkage phase analysis, sequences were automatically trimmed to focus on the specific regions of HV1 and HV2 which are used by forensic practitioners to report mtDNA haplotypes. For HV1 this included bases 16024 - 16,365 and for HV2 this included bases 73 - 340. Linkage phase analysis was then performed, the respective contributor haplotypes were determined and the results were compared to the sequence of the respective contributors. Project files, which included analyzed data and linkage phase information, were saved for each set. Exports in the form of .XML files were also saved to facilitate the automated analysis of the linkage phase results. The .XML files contained information on the linkage phase analysis of the mixed-base positions present in the analyzed data set, including every possible unique pairwise comparison of fractions; their associated peak shifts; and mixture deconvolution results (*i.e.*, contributor HV1 and HV2 haplotypes).

Table 1 - Total number of linkage phase analyses by FLiPARS 2.0

% First Fraction - % Second Fraction	99 – 1	148	147	147	146	149	147	149	146	148	145	150	147	
	95 – 5	148	149	146	148	149	149	148	149	147	149	150		
	90 – 10	150	150	149	149	152	150	151	149	150	148			
	80 – 20	147	148	145	147	148	148	147	147	147				
	70 – 30	149	148	148	147	150	148	150	147					
	60 – 40	148	149	146	148	148	149	147						
	50 – 50	150	149	149	148	151	149							
	40 – 60	150	151	148	150	150								
	30 – 70	150	150	149	149									
	20 – 80	149	150	147										
	10 – 90	149	148											
	5 – 95	150												
	1 – 99													
		1 – 99	5 – 95	10 – 90	20 – 80	30 – 70	40 – 60	50 – 50	60 – 40	70 – 30	80 – 20	90 – 10	95 – 5	99 – 1
	% First Fraction - % Second Fraction													

Count 11581

 = Comparison with identical fraction concentrations - no shift (and thus no linkage phase) would be expected here.

 = Inverse of another comparison on the opposite side of the diagonal (ie a comparison of 99-1 to 50-50 is the same as comparing the 50-50 to 99-1)

Without manual editing of sequence alignments by a human analyst, the base calling and sequence alignment algorithms implemented in FLiPARS 2.0 were able to resolve nearly 70% of all aligned sequences (Table 2). Of sequences which were not resolved, nearly all involved length variants characterized by stretches of mixed bases at nearly every position. The deconvolution of such “out-of-register” sequence by FLiPARS 2.0 often required extensive manual editing. It should be noted that when practitioners encounter such out-of-register sequence, their only option for interpretation is to examine the sequence from the opposite direction. Using this strategy, FLiPARS 2.0 can also readily make an accurate linkage phase determination.

The second class of alignments that were not amenable to deconvolution by FLiPARS 2.0 were those where only a minimal shift in peak height ratio was present – *e.g.*, comparisons between a 99:1 and a 95:5 mixture. A third and smallest class of sequences that contained deconvolution errors were those that included mixed-base positions where fluorescence peak height ratios deviated significantly from the underlying molar ratio of DNA. This phenomenon, while rare, occurs reproducibly at known positions in HV1 and HV2^[34]. It results in single base drop-out from what should be a mixed position when the minor contributor accounts for less than 24% of the mixture. In cases where mixtures were successfully deconvolved, the statistical confidence associated with the linkage phase determination typically exceeded 99.9% for each base (Table 3).

While the ability of FLiPARS 2.0 to deconvolve many mtDNA mixtures without the assistance of a practitioner is encouraging, it should be emphasized, however, that this software application was not designed to replace forensic practitioners. Rather it is a tool to assist the practitioner in resolving mtDNA mixtures by automating the computationally intensive process of measuring coordinated shifts in peak height ratio at mixed base positions and then determining the statistical confidence that the observed shifts represent actual changes in the underlying DNA ratios. Most importantly, however, successful mixture deconvolution of sequence traces approached 100% for files that were first examined by a human analyst and edited based on their experience. Typical manual edits to sequence files included merging mixed bases whose peaks occurred too far from one another to have been automatically called as a mixed base position and elimination of false positive base calls and false positive mixed position calls from especially noisy sequence electropherograms. Here too, the statistical confidence associated with the linkage phase determination of successfully deconvolved mixtures typically exceeded 99.9% for each mixed base position.

Table 2 - Percentage of successful linkage phase analyses by FLiPARS 2.0 using raw (i.e., not practitioner edited) sequence electropherograms.

% First Fraction - % Second Fraction	99 - 1	64.9%	64.6%	62.6%	74.7%	84.6%	85.0%	85.9%	84.2%	81.8%	55.2%	8.000%	1.361%	
	95 - 1	64.9%	65.1%	63.0%	75.0%	83.9%	83.9%	84.5%	83.9%	81.0%	56.4%	8.667%		
	90 - 10	63.3%	63.3%	62.4%	74.5%	84.9%	84.7%	85.4%	84.6%	82.0%	55.4%			
	80 - 20	62.6%	62.8%	62.8%	74.8%	83.8%	83.8%	84.4%	83.7%	79.6%				
	70 - 30	66.4%	66.2%	66.9%	78.2%	87.3%	87.8%	88.7%	87.1%					
	60 - 40	66.9%	67.1%	67.1%	79.1%	87.8%	88.6%	86.4%						
	50 - 50	61.3%	61.1%	61.1%	73.0%	83.4%	85.2%							
	40 - 60	59.3%	58.9%	59.5%	71.3%	80.7%								
	30 - 70	56.0%	55.3%	55.7%	66.4%									
	20 - 80	35.6%	34.7%	35.4%										
	10 - 90	4.0%	3.4%											
	5 - 95	0.0%												
	1 - 99													
			1 - 99	5 - 95	10 - 90	20 - 80	30 - 70	40 - 60	50 - 50	60 - 40	70 - 30	80 - 20	90 - 10	95 - 5

= Comparison with identical fraction concentrations - no shift (and thus no linkage phase) would be expected here.
 = Inverse of another comparison on the opposite side of the diagonal (ie a comparison of 99-1 to 50-50 is the same as comparing the 50-50 to 99-1)

Table 3 - Average confidence for linkage phase determination by FLiPARS 2.0

% First Fraction - % Second Fraction	99 - 1	99.99998%	99.99995%	99.99942%	99.84686%	99.92842%	99.96270%	99.76725%	99.90528%	98.84425%	97.34274%	84.72230%	99.82093%	
	95 - 1	99.99982%	99.99988%	99.99936%	99.99783%	99.99640%	99.99577%	99.98365%	99.98722%	99.96453%	99.91154%	99.85614%		
	90 - 10	99.99924%	99.99921%	99.99866%	99.99783%	99.99711%	99.99571%	99.99254%	99.98314%	99.96331%	99.83391%			
	80 - 20	99.99783%	99.99780%	99.99739%	99.99644%	99.99495%	99.99224%	99.98307%	99.92734%	99.53825%				
	70 - 30	99.99706%	99.99697%	99.99304%	99.99486%	99.99150%	99.98350%	99.94751%	99.54715%					
	60 - 40	99.99586%	99.99583%	99.99554%	99.99097%	99.98378%	99.96048%	99.77990%						
	50 - 50	99.99316%	99.99303%	99.99179%	99.98276%	99.95835%	99.66834%							
	40 - 60	99.98540%	99.98524%	99.98229%	99.92625%	99.56661%								
	30 - 70	99.97075%	99.97001%	99.95906%	99.56379%									
	20 - 80	99.95411%	99.95220%	99.85811%										
	10 - 90	99.91946%	99.68254%											
	5 - 95	0.00000%												
	1 - 99													
			1 - 99	5 - 95	10 - 90	20 - 80	30 - 70	40 - 60	50 - 50	60 - 40	70 - 30	80 - 20	90 - 10	95 - 5

Casework-Type Samples (Body Fluid Mixtures on Varied Substrates) – Eighteen two-component body fluid mixtures were stained onto a variety of substrates including denim, leather, wood, carpet, nylon and wallboard and aged at room temperature for four weeks followed by DNA extraction. No two donor samples used for this study had identical mtDNA haplotypes. This was confirmed by sequencing known reference buccal swabs. Amplification of control region fragments for the mixtures was performed using forensically validated primer sets and conditions. All amplified samples were fractionated by DHPLC into a sequential series of up to twenty fractions. Up to nine fractions for each mixture were sequenced and the resulting electropherograms were analyzed by FLiPARS 2.0 to determine the haplotypes of the individual contributors. The individual contributor haplotypes identified were in full concordance with sequencing results obtained from reference samples for the individual contributors. The FLiPARS 2.0-generated linkage phase determination was characterized by a high degree of base resolution confidence (99.94% to 99.99%) (Table 4).

Casework-Type Samples (Environmental Insult Mixtures) – Seventeen two-component body fluid mixtures were subjected to a variety of environmental insults including gasoline, soil, laundry detergent, used motor oil, sodium hydroxide and acetic acid followed by DNA extraction. No two donor samples used for this study had identical mtDNA haplotypes as previously determined by sequencing reference buccal swabs. Amplification of control region

fragments for the mixtures was performed using forensically validated primer sets and conditions. All amplified samples were then fractionated by DHPLC into a sequential series of up to twenty fractions. Five to nine fractions for each sample were sequenced and the resulting electropherograms were analyzed by FLiPARS 2.0 to determine the haplotypes of each individual contributor. The individual contributor haplotypes identified were in full concordance with sequencing results obtained from individual non-mixed samples. The FLiPARS 2.0-generated linkage phase determination was characterized by a high degree of base resolution confidence (99.80% to 99.99%) (Table 4).

Casework-Type Samples (Hair and Bone Mixtures) – Seven casework-type mixed samples were prepared from hair and bone samples mixed with semen or blood to simulate conditions likely to be encountered by forensic mtDNA practitioners working on challenging samples. All mixtures were amplified and then DHPLC-fractionated as described above. Sequence electropherograms from the resulting DHPLC fractions were then analyzed by FLiPARS 2.0 to determine the haplotypes of each individual contributor. The individual contributor haplotypes identified were in full concordance with sequencing results obtained from individual non-mixed samples. The FLiPARS 2.0 - generated linkage phase determination was characterized by a high degree of base resolution confidence (99.77% to 99.99%) (Table 4).

Table 4 - Linkage phase analysis of casework type samples by FLiPARS 2.0

Tissue Mixture	Substrate/Contaminant	Primer	Mixture Resolved	Confidence Range
Blood/Blood	Denim	B1	Yes	99.98-99.99%
Blood/Blood	Leather	B1	Yes	99.98-99.99%
Blood/Blood	Wood	B1	Yes	99.98-99.99%
Semen/Semen	Carpet	B1	Yes	99.98-99.99%
Semen/Semen	Nylon	B1	Yes	99.98-99.99%
Saliva/Saliva	Carpet	A2	Yes	99.95-99.98%
Saliva/Saliva	Leather	A2	Yes	99.94-99.99%
Saliva/Saliva	Wallboard	A2	Yes	99.96-99.97%
Saliva/Saliva	Wood	A2	Yes	99.94-99.97%
Blood/Blood	Gasoline	B1	Yes	99.98-99.99%
Blood/Blood	Motor Oil	B1	Yes	99.99%
Blood/Blood	Detergent	B1	Yes	99.98-99.99%
Blood/Blood	Acetic Acid	B1	Yes	99.97-99.99%
Semen/Semen	Sodium Hydroxide	B1	Yes	99.98-99.99%
Saliva/Saliva	Gasoline	A2	Yes	99.95-99.99%
Saliva/Saliva	Motor Oil	A2	Yes	99.97-99.99%
Saliva/Saliva	Acetic Acid	A2	Yes	99.95-99.98%
Blood/Semen	Denim	C1	Yes	99.89-99.95%
Blood/Semen	Leather	C1	Yes	99.89-99.97%
Blood/Semen	Wood	C1	Yes	99.95-99.99%
Saliva/Semen	Carpet	C1	Yes	99.95-99.99%
Saliva/Semen	Nylon	C1	Yes	99.95-99.99%
Blood/Saliva	Carpet	A2	Yes	99.99%
Blood/Saliva	Leather	A2	Yes	99.99%
Blood/Saliva	Wallboard	A2	Yes	99.99%
Blood/Saliva	Wood	A2	Yes	99.99%
Blood/Semen	Gasoline	C1	Yes	99.95-99.97%
Blood/Semen	Soil	C1	Yes	99.87-99.97%
Blood/Semen	Detergent	C1	Yes	99.94-99.97%
Blood/Semen	Sodium Hydroxide	C1	Yes	99.94-99.97%
Saliva/Semen	Gasoline	C1	Yes	99.80-99.96%
Saliva/Semen	Sodium Hydroxide	C1	Yes	99.97-99.98%
Blood/Saliva	Motor Oil	A2	Yes	99.99%
Blood/Saliva	Detergent	A2	Yes	99.99%
Blood/Saliva	Acetic Acid	A2	Yes	99.99%
Bone/Bone	N/A	C1	Yes	99.99%
Bone/Bone	N/A	C1	Yes	99.77-99.99%
Dyed Hair/Blood	N/A	A1	Yes	99.88-99.96%
Head Hair/Blood	N/A	A1	Yes	99.99%
Axial Hair/Blood	N/A	A1	Yes	99.87-99.97%
Pubic Hair/Semen	N/A	A1	Yes	99.98-99.98%
Permed Hair/Blood	N/A	A1	Yes	99.78-99.89%

Alignment Ambiguities: It was discovered during the validation phase that certain mutation patterns may cause alignment ambiguities. According to published FBI guidelines^[35], the fewest number of changes to a sequence to produce the reference should be used as the first criteria for

alignment. Additional criteria are also defined in order to break potential “ties” in the number of edits required, including preferring transitions over transversions and substitutions over indels. A strict implementation of these rules, however, does not always produce an alignment that is biologically reasonable. An example of a potentially ambiguous alignment that was observed follows:

```
rCRS>                292 - TTTCCACCAAACCCCCCTCCCCCGCTTCTGG - 323
Query Sequence>      292 - TTTCCACCAAACCCCCCTCCCCCGCTTCTGG - 323
```

Aligning these two sequences together can produce two different alignment results.

Alignment 1: 309D, 315.1C

Alignment 2: 309T, 310C

Two edits are required to convert the query sequence into the reference in both cases, so further criteria from the alignment definitions are required to determine which is correct. Namely, the preference of substitutions over indels “technically” breaks the tie here in favor of the second alignment even though the resulting alignment is not biologically favored. After consulting with expert forensic practitioners (Terry Melton of Mitotyping Technologies, Suni Edson of AFDIL, and Walther Parson of Institute of Legal Medicine, Innsbruck Medical University), we were able to confirm that the first alignment is actually the correct alignment. The second alignment is an inaccurate artifact created as a consequence of alignment definitions that do not always reflect biological patterns of mutation.

This presented a challenge for the implementation of FLiPARS 2.0, as either alignment could be produced depending on the scoring scheme used. The forensic version of Sequencher[®] 4.10.1 by Gene Codes, true to the published alignment guidelines, mistakenly yields the second alignment in the example above. In order to avoid this error, the scoring scheme for the alignment algorithm of FLiPARS 2.0 was modified to better reflect evolutionary processes and to thereby yield the first alignment. It is important to emphasize that the opposite case was also observed - *i.e.*, a case in which the modified FLiPARS 2.0 scoring scheme was found to yield an incorrect alignment, while the guidelines employed by Sequencher[®] produced the correct alignment. Fortunately, FLiPARS was designed to employ a modular plug-in system. In this case, the plug-in system was extended to the alignment algorithms as well as the file parsing system. Thus, two alignment scoring schemes (“Default” and “Forensic D”) are now included with FLiPARS 2.0. This provides the practitioner with the option of double-checking the default alignments (*i.e.*, strict FBI guidelines/Sequencher[®]-like alignments) with a second scoring scheme that may yield a more accurate alignment. The “Forensic” alignment algorithm of Sequencher favors substitutions over indels, while the optional “Forensic D” alignment algorithm of FLiPARS 2.0 favors indels over substitutions. It should be noted that Budowle, *et al.*, have recently advocated for the anchoring of the T at position 310 when aligning this region^[36]. This would force insertions/deletions to be placed at position 309 resulting in the alignment favored by the aforementioned practitioners. While such an anchoring approach does serve to correctly resolve alignment ambiguities at this specific location within HV2, this solution does not reflect evolutionary processes thus necessitating that additional “patches” are applied wherever comparable alignment ambiguities arise outside of this region.

It should be noted that in the vast majority of cases, both aligners were found to produce identical alignments. It is an unfortunate limitation of our current alignment guidelines that such ambiguities may arise, but FLiPARS does not make the “one size fits all” assumption with respect to sequence alignments. This is also true of the base-calling system, which provides the practitioner with the tools (and thus the responsibility) to ensure that the biologically and forensically correct results are being obtained from the data. This is in contrast to other commercially available sequence analysis suites (*e.g.*, basecalling algorithms that employ Phred which endeavors to fit real-world electrophoretic data to an artificial sine wave) which may mask such potential difficulties and thus give users a false sense of the accuracy in the results being displayed.

Noisy Sequencing Reactions – The simulated casework samples used in the current study included several examples of ‘noisy’ (*i.e.*, high baseline) sequences. These are expected with DNA samples that are degraded or which have been subjected to a variety of environmental insults. Such datasets reflect conditions indicative of what many forensic practitioners encounter when processing casework samples. For some of these samples, base-calling and sequence alignment were clearly not as efficient or as accurate as for less challenging/(pristine) samples.

To improve the analysis of such noisy sequences, it was necessary to modify FLiPARS 2.0 to allow for greater analyst flexibility in editing problematic sequences. One of the most important functionalities added was the ability to merge peaks that constituted mixed bases that might be missed by FLiPARS 2.0 due to slight but reproducible irregularities in base mobility (*i.e.*, bases in a mixture which should have produced a clear mixed-base position were instead not identified as overlapping because they were spaced too far apart. While the critique could be made that such manual modifications might constitute massaging of the sequence data, the reality is that current practices rely solely on the automated base-calling by programs such as ABI Base-caller. In fact, numerous instances were observed where ABI’s base-caller missed base-calls because it tries to fit the chromatographic data to a completely artificial sine wave. Without FLiPARS 2.0 or the known sequences of the contributors, it would have been nearly impossible without manually calling each peak for an analyst to identify such a base-calling mistake.

In short, computers and software-associated algorithms cannot substitute for human analysis. They remain tools to speed and improve the accuracy of sequence analysis, not to automate it completely away from human input. FLiPARS 2.0 has been engineered to embrace this view. Accordingly, it provides the practitioner with the tools to define what is and isn’t biologically and forensically real and/or relevant.

Implications for Policy and Practice

A mixture of different mtDNA molecules in a single sample presents an often insurmountable challenge to successful sequence analysis. In a forensic context, mtDNA mixtures are most typically “situational”, in which mtDNA from separate humans is found in association with a single evidentiary sample. Additionally, individual humans can possess more than one mtDNA haplotype, *i.e.*, “heteroplasmy”. Interindividual differences in mtDNA may involve base substitutions or small insertions/deletions. These result in ambiguous base calls.

Conventional methods of mixture deconvolution (*e.g.*, denaturing gradient gel electrophoresis, single-strand conformational polymorphism analysis and subcloning into bacterial vectors)

involve tedious and time consuming processes which are often difficult to automate. This has effectively deterred forensic laboratories from implementing these technologies.

The developmental validation of linkage phase analysis as a powerful and extremely accurate means of resolving mtDNA mixtures offers practitioners the opportunity to obtain potentially useful information from what might otherwise be uninterpretable results. Coupled with publication of findings in peer-reviewed journals linkage phase-based mixture deconvolution will be on sound legal footing with regard to the Frye and Daubert standards. The approach has been tested; the underlying reasoning is scientifically valid and the potential error rates and limitations have been evaluated. The FLiPARS 2.0 software application developed and tested under the current DNA Research and Development Award (2009-DN-BX-K047) has automated an extremely laborious and computationally-intensive process. The availability this software application serves to remove a significant and immediate obstacle to the use of this technology by forensic practitioners. As a result, forensic practitioners will be able to readily ascribe specific haplotypes to the individual contributors to a mixture and to do so with a reportable degree of statistical confidence. In turn, investigators may be able to obtain potentially probative genetic information from samples that have historically not been amenable to analysis by direct sequencing. This has the potential to expand the potential applicability of mtDNA testing to a broader range of criminal investigations.

Implications for Further Research

The central objective of this research program was to develop and validate a software application to automate the computationally intensive analysis of electrophoretic data that is necessary to determine the linkage phase (*i.e.*, haplotypes) of individual contributors to an mtDNA mixture. This software leverages previously validated DHPLC technology for the differential fractionation of mixed (two component) DNA samples. Through a comparative analysis of the sequence electropherograms representing DNA from two or more chromatographic fractions it has been shown that the linkage phase (and thereby the specific haplotype) of the individual components of a mixture can be determined. If adopted by forensic practitioners, this software application has the potential to increase the number of forensic samples for which definitive sequence analyses can be conducted. The current research program has been completed in accordance with DNA Advisory Board developmental validation standards for sensitivity, reproducibility and accuracy especially with casework-type samples^[37]. The future of this research, therefore, will depend on achieving three critical objectives. These are:

- 1) **Rigorous Interlaboratory Validation Studies by Practitioners** of the FLiPARS 2.0 software application for resolving mtDNA mixtures. This should involve participation by established practitioners with extensive experience in forensic mtDNA analysis. The NIJ could play an important role in facilitating these studies by providing resources needed to enable interested practitioner laboratories to participate. This might include funding to help with the lease of DHPLC instrumentation and the hiring of additional personnel that would need to be dedicated to evaluating this technology and the associated software tools.

- 2) **Further Refinement the Software Application** to meet the needs of forensic practitioners is essential. Under the current DNA Research and Development award, a robust and reliable software application (FLiPARS 2.0) has been developed which automates the process of linkage phase analysis. While FLiPARS 2.0 is a fully-functional software application, it will certainly be improved upon through feedback from practitioners. Among the most important features that could be improved are the tools that enable practitioners to use their own expertise in evaluating the validity of peak calls and alignments made by the software. Developing an optimal interface between the software and its users will necessitate a continuing conversation between the developer and forensic practitioner community.
- 3) **Extend Validation to Additional Amplicons** in recognition of the fact that forensic practitioners often employ a variety of primer pairs beyond those with which DHPLC was validated. While the standard HV1A through HV2B primer sets yield amplification products that are approximately 270bp in length, highly degraded samples frequently contain DNA molecules that are severely restricted in size (*e.g.*, <150bp). To facilitate the analysis of such highly degraded material, a variety “mini-primer” pairs have been developed that span each of the HV regions with an average amplicon size of 140 bp. Although these were not evaluated as part of the current research, it is anticipated that FLiPARS 2.0 will perform as well with these primer pairs as it did in the current study.

Cited References

1. Budowle, B., *et al.*, *DNA Typing Protocols: Molecular Biology and Forensic Analysis*. Forensic Science Series. 2000: Biotechniques Books. 36.
2. Holland, M.M. and T.J. Parsons, *Mitochondrial DNA Sequence Analysis-Validation and use for forensic casework*. Forensic Science Review, 1999. **11**(1): p. 22-49.
3. Holland, M.M., *et al.*, *Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War*. Journal of Forensic Sciences, 1993. **38**(3): p. 542-53.
4. Ivanov, P.L., *et al.*, *Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II*. Nat Genet, 1996. **12**(4): p. 417-20.
5. Hagelberg, E., I.C. Gray, and A.J. Jeffreys, *Identification of the skeletal remains of a murder victim by DNA analysis. [see comments.]* Nature, 1991. **352**(6334): p. 427-9.
6. Allen, M., *et al.*, *Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities*. J Forensic Sci, 1998. **43**(3): p. 453-64.
7. Budowle, B., *et al.*, *Mitochondrial DNA regions HVI and HVII population data*. Forensic Science International, 1999. **103**(1): p. 23-35.
8. Parsons, T.J., *et al.*, *A high observed substitution rate in the human mitochondrial DNA control region*. Nat Genet, 1997. **15**(4): p. 363-8.
9. Fournay, R.M., *Mitochondrial DNA and Forensic Analysis: A Primer for Law Enforcement*. Can. Soc. Forens. Sci. J, 1998. **31**(1): p. 45-53.
10. Wilson, M.R., *et al.*, *Validation of mitochondrial DNA sequencing for forensic casework analysis*. Int J Legal Med, 1995. **108**(2): p. 68-74.
11. Melton, T., *et al.*, *Forensic mitochondrial DNA analysis of 691 casework hairs*. J Forensic Sci, 2005. **50**(1): p. 73-80.
12. Hanekamp, J.S., W.G. Thilly, and M.A. Chaudhry, *Screening for human mitochondrial DNA polymorphisms with denaturing gradient gel electrophoresis*. Hum Genet, 1996. **98**(2): p. 243-5.
13. Barros, F., *et al.*, *Rapid and enhanced detection of mitochondrial DNA variation using single-strand conformation analysis of superposed restriction enzyme fragments from polymerase chain reaction-amplified products*. Electrophoresis, 1997. **18**(1): p. 52-4.

14. Alonso, A., *et al.*, *Rapid detection of sequence polymorphisms in the human mitochondrial DNA control region by polymerase chain reaction and single-strand conformation analysis in mutation detection enhancement gels*. Electrophoresis, 1996. **17**(8): p. 1299-301.
15. Steighner, R.J., *et al.*, *Comparative identity and homogeneity testing of the mtDNA HV1 region using denaturing gradient gel electrophoresis*. Journal of Forensic Sciences, 1999. **44**(6): p. 1186-98.
16. Andreasson, H., *et al.*, *Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology*. Forensic Sci Int Genet, 2007. **1**(1): p. 35-43.
17. Andreasson, H., *et al.*, *Mitochondrial sequence analysis for forensic identification using pyrosequencing technology*. Biotechniques, 2002. **32**(1): p. 124-6, 128, 130-3.
18. Hall, T.A., *et al.*, *Base composition analysis of human mitochondrial DNA using electrospray ionization mass spectrometry: a novel tool for the identification and differentiation of humans*. Anal Biochem, 2005. **344**(1): p. 53-69.
19. Steven A. Hofstadler, P.D., *Analysis of DNA Forensic Markers Using High Throughput Mass Spectrometry*, in *Final report: NIJ Award #2006-DN-BX-K011*. 2008. p. 117.
20. Andreasson, H., *et al.*, *Quantification of mtDNA mixtures in forensic evidence material using pyrosequencing*. Int J Legal Med, 2006. **120**(6): p. 383-90.
21. Huber, C.G., P.J. Oefner, and G.K. Bonn, *High-resolution liquid chromatography of oligonucleotides on nonporous alkylated styrene-divinylbenzene copolymers*. Anal Biochem, 1993. **212**(2): p. 351-8.
22. Huber, C.G., P.J. Oefner, and G.K. Bonn, *Rapid and accurate sizing of DNA fragments by ion-pair chromatography on alkylated nonporous poly(styrene-divinylbenzene)*. Analytical Chemistry, 1995. **67**: p. 578-585.
23. Emmerson, P., *et al.*, *Characterizing mutations in samples with low-level mosaicism by collection and analysis of DHPLC fractionated heteroduplexes*. Hum Mutat, 2003. **21**(2): p. 112-5.
24. Etokebe, G.E., *et al.*, *Physical separation of HLA-A alleles by denaturing high-performance liquid chromatography*. Tissue Antigens, 2003. **61**(6): p. 443-50.
25. Steighner, R.J., *et al.*, *Comparative identity and homogeneity testing of the mtDNA HV1 region using denaturing gradient gel electrophoresis*. J Forensic Sci, 1999. **44**(6): p. 1186-98.
26. Budowle, B., *et al.*, *Forensics and mitochondrial DNA: applications, debates, and foundations*. Annu Rev Genomics Hum Genet, 2003. **4**: p. 119-41.
27. Danielson, P.B., *et al.*, *Separating human DNA mixtures using denaturing high-performance liquid chromatography*. Expert Rev Mol Diagn, 2005. **5**(1): p. 53-63.
28. Danielson, P.B., *et al.*, *Resolving mtDNA mixtures by denaturing high-performance liquid chromatography and linkage phase determination*. Forensic Sci Int Gen, 2007. **1**(2): p. 148-53.
29. Danielson, P.B., *Mitochondrial DNA Analysis by Denaturing Liquid Chromatography for the Separation of Mixtures in Forensic Samples*. Final Technical Report for 2003-IJCX-K104 submitted to the National Institute of Justice, Forensic DNA Research and Development Program, 2008: p. 1-105.
30. Troelsen, A., *Pro C# 2008 and the .NET 3.5 Platform (Windows.Net)*. Fourth Edition ed. 2007 Berkely, CA Apress.
31. www.microsoft.com/netframework [cited].
32. Vial, P., *et al.*, *Software tool for portal dosimetry research*. Australas Phys Eng Sci Med, 2008. **31**(3): p. 216-22.
33. Kim, K.W., *et al.*, *PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets*. Bioinformatics, 2007. **23**(16): p. 2155-62.
34. Danielson, P.B., *Mitochondrial DNA Analysis by Denaturing Liquid Chromatography for the Separation of Mixtures in Forensic Samples*, Final Technical Report 2009 [Online] <http://www.ncjrs.gov/pdffiles1/nij/grants/226638.pdf>
35. Federal Bureau of Investigation "Further Discussion of the Consistent Treatment of Length Variants in the Human Mitochondrial DNA Control Region" *Forensic Science Communications* October 2002. 4 (4).
36. Budowle, *et al.*, *Automated alignment and nomenclature for consistent treatment of polymorphisms in the human mitochondrial DNA control region* *J. Forensic Sci*, 2010. **55**, 1190-1195.
37. Federal Bureau of Investigation "Quality assurance standards for forensic DNA testing laboratories", *Forensic Science Communications* July 2000. **2** (3).

Dissemination of Research Findings

A total of two semiannual progress reports on this research program have been provided to the National Institute of Justice. Research findings were also disseminated through invited research seminars listed below. With the completion of the core research and development objectives under award 2009-DN-BX-K047, a manuscript will be prepared for submission to the Journal of Forensic Science or equivalent publications.

Invited Research Talks and Poster Presentations – Although the bulk of the software development was only completed in July of 2010, we been fortunate to have already had two opportunities to give invited talks to introduce the forensic community to the potential utility of FLiPARS 2.0. These were:

- June 2010 “Resolving mtDNA Mixtures by Denaturing HPLC and Linkage Phase Analysis”, The National Institute of Justice Conference 2010 - Forensic DNA: Tools, Technology, and Policy. Washington, D.C.
- Sept. 2010 “De-Convoluting mtDNA Mixtures Using HPLC” Virginia Institute of Forensic Science, Second Annual Current and Future Advances in Human Identification Conference, Hampton, VA

Application to Casework – In 2010, the FLiPARS software application was employed for the linkage phase analysis of a mixed mtDNA sample as part of a sexual assault / homicide investigation in the Netherlands. The mixture was accurately deconvolved and the approach was ruled admissible by the second section of the High Court of the Netherlands.

Software Download and User Guide – The central deliverable of this project is a bioinformatic application entitled FLiPARS 2.0 (Fractionated Linkage Phase Analysis Resource System). This software automates that computationally intensive process of quantitatively comparing and then confidently deconvolving mixed sequence electropherograms from direct sequencing of mtDNA mixtures. A website has been established (www.flipars.com) which is registered in the research director’s name. The website is meant to promote the technology and software, to serve as a resource for users of the program and to distribute the program to laboratories across the country and globe. The software application can be downloaded from this website along with a User’s Guide (Appendix I) under the “documentation” tab.

The FLiPARS 2.0-generated linkage phase determination was characterized by a high degree of base resolution confidence FLiPARS is licensed under the Apache 2.0 License, providing free access to the source code for virtually any purpose. Developers can modify and distribute the software as they see fit, so long as the appropriate licensing matters are followed. Modifications or additions to FLiPARS may be published under any license, again assuming that the original portions of the software retain and follow the Apache 2.0 Licensing guidelines.

Appendix I

User's Manual FLiPARS 2.0

FLiPARS 2.0 Manual

Jesse Goeglein
jesse@flipars.com

July 29, 2010

Contents

1	Introduction	6
2	Linkage Phase Analysis Background	7
3	Basic Workflow	9
3.1	Input Fractionated Files	9
3.2	Parse the Files	9
3.3	Inspect the Parsing	10
3.4	Make Necessary Edits	10
3.5	Run Linkage Phase Analysis	10
4	Main Window	12
4.1	Project Options	12
4.1.1	New Project	12
4.1.2	Open Project	12
4.1.3	Save Project	12
4.1.4	FLiPARS Settings	13
4.1.5	Help	13
4.2	Input File List	13
4.2.1	Edit Primers	13
4.2.2	Re-order Files	13
4.2.3	Choose Reference	13
4.2.4	Choose Aligner	13
4.3	Parse Files	14
4.4	Visualize	14
4.5	Run Linkage Phase Analysis	14
4.6	Sequence Display Box	14
4.7	Re-examine	15
4.8	Resolved Sequences	15
4.8.1	Linkage Phase Resolution View	15
5	Chromatogram Display	16
5.1	Edit Sequence Controls	16
5.1.1	Trimming Left and Right	17
5.1.2	Keep Tall	17
5.2	Zooming and Navigation Controls	17
5.2.1	Zooming	17
5.2.2	Go To Selected	18
5.2.3	Go To Alignment	18
5.3	Display Options Controls	18
5.3.1	Display Called Peaks	18
5.3.2	Display Uncalled Peaks	18
6	Pop-up Sequence Editor	19
6.1	Validating or Invalidating Peaks	19
6.1.1	Selecting Overlapping Peaks	19
6.1.2	Creating a Mixed Base Position	20
6.2	Examining Signal Data	20
6.3	Consider Mixed	21

7	Linkage Phase Analysis Window	22
7.1	First and Second Fractions	22
7.1.1	Drop-down File Lists	22
7.1.2	Select In Alignment	22
7.2	Fraction Comparisons	23
7.3	Resolved Sequences	23
8	Settings	24
8.1	Available Parsers	24
8.1.1	Add Plugin Form	24
8.2	Peak Stringency	25
8.3	Available Primers	25
8.4	Available Reference Sequences	25
8.4.1	Add Reference Sequence Dialog	25
8.5	Available Aligners	25
8.5.1	Default 'Forensic' Aligner	25
8.5.2	Alternate 'ForensicD' Aligner	26
8.5.3	Add Aligner Form	26
9	Common Tasks	27
9.1	Add New Reference Sequences	27
9.2	Add New Primers	27
9.3	Change Existing Primers	27
9.4	Edit Sequences	27
9.5	Adding New File Input Plugins	28
9.6	Align Sequences	28
10	For Developers	29
10.1	File Input Plugins	29
10.2	Sequence Alignment Plugins	29
11	Acknowledgements	31

List of Figures

1	Main Window	12
2	Project Options Bar	12
3	Parsed Main Window	14
4	Linkage Phase Resolution View	15
5	Linkage Phase Resolution View - Off	16
6	Chromatogram Display	16
7	Chromatogram Trimming	17
8	Pop-up Sequence Editor	19
9	Pop-up Sequence Editor	20
10	Linkage Phase Analysis Window	22
11	Linkage Phase Analysis w/ Graphs	23
12	Edit Settings Window	24

License For Use and Redistribution

Copyright 2010 Jesse Goeglein

Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

1 Introduction

Welcome to FLiPARS - the Fractionated Linkage Phase Analysis and Resource System! Picking up any new piece of software will always incur some sort of learning curve, and this instruction manual is meant to be a user's guide to effectively coming to grips with what FLiPARS can, and maybe more importantly, cannot do. There is also information for interested developers at the end of this guide with references to full programmatic documentation online. So what exactly does FLiPARS do and what new features does it bring to the sequence assessment landscape?

In short, this software package was designed from the ground up to support forensic technicians in particular and the scientific community in general to resolve mixtures of DNA, without any *a priori* knowledge of the contributors, as seamlessly as possible. FLiPARS uses a unique base calling algorithm to ensure that all peaks are available to you for editing the appropriate and experimentally verified bases into (or out of!) your sequences. It also inverts the traditional process of modifying those sequences: Rather than editing the ATGC text of your sequences directly, you simply modify what the system sees as legitimate base calls from the chromatograms. Making changes directly in the chromatogram forces the user to analyze the data itself to see if their proposed changes can be backed up by the data that they are looking at - or if what they are wanting to see isn't really there. Alignments of the newly modified sequences against a chosen reference sequence are done on the fly so that you can immediately see the implications of those modifications. Chromatogram editing functions and linkage phase analysis are segregated into separate windows so that the user doesn't become overwhelmed with all of the information being generated by the software. We hope you'll find using FLiPARS to be an useful and enjoyable experience. Feedback is always welcome, whether it's good, bad, or suggestions for additional capabilities that you wish the system possessed - please send those comments to jesse@flipars.com . Again, welcome to FLiPARS!

2 Linkage Phase Analysis Background

The quantity and quality of DNA are critical factors in forensic investigations. When the use of short tandem repeat (STR) nuclear loci[1] fails, however, mitochondrial DNA (mtDNA) often offers investigators the only remaining opportunity to obtain potentially probative genetic information[2]. Mitochondrial DNA analysis has often been used for the analysis of especially challenging samples. For example, it has been used to identify severely weathered/degraded remains from the Vietnam War[3], Czar Nicholas II[4] and murder victims[5] as well as to provide useful forensic information on shed head hairs and saliva from robbery caps[6]. It is also frequently used in DVI cases such as plane crashes where remains may be exposed to conditions that compromise DNA quality. Its exonuclease-resistant nature and the presence of up to several thousand copies of mtDNA per cell facilitates the analyses of degraded and/or low-copy number material[1, 7]. Additionally, the uniparental inheritance of mtDNA, allows reference material to be obtained from maternal relatives[8].

Analysis of mtDNA is currently accomplished almost exclusively by sequencing the DNA from hypervariable regions 1 and 2 (HV1/HV2) of the control region. This approach has been rigorously validated and has withstood several court challenges (see www.denverda.org for specific case law examples). Although the protocol for sequencing mtDNA is well established, the presence of a mixture of different mtDNA molecules in a single sample can present a significant obstacle to successful mtDNA analyses by standard methods. In fact, a situational mixture from two or more individuals in a single sample, and even naturally occurring heteroplasmy, typically precludes successful mtDNA analysis[2, 9, 10]. This roadblock occurs because sequencing a mixture of two or more DNA amplicons yields electropherograms characterized by overlapping peaks at sites where the amplicons differ in sequence. Because peak height is sequence context dependent, it cannot be used by itself to determine the absolute or even relative quantities of DNA from the individual contributors to the mixture. This can impede the forensic use of mtDNA.

Underscoring the fact that this is not a minor problem is the observation from extensive casework records that a significant proportion of evidentiary hairs examined were heteroplasmic (11.4%) or displayed a mixed profile (8.7%). Moreover, the occurrence of mixed mtDNA profiles appears to increase with the age of a sample and is usually not ameliorated even following extensive validated cleaning methods[11]. This likely represents only the tip of the iceberg since samples which are suspected to yield mixtures are often not even submitted for analysis. A reliable means of resolving the individual sequences within a mixture could greatly aid investigators by increasing the range of casework samples suitable for mtDNA testing.

There are a number of established molecular strategies that could be employed to separate DNA mixtures into their individual components. These include separation by denaturing gradient gel electrophoresis (DGGE)[12] or single-strand conformational polymorphism (SSCP) analysis[13, 14, 15] and subcloning into bacterial vectors. These approaches are generally time consuming, necessitate multiple handling steps, require laborious product purification and are not readily adaptable to automation. These factors have all been obstacles to the implementation of these technologies by forensic laboratories. Both DGGE and SSCP require manual recovery of fractionated DNA from polyacrylamide gels and a second round of PCR amplification to generate enough template for DNA sequencing. Subcloning is an even more time and labor-intensive approach. It would require forensic scientists to screen and sequence DNA from multiple transformed bacterial colonies to ensure that observed sequence differences reflect genuine differences in the starting template rather than artificial variants introduced as a result of nucleotide misincorporation during PCR. More recently, both pyrosequencing[16, 17] and electrospray ionization-mass spectrometry[18] have attracted significant interest as a means of resolving mtDNA mixtures. Both approaches have the significant advantage of being able to more precisely quantitate heteroplasmic and other mixed samples than earlier nucleic acid analysis methods. In the case of mass-spectrometry, it is possible to take advantage of the fact that the exact mass of each deoxyribonucleotide is a precisely known value. Accordingly, it becomes a relatively straight forward matter to generate a tightly constrained list of possible deoxyribonucleotide combinations that can account for the total mass of a given PCR amplified fragment. Therein, however, is rooted the most significant limitation of a mass-spectrometry based approach. While it is possible to predict the base composition of an assayed

fragment, it is not possible to know with absolute certainty the precise base sequence of a fragment. Given the frequency with which new mtDNA variants are reported, it is not inconceivable that fragments having identical base compositions can have different sequences. Furthermore, mass spectrometry approaches based on post amplification restriction digests necessitate additional sample handling while those based on tiled sets of PCR primers[19] require multiple PCR reactions that in some cases may require the consumption of more of a precious DNA extract than is available. Pyrosequencing, on the other hand does allow an analyst to determine the actual base sequence of an amplified fragment and it can provide quantitative information on the composition of a mixture[20]. This approach, however, is limited by the fact that read lengths are limited to approximately 50 nucleotides and the interpretation of mixtures is complex - typically necessitating manual interpretation to ensure accuracy. While both approaches are certainly promising, dideoxy sequencing remains the method of choice for mtDNA analysis in most forensic laboratories that process such samples.

Denaturing High-Performance Liquid Chromatography (DHPLC) [21, 22] is a chromatographic means of fractionating natural (heteroplasmic) or situational (multi-contributor) DNA mixtures prior to sequencing[23, 24]. In contrast to the alternative approaches that have been proposed for the separation of mtDNA mixtures[12, 13, 19, 20, 25, 26], DHPLC does not require secondary amplification or excessive sample manipulation to resolve a mixture of different mtDNA haplotypes. Furthermore it is completely consistent with established forensic SOPs for standard fluorescence-based sequencing. Under NIJ DNA Research and Development Award 2003-IJCX-K104 we have rigorously validated this approach to resolving mtDNA mixtures [27, 28, 29] using both reference and casework type samples. Specifically, we have demonstrated:

- (1) the sensitivity of DHPLC to detect and fractionate mixtures involving all classes of polymorphisms;
- (2) the reproducibility and statistical correlation between peak height and DNA quantity ratios in sequencing electropherograms;
- (3) the reliability of determining individual haplotypes by linkage phase analysis of sequence data from DHPLC fractionated samples.

Finally, we have developed standard operating procedures and statistically-grounded interpretation guidelines for the use of DHPLC to resolve mtDNA mixtures.

3 Basic Workflow

Resolving or determining the two distinct DNA sequences that compose a mixture using FLiPARS will generally proceed through a systematic process. This section outlines that process and the steps in detail. The following sections assume that the fractionation and sequencing reactions have already been carried out using the appropriate SOPs.

3.1 Input Fractionated Files

The first step is to tell FLiPARS where to find the sequencing data via the 'Add Files' button in the main window. A familiar dialog will appear that will enable you to browse your computer system to find the appropriate files. You may select multiple files to add at a single pass. When you press 'Open' from that dialog, FLiPARS will attempt to add the selected files to the current project by checking for an installed plugin that knows how to parse those files (based on the file extension). By default, FLiPARS knows how to parse .SCF, .AB1 and certain tab-delimited .TXT files. However, if the provided files are not formatted correctly, no error will be shown here.

Once the files are added, you may reorder the files for display purposes as you wish using the up and down arrows. You can also select which reference sequence the input sequences were generated against (FLiPARS defaults to the revised CRS). You can also (optionally) tell the system about the primers that were used to generate each sequence. FLiPARS uses this primer information to determine whether or not to look at the reverse complement of each sequence when attempting to align them against a reference sequence. FLiPARS will also reverse the raw data points when viewing the sequences in the chromatogram editor so that the aligned sequence is viewable as the system is looking at it relative to the selected reference sequence. 'Forward' primers are expected to be sequencing in the same direction as the reference sequence you have chosen. Reverse primers, then, are expected to be sequencing in the opposite direction (on the opposite strand to the reference) and thus will need to be reverse complemented in order to find their position within the reference. If no primer is set, the sequence is assumed to be forward with respect to the reference.

It should also be noted that if the primers in use in your laboratory or the reference sequence that you are sequencing against are different from those provided by default in FLiPARS, you can add your own to the system: See Section 8 on page 24 for more information on how to do so.

3.2 Parse the Files

The next step that FLiPARS will always take is to parse the files. This will occur when you press the 'Parse Files' button in the main window, or if this step isn't taken manually, they will be parsed when pressing the 'Visualize' or 'Run Linkage Phase Analysis' buttons. Newly added files will also only be parsed when one of these actions takes place.

While the files are being parsed, the progress bar in the lower left hand corner of the main window will be updated and brief messages regarding what the system is working on will be displayed below that progress bar. Once the parsing is completed, the main window will be updated with what FLiPARS automatically made for base calls in the provided sequence data files as well as displaying the portion of the selected reference sequence in which the sequences were found. Provided sequences will be labeled with the appropriate file name and primer information and will have alternating black/blue backgrounds. The reference sequence will always be displayed above those sequences with a red background, and it also will be labeled appropriately. Labels are interspersed throughout the sequence so that you can see exactly what sequence you're looking at without having to figure out which file is where (5th from the top, for example).

3.3 Inspect the Parsing

The end user will almost certainly need to inspect the automated base calls. Mixed positions are marked as N and are highlighted in orange, making them easy to spot. Mis-matched bases against the reference sequence are highlighted in gray, and these may often indicate a missed mixed base call by the system due to a weak signal from the smaller contributor at that position. Clicking on any letter in the sequence will cause that base-call to be highlighted and its aligned position within the reference sequence will be shown. If it is a mixed base, the two signals composing that mixture will be shown as well. Incorrect base-calls or noisy sequences can dramatically affect the alignment of each sequence, thus affecting which mixed-base positions are compared. For mixed bases to be compared during linkage phase analysis, they must be aligned to the same position in the reference sequence. Visually, the orange N's should create a column in the main window when the sequences are parsed and called correctly. In addition, if some sequences contain a mixture at one position while others do not due to weak secondary mixture signals, then it may be advantageous (if the data permit it) to go in and tell the system that there is in fact a mixed base at that position.

NOTE: FLiPARS uses a novel algorithm for calling bases in order to avoid missing any mixed base positions. Current base-calling algorithms assume (rightly) that you probably DON'T want mixes, and as such, would require the system to go back into the raw data to re-examine any potential mixed base positions that were missed. FLiPARS also provides ALL potential base-calls to the end-user to select from as being 'real' or 'noise'.

3.4 Make Necessary Edits

If anomalies are identified in the parsed sequences, there are two ways to examine the raw data and make further changes. You may either select a base by clicking on it and then clicking on the 'Visualize' button, or you can click a second time inside the box surrounding the selected base. This is NOT a double click, so please give the system a moment to process your selection before clicking a second time within the chosen box. The view will automatically be centered over your selected base call in the chromatogram display. From here, you may open up another, smaller editor that allows you to add or remove base-calls. Alternatively, large-scale modifications may be made from the chromatogram display, such as trimming 'noisy' ends from the sequences (see section 5 on 16 for full details).

NOTE: Modifications in FLiPARS are not done by editing textual 'ATGC' information. Rather, modifications are made by going through the chromatogram information and either verifying that a signal peak is either real or noise. Alignment definitions exist so that the end-user doesn't have to think about them, and because computers are good at such calculations. The system will re-calculate the necessary information on the fly as you make changes and will automatically ensure that the alignment information is corrected. Furthermore, inserting textual 'ATGC' information may or may not be supported by the underlying signal data. Forcing the user to actually look at the data while making the changes helps to encourage good QA/QC practices, and it also helps to reduce human errors in finding exactly where in the sequence the modifications belong.

3.5 Run Linkage Phase Analysis

Once the sequences are manually examined and modified, pressing the 'Run Linkage Phase Analysis' button will bring up the linkage phase information window. All comparisons among the various fractions will have been made and the two fractions producing the largest average shift amongst the most number of mixed base positions will be automatically selected and displayed for you. Drop-boxes containing each input sequence are available as well if you would like to manually verify the accuracy of the system. The indices (from 1 to the number of sequences that you provided) of the selected fractions as well as the fractions determined to give the largest average shift over the most number of mixed base positions are displayed at the top. Clicking on the mixed base positions for each respective fraction will display a zoomed-in graph of the mixed base

signal. The two distinct resolved sequences are shown at the bottom. More information on the linkage phase analysis window can be found in Section 7 on page 22

4 Main Window

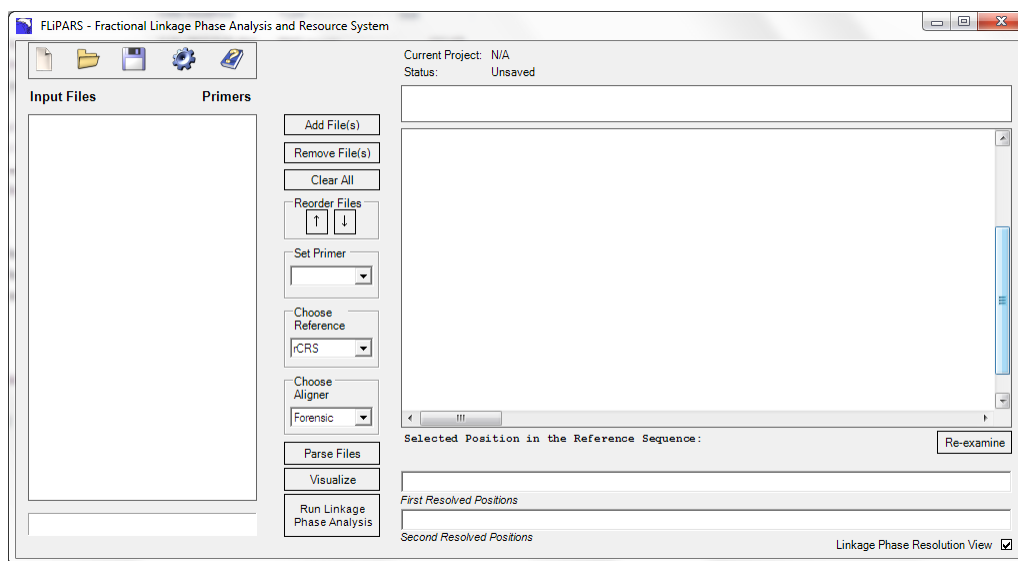


Figure 1: The main window as it appears when FLiPARS is first launched

The main window serves as the starting point for sequence analysis. It allows you to input sequence data, to view the aligned sequence in a textual format, to view the resolved DNA sequences, and to access the underlying data as you deem necessary.

4.1 Project Options

4.1.1 New Project

Pressing the 'New Project' icon will clear the FLiPARS window of any data in the current project. If it hasn't been saved, you will be prompted regarding whether or not you wish to save your changes.



Figure 2: Project Options Bar. Icons in order from left to right: New Project, Open Project, Save Project, Settings, Help

4.1.2 Open Project

The 'Open Project' button allows you to re-open previous FLiPARS project files. FLiPARS project files end with the *.flprs extension and can also be opened from your computer's filesystem directly by double-clicking on them (assuming FLiPARS is already installed).

4.1.3 Save Project

The 'Save Project' button allows you to save changes to a project, or to save that project for the first time. You will be prompted with a dialog to determine where to save the project file. If the current project has already been saved in a certain location once, subsequent clicks on the 'Save' icon will automatically write

any changes that you have made since the last change to that file. Progress of the saving operation is displayed in the progress bar in the lower-left corner of the main window.

4.1.4 FLiPARS Settings

The 'FLiPARS Settings' button brings up the control panel for global FLiPARS settings, such as available reference sequences and available primers. Please see section 8 on page 24 for more information on using this window.

4.1.5 Help

Pressing the 'Help' icon will bring up the FLiPARS documentation website if your computer is connected to the internet. Otherwise, it will bring up the local .pdf manual that was distributed when FLiPARS was installed.

4.2 Input File List

Pressing the 'Add Files' button will open a dialog that enables you to browse your computer to the appropriate data files. A basic sanity check on the file extensions is done before adding the files to the current set to ensure that the file is supported by an installed plugin. You can also select multiple files in the current project and remove them via the 'Remove Files' button.

4.2.1 Edit Primers

Setting the primers for your sequence data may be necessary if the sequence information is on the opposite strand from what the reference sequence was generated from. You can select multiple files in the file input box, and using the 'Edit Primers' drop-down menu, select which installed primer was used to generate the sequences. There is no requirement that all primers in a single file set must be the same, but linkage phase analysis will not work properly if the sequences are not from the same DNA region.

4.2.2 Re-order Files

Selecting files and pressing the 'Up' or 'Down' buttons will reorder those files in the sequence and chromatogram displays. All sequences are labeled with their file name and primer as well, regardless of their ordering. If files are re-ordered while the chromatogram display is open, that display will need to be closed and re-opened in order for the changes to take effect in that window.

4.2.3 Choose Reference

Different reference sequences may be added to the software (see Section 8 on page 24) if necessary. This simply changes what the system is attempting to align each sequence to. FLiPARS may not function properly if there is no valid reference sequence for the region that you're examining.

4.2.4 Choose Aligner

Different alignment algorithms may also be added to the software (see Section 8 on page 24) if necessary. The default algorithm provided with FLiPARS uses a simple Smith-Waterman, end-space free variant to align sequences against the provided reference. Indels are placed towards the 3' end of the reference sequence when ambiguities arise. An alternative 'ForensicD' algorithm is also provided. The difference between these two aligners is described on page 25.

4.3 Parse Files

Pressing the 'Parse Files' button will cause the system to make base-calls for each of the provided input files as well as alignments against the current reference sequence based on those base calls. If, after parsing, two or more mixed bases are called and aligned at the same position, all other single-base calls at that position will automatically be reclassified as pure separations for later analysis.

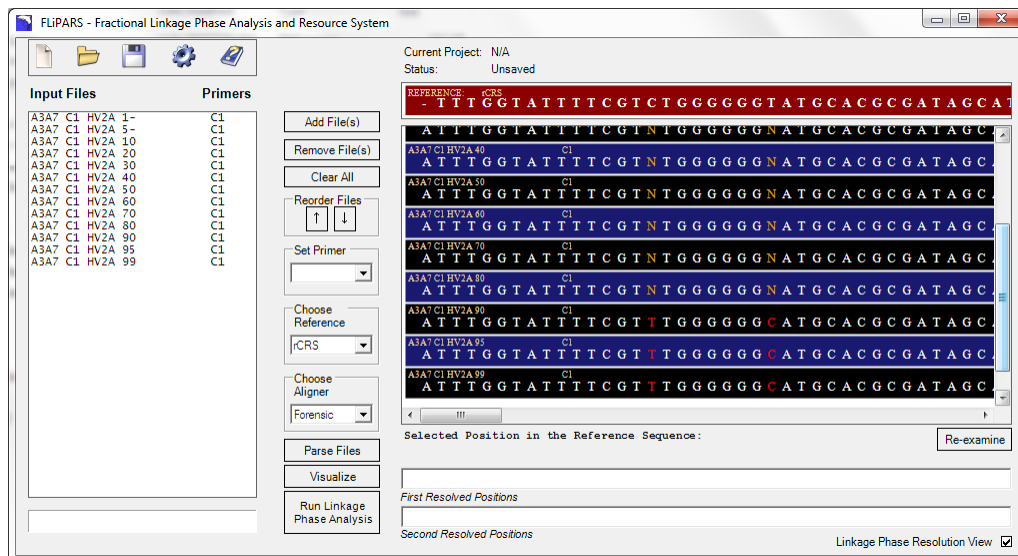


Figure 3: The main window after some files have been input and parsed. Base calls that are considered to be pure separations (100%/0% relative peak heights) are colored in red.

4.4 Visualize

Pressing the 'Visualize' button opens the chromatogram display window (see Section 5 on page 16) for viewing the raw signal data and the associated peak calls. The chromatogram window is the gateway to making modifications to the sequences (if needed). If the sequences have not already been parsed, they will be before the chromatogram display window is shown.

4.5 Run Linkage Phase Analysis

Pressing the 'Run Linkage Phase Analysis' button will open the linkage phase analysis window (see Section 7 on page 22). If the sequences have not already been parsed, they will be before the linkage phase analysis window is shown. Once this button has been pressed, the resolved sequences boxes at the bottom of the main window should be populated if the analysis was successful.

4.6 Sequence Display Box

The sequence display box displays both the reference sequence (at the top with a red background) and the aligned and parsed sequences with alternating black/blue backgrounds. Each individual sequence is labeled with its file name followed by the primer that was associated with it so that some portion of that information is always visible. Mixed base positions are highlighted in orange and are labeled as 'N'. Mismatched bases (when aligned with the reference sequence) are colored gray to distinguish them from a match. Clicking on a base-call letter in the sequence display will select that base call, boxing it in orange, and display its aligned position in the reference sequence just below the sequence display box. If it is a mixed base, the signals that

make up that mixture will also be displayed below its aligned position within the reference. Clicking within an already selected base-call will open up the chromatogram editor, select that base-call, and attempt to center the view over the selected peak.

4.7 Re-examine

The 'Re-examine' button is provided solely for convenience. It is possible to perform all operations without it, but it will greatly speed up some potentially monotonous and time-consuming tasks. For certain sequences, it has been observed that legitimate mixed peaks will occasionally be offset horizontally some distance from each other, causing the automatic base calling to miss the fact that there should be mixed calls present. When this occurs, it is useful for the analyst to be able to combine the appropriate peaks in those sequences where they are visible and to allow the system to automatically mark the other aligned (pure) single base calls at the same position to be 'Considered Mixed' for linkage phase analysis. The 'Re-examine' button simply tells the system to go back into the current set of base calls and to search for *aligned* positions containing 2 or more *mixed* base calls. If it finds such a position, it will first find the same base call(s) at the aligned position in every other sequence in the set. It will then try to 'promote' each of these base calls to a mixture if it can find an underlying peak strong enough and close enough (horizontally) to combine them. If it cannot promote any underlying peaks, it will automatically mark the single base-calls at that position to be 'considered mixed'.

4.8 Resolved Sequences

The resolved sequences boxes will only be populated once linkage phase analysis has been run. Confidence values for those determinations can be obtained in the linkage phase analysis window (see Section 7 on page 22).

4.8.1 Linkage Phase Resolution View

Two views for the results of the linkage phase analysis are provided in the main window via a checkbox below the first and second resolved sequence boxes. When the checkbox is checked (the default) the software will display the results in a so-called 'linkage phase resolution' mode. This display shows each mixed base that was resolved, regardless of whether or not the resolution matches the reference sequence. When the box is not checked, the display will switch to showing the standard polymorphic designations of each sequence as compared against the reference sequence (in other words, only the differences of each sequence with regard to the reference are shown - otherwise, the other positions are assumed to be identical to the reference).



Figure 4: Each resolved (mixed) position is displayed for both sequences.

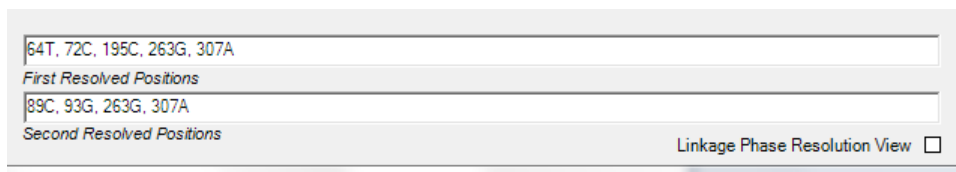


Figure 5: Only differences from the reference (including non-mixed positions) are displayed for both sequences.

5 Chromatogram Display

The chromatogram display shows a side by side view of the raw data for the project files you provided. Note that no alignment information is used here - any appearance of the sequences being aligned is purely coincidental. It is also important to remember that if your data was generated using a 'reverse' primer (that is, one that sequences the opposite strand in the opposite direction of the reference strand) then the sequence data will automatically be reverse complemented. Hence, when the chromatogram display shows a 'A' signal, for instance, that is in reality the 'T' signal for the actual sequencing data. In addition, the data will be reversed so that you can see the corresponding 5' to 3' (on the 'forward' strand) sequence.

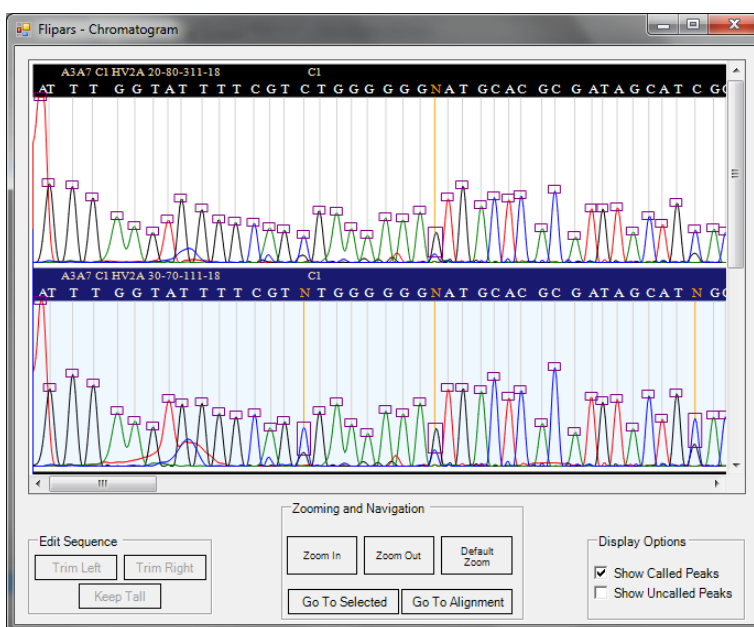


Figure 6: The chromatogram display for a set of fractionated sequence data

5.1 Edit Sequence Controls

The 'Edit Sequence' controls allow you to make some large scale modifications to the sequence. In particular, these editing tools are intended to be used on noisy (the beginning and/or end of the sequence, typically) portions of sequences. Fine-tuned editing of specific positions is handled via the pop-up editor dialog: See Section 6 on page 19 for more information. To open the pop-up editor for a specific base, you must first select that base in the chromatogram display by clicking on its corresponding peak (un)call box. Clicking in the selected box again will bring up the appropriate editor dialog.

5.1.1 Trimming Left and Right

Trimming left and right will invalidate all peak calls to the right or left of the selected base (including that base). This is useful for removing the noisy ends of a particular sequence.



Figure 7: An (unrealistic) example of how trimming in the chromatogram works. Trim left was used to remove all base calls to the left of a particular position - the invalidated peak calls are circled in yellow.

5.1.2 Keep Tall

The 'Keep Tall' button is used to invalidate a mixed base position by removing the shorter of the two peaks in the mixture. This situation can occur in noisy portions of sequences, where the normal base-calling criteria may not be sufficient to differentiate between a real mixture and a 'fake' one. Practitioner discretion is needed in using this particular tool, and as such, it should *not* be used to massage the sequence information to give you 'what you expect'. Also note that whenever these functions are used, the appropriate sequence will automatically be re-aligned against the reference sequence - the user doesn't have to worry about where in the sequence their change(s) occur: Rather, the user should only be concerned with what the data is really showing.

5.2 Zooming and Navigation Controls

The zooming and navigation controls provide chromatogram movement functions to enable the practitioner to examine relevant portions of the sequence easily.

5.2.1 Zooming

The zooming functions will shrink or expand the visible chromatograms appropriately. The 'Default Zoom' button will return the zoom to its original setting.

5.2.2 Go To Selected

The 'Go To Selected' button is used when you have selected a particular peak (un)call box in the chromatogram and scrolled elsewhere within the sequences. Oftentimes, it can be difficult to relocate where you had something selected simply by scrolling back through the sequences. Clicking on the 'Go To Selected' button will automatically center (if possible) the selected peak box in the viewable portion for you.

5.2.3 Go To Alignment

The 'Go To Alignment' button will automatically center the sequence view (Section 4 on page 12) of the main window over the selected peak (un)call box. It will also select that peak's corresponding base call in the main window for you.

5.3 Display Options Controls

5.3.1 Display Called Peaks

Draws violet colored boxes over all called peaks when checked

5.3.2 Display Uncalled Peaks

Draws cyan colored boxes over all uncalled peaks when checked

6 Pop-up Sequence Editor

The pop-up sequence editor provides a zoomed-in version of the sequence data where you may be considering making changes to the automated base-calls. This encourages good QA/QC practices by forcing the practitioner to examine in detail the data underlying their proposed changes.

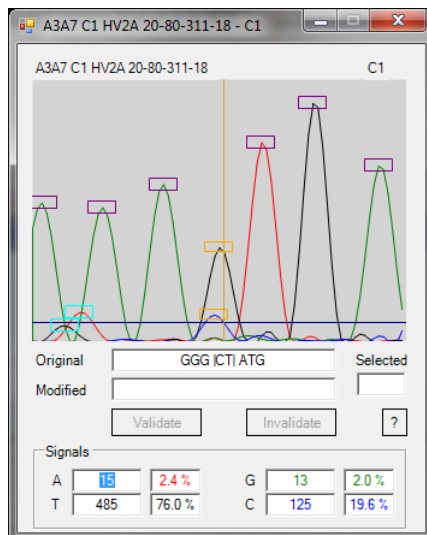


Figure 8: The pop-up sequence editor for a particular region containing a mixed base

6.1 Validating or Invalidating Peaks

The main function of the pop-up editor window is to allow you to either validate (insert) peaks that the automated base-calling missed, or to invalidate (delete) some peaks that the base-calling made. Such determinations should be made with caution by an expert at sequence analysis.

Violet boxes represent called peaks, while cyan boxes represent uncalled peaks. Original (when the pop-up editor was opened) sequence information in the displayed region is shown below the graph, along with the modified sequence (if any) in the event that you want to 'undo' your changes. Clicking in any one of these boxes will select that peak, highlight it, position the vertical slider over the peak, update the signal information at the bottom of the editor for that position, and display what peak signal you have selected. If an uncalled peak is selected, only the 'Validate' button will be enabled. Likewise, if a called peak is selected, only the 'Invalidate' button will be enabled. Using either button in their respective situations will either insert or remove that peak into or out of the sequence, and the sequence will then be realigned. After being realigned, all of the various displays will be updated accordingly with the updated information.

NOTE: There is no undo functionality built into FLiPARS! While convenient, it also makes modifying the sequences without regard to the biological reality trivial. As such, if you make a mistake in editing your sequences, YOU have to go back in to fix it, one base at a time. Modify with care!

6.1.1 Selecting Overlapping Peaks

On occasion, peaks may be located at nearly the same index and be of nearly the same height, causing the boxes that the system draws over those calls to overlap. When you click inside one of these overlapping

positions, the system will prompt you regarding which peak signal you wish to select. The editor again will display the selected signal so that you can remember exactly which peak you have selected.

6.1.2 Creating a Mixed Base Position

If you validate two peaks that are close enough to each other, the system will prompt you to see if you want to promote those two peak calls into one single mixed base position. Creating mixed base calls where only one peak exists, however, is not possible. You may also hold the CTRL key and click on a second peak (which will be highlighted in green instead of the normal red) if the two peaks that you wish to merge into a single mixed base call are too far apart. Once a second peak is selected, the 'Merge' button will be enabled - it simply creates a single mixed base call out of the two selected peaks. Note that one of the selected peaks must already be a validated peak call before merging will work - merging two non-calls together is not possible.

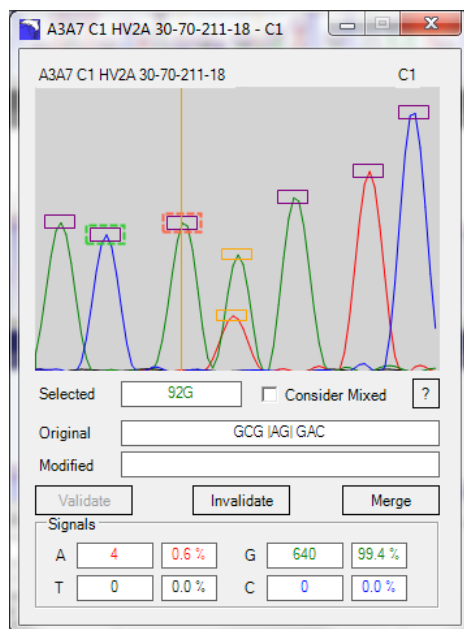


Figure 9: Multiple selections and merging. The red box shows the 'Selected' peak call (the call whose information is displayed in the editor) while the target peak to merge with the selected peak is highlighted in green. For merging to work, one of the two selected peaks must have already been validated as a legitimate base call.

6.2 Examining Signal Data

The pop-up editor automatically positions the vertical slider over the selected peak positions and updates the signal information accordingly. You can also move this slider manually anywhere within the displayed data region. To do so, *right-click* anywhere in the graph. The background will turn white, and moving your mouse within the graph will also move the vertical slider and update the signal information at the bottom. You are not able to select peak calls while examining the graph. To stop the vertical slider at a specific position, right-click again, and the background will once again turn gray and the selection and editing functionality of the editor will again be available.

6.3 Consider Mixed

An option is provided to manually consider some single base calls as a 'mixed' position. This has the effect of producing a 100% / 0% (pure) statistical value if linkage phase analysis is run on the aligned position of that base. This situation occurs when a pure separation is obtained. In known control sequences, mixed peaks will usually begin to appear when the minor contributor comprises 20% of the DNA in the mixture. Base calls that are considered to be mixed are colored in red in the main window parsed sequences display.

7 Linkage Phase Analysis Window

The linkage phase analysis window exposes all of the underlying mathematical data that was calculated for any combination of the fractions and/or mixed bases found in the current project. When first opened, it automatically selects the two fractions that provide the largest average shift across the most number of mixed base positions. In other words, if a pair of fractions only contains 3 *aligned* mixed base positions while another contains 4, the pair of fractions containing the 4 mixed base positions will always be chosen over the pair of fractions with 3, regardless of the average shift values.

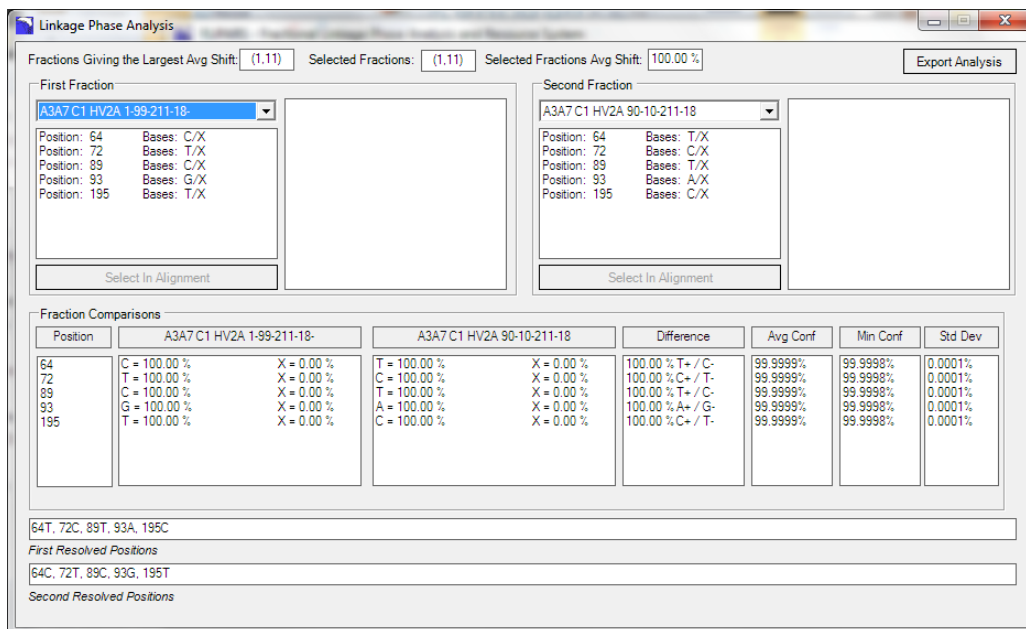


Figure 10: The linkage phase analysis window as it appears when first opened

7.1 First and Second Fractions

The top of the linkage phase analysis window provides display and selection tools for two of the fractions provided in the data set. At the very top of the window, the indices (starting with 1 at the top of the list) of the fractions currently selected as well as the two fractions detected as providing the largest average shift over the most aligned mixed bases. You can also select a specific mixed base position in either fraction and a zoomed in graph of the mixed signal being referenced will be drawn immediately next to the data so that you can literally view the signal shifts between any aligned mixed-base positions in any two fractions.

7.1.1 Drop-down File Lists

The drop-lists allow you to select any two fractions in the data set for comparison. The indices referenced at the top of the linkage phase analysis window start with 1 at the top of each list and increase downward.

7.1.2 Select In Alignment

When you have a specific mixed base position in a specific fraction selected, the 'Select in Alignment' button will be enabled. This selects that base in the main window's sequence display and centers your view over that selected position. This enables you to jump from the linkage phase analysis window to the main window to the chromatogram display window (using the Main Window's 'Visualize' button) should you desire to.

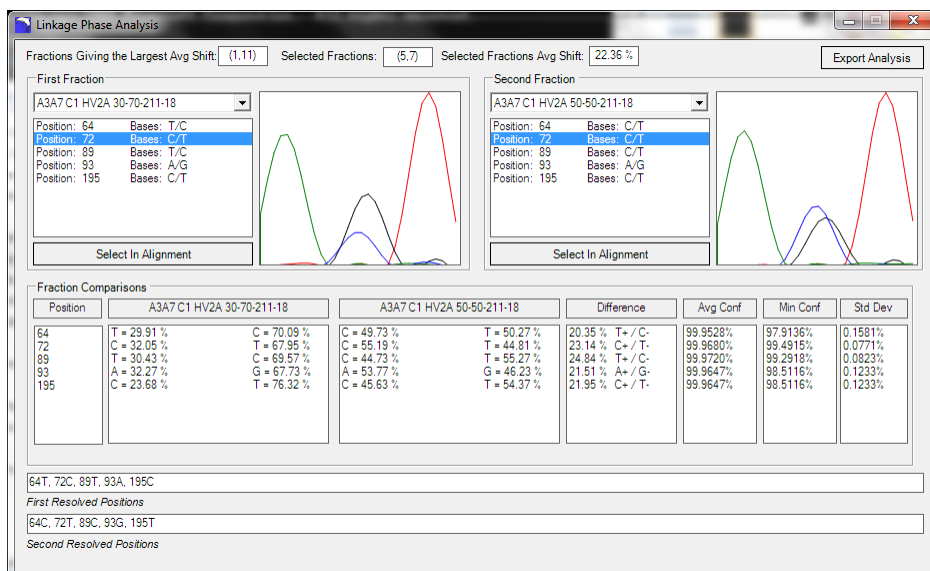


Figure 11: The linkage phase analysis window after selecting two particular mixed base positions from each displayed fraction

7.2 Fraction Comparisons

The fraction comparisons box displays the raw data for each fraction and mixed base position. It shows the percentage of the peak height that each signal represents in that particular fraction, the difference between the two, and the confidence values associated with those calculations.

7.3 Resolved Sequences

The resolved sequences boxes displays *ONLY the resolved mixed base positions*. If all of the fractions possess a polymorphism at a non-mixed site, that change will not be listed here. Consistent polymorphisms will be shown in the main window's resolved sequences boxes when the 'Linkage Phase Resolution View' checkbox is *disabled*.

8 Settings

The settings window provides various options that can be set and/or adjusted by the user.

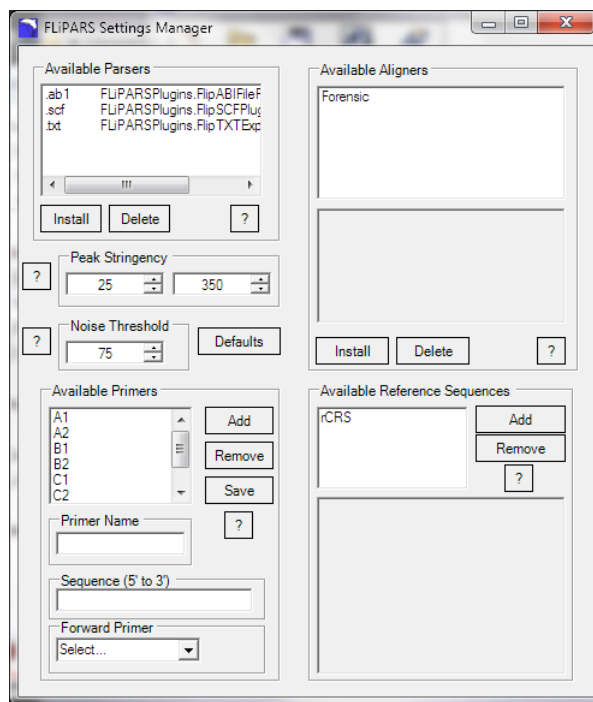


Figure 12: The default settings window when FLiPARS is first installed

8.1 Available Parsers

FLiPARS provides a plugin system that provides extensibility in the future for support parsing of new file types. Developers can write plugin libraries using the .NET platform, at which point those libraries can be added to your FLiPARS installation with minimal effort. These plugins can be installed or removed as seen fit for your laboratory's uses. Once a plugin has been added, the system knows what types of files it can parse and will automatically utilize the plugin as needed.

WARNING: DO NOT INSTALL PLUGINS FROM UNTRUSTED SOURCES. MALICIOUS PLUGINS COULD DO HARM TO YOUR COMPUTER - THE SAFETY OF YOUR DATA AND COMPUTER IS YOUR RESPONSIBILITY.

8.1.1 Add Plugin Form

Pressing the 'Install' button in the plugins box of the settings page brings up another window that asks for a plugin file. Plugin files will end with the '.dll' extension. Once you locate the file on your computer and provide it to the plugin form, some diagnostic information will be displayed in the window regarding the provided file. If the file *appears* to be a valid FLiPARS plugin, a "Success!" message will be displayed and the 'Add Plugin' button will be available. Pressing that button will add that plugin to the FLiPARS system.

8.2 Peak Stringency

Peak stringency settings affect the base-calling algorithm that FLiPARS uses. In short, the system looks at the derivative of the sequence data and identifies peaks where the sign of that derivative changes. The amplitude of this sign-change is what the 'Upper' and 'Lower' threshold values relate to. In order to qualify as a peak, that amplitude must be at least the lower threshold, but less than the upper threshold. The 'Noise Threshold' determines the minimum height needed to qualify as a legitimate peak. Any peaks under this height, regardless of the amplitude of the changes of those peaks, will not be called. The units are assumed to be RFUs.

8.3 Available Primers

The available primers settings allow you to add or remove primers in use in your particular laboratory. Primer information is used in FLiPARS to determine whether or not to reverse-complement the sequences being analyzed for alignment purposes. Primer sequences are purely present for housekeeping purposes. Forward primers are defined to be sequencing 5' to 3' in the same direction as the reference that it is sequencing in. Reverse primers, then, are sequencing the opposite strand 5' to 3' and will be reverse-complemented as needed.

8.4 Available Reference Sequences

Available reference sequences give laboratories flexibility in what types of scientific applications they might be using FLiPARS for. By default, FLiPARS only comes with the revised Cambridge Reference Sequence for forensic applications. Selecting an installed sequence will display a textual representation of that reference. Each reference can be removed if desired, though FLiPARS will NOT function without a reference to align fractions against.

8.4.1 Add Reference Sequence Dialog

Adding a reference consists of simply providing an identifying (unique) name for that reference sequence, and the raw ATGC (N if necessary) sequence, without numbers or spaces.

8.5 Available Aligners

Available aligners provide customizable alignments to be performed if the Smith-Waterman end-space free variant doesn't work for your application. The alignment algorithm provided by default with FLiPARS strives to conform to the FBI's published guidelines for forensic consistency[31].

Unfortunately, not all alignments done either by FLiPARS or other commercial forensic applications get every single alignment correct. Because of this, two alignment plugins are provided to the end user with FLiPARS by default. As a general rule of thumb, it may be good practice if you see mixed bases and/or indels within 10 bases of each other to look at the final aligned results with both aligners and to make a judgment about the correct alignment from there. Descriptions of both aligners follow.

NOTE: Be aware that in *most* cases your choice of the two aligners provided by FLiPARS will not matter. Both of them first search for the shortest 'edit distance' between the sequence and the reference. Resolution of ambiguities when multiple alignments are possible with the same number of edits is what differentiates these two aligners.

8.5.1 Default 'Forensic' Aligner

The default aligner, 'Forensic', produces alignments that are very similar to those produced by certain commercially available forensic sequence analysis applications. Some cases have been observed, however, where

this algorithm provides an artificial alignment that does not reflect the true nature of the sequence. Generally speaking, the 'Forensic' aligner will try to use substitutions (transitions are favored over transversions) to produce an optimal alignment when ambiguities arise.

8.5.2 Alternate 'ForensicD' Aligner

The alternative aligner, 'ForensicD' will *usually* give the same alignment as the 'Forensic' aligner. However, it will generally try to use indels instead of substitutions to provide an optimal alignment when ambiguities arise, which in some cases is preferable.

8.5.3 Add Aligner Form

Pressing the 'Install' button in the plugins box of the settings page brings up another window that asks for a plugin file. Plugin files will end with the '.dll' extension. Once you locate the file on your computer and provide it to the plugin form, some diagnostic information will be displayed in the window regarding the provided file. If the file *appears* to be a valid aligner file, the 'Install' button will be enabled. Enter a unique (to the system) name for the new aligner ('Forensic' is the name of default aligner provided by FLiPARS). Pressing that button will permanently add that plugin to the FLiPARS system.

9 Common Tasks

This section provides brief, step-by-step instructions for common tasks - in case you don't / didn't want to dive into the full-fledged documentation in previous sections.

9.1 Add New Reference Sequences

Step 1: Open the Settings Window by pressing the 'Settings' icon in the main window.

Step 2: Click on the 'Add' button in the 'Available Reference Sequences' box.

Step 3: Enter a unique name for the reference sequence you are adding.

Step 4: Copy / paste the ATGC(N) ONLY textual sequence into the 'Sequence' text box. This text is assumed to be in 5' (beginning) to 3' (end) order.

Step 5: Press the 'Add Sequence' button.

9.2 Add New Primers

Step 1: Open the Settings Window by pressing the 'Settings' icon in the main window.

Step 2: Click on the 'Add' button in the 'Available Primers' box.

Step 3: Enter a unique name for the new primer.

Step 4: Set whether or not the primer is 'Forward' (sequences in the same direction the associated reference sequence) or 'Reverse' (sequences the strand opposite the reference sequence).

Step 5: Optionally add the primer's ATGC sequence.

Step 6: Press the 'Save' button in the 'Available Primers' box.

9.3 Change Existing Primers

Step 1: Open the Settings Window by pressing the 'Settings' icon in the main window.

Step 2: Select the primer that you want to change from the list of primers in the 'Available Primers' box.

Step 3: Make the appropriate modifications (change the sequence, name, or forward/reverse). A message will appear saying that the changes are 'Unsaved'

Step 4: Press the 'Save' button in the 'Available Primers' box.

9.4 Edit Sequences

There are a couple of ways to edit sequences in FLiPARS. You can use the high-level tools provided by the chromatogram display (section 5 on page 16). The following steps refer to using the pop-up editor's tools, as any edit may be made in that manner, unlike the chromatogram display edits. NOTE: Your sequences will be realigned on the fly as you make changes so that you can be concerned with what the experimental data shows instead of having to worry about how to make any necessary alignment changes.

Step 1: Add sequence files to the current project.

Step 2: Press the 'Parse Files' button in the main window.

Step 3: Find a region of questionable base-calling accuracy in the main window sequences. (Alternatively, you can press the 'Visualize' button and search for questionable calls in the chromatograms).

Step 4: Click on the questionable call that you'd like to examine. If it is a deletion that appears out of the ordinary, click on an adjacent call to that deletion.

Step 5: Either click on the 'Visualize' button in the main window or click again inside of the selected box around the questionable base-call. (Skip this step if you went straight to the chromatogram display to locate an area of the sequences to edit)

Step 6: The peak call that was selected will be highlighted with a dashed red/orange line and close to the center of the display. Click inside the peak-call box to open the pop-up editor.

Step 7: Select a peak call that you want to delete (Invalidate - called peaks will be boxed in violet) or to insert (Validate - these peaks will be boxed in cyan).

Step 8: Press the appropriate Valildate/Invalidate button in the pop-up editor window. The system will pause while it processes the modified sequence data.

That's it! Alignments and where exactly in the sequence to add the appropriate base call is handled internally for you so that you can focus on what is scientifically relevant and not worry about how to make the computational aspects work.

9.5 Adding New File Input Plugins

Step 1: Open the Settings Window by pressing the 'Settings' icon in the main window.

Step 2: Click the 'Install' button in the 'Available Plugins' box.

Step 3: Press the 'Browse' button in the 'Add New File Parser' dialog.

Step 4: Locate and select the desired plugin (ending in .dll) on your computer and press 'Open'.

Step 5: FLiPARS will perform some brief sanity checks on the provided file. If all is well and good, you can press the 'Add Plugin' button.

9.6 Align Sequences

Sequences NEVER need to be aligned manually. Whenever you either validate an uncalled peak (make an insertion into the experimentally supported dataset) or invalidate (delete a base-call from the experimental dataset) FLiPARS will automatically realign the sequence against its reference based on the new modifications. Standard alignment definitions exist, and computers are good at calculating that kind of information. FLiPARS gives the practitioner the tools to look at the scientifically relevant information that the sequencing reactions are providing and tries to hide details that relate to the computational underpinnings as much as possible. While important to understand, the practitioner should not be burdened with something that all of our computational tools were designed to automate to the greatest extent possible.

10 For Developers

FLiPARS was designed to be extensible for future sequencing and alignment technologies. The FLiPARS source code is freely available online at www.flipars.com and is released under a BSD license (See page 5). In particular, FLiPARS 2.0 supports plugins for parsing new (or old) sequencing file formats as well as for performing customized alignments. This section outlines in brief the requirements for developing such a plugin.

10.1 File Input Plugins

New file parsing plugins may be developed in most languages, provided they are .NET compatible. The vast majority of popular languages already have their own .NET compilers, including (but not limited to) C, C++, C#, Basic, Python, and Perl. Many uncommon languages are available as well, including Prolog, Haskell, LISP. See www.dotnetpowered.com/languages.aspx for an unofficial list of available compilers and their respective homepages. Wikipedia also has a list of .NET compatible languages for your perusal.

FLiPARS requires all new file parsing plugins to support the `libFLiPARS.IFliparsInputPlugin` interface, which consists of 3 methods: `retrievePoints`, `retrieveCalls`, and `retrieveExtensionSupport`. That's all there is to it. Develop your .NET compatible plugin to implement the `IFliparsInputPlugin` interface, build it into a .NET library (*.dll) and distribute/install it on the machines that will need to use it.

```
public interface IFliparsInputPlugin
{
    /// <summary>
    /// Retrieves the raw sequencing data points from the location specified in x
    /// </summary>
    /// <param name="Location">Location of the data for the plugin to parse</param>
    /// <returns>A list of 4 lists of double values corresponding to the raw
    /// electropherogram data. 0=A, 1=T, 2=G, 3=C</returns>
    List<List<double>> retrievePoints(FileInfo Location);

    /// <summary>
    /// Retrieves only the called bases present in the location specified in x (if
    /// present). The method should return null if the sequencer did not store such
    /// information in the specified location. In this case, the data parser will call
    /// retrievePoints() and proceed that way regardless of the users settings
    /// </summary>
    /// <param name="Location">Location of the data for the plugin to parse</param>
    /// <returns>List of constructed nucleotides from the file. Null if calls are not
    /// present</returns>
    List<Nucleotide> retrieveCalls(FileInfo Location);

    /// <summary>
    /// Retrieves the file extension that this plugin supports. Each plugin is assumed
    /// to support only one file extension
    /// </summary>
    /// <returns>String representing the supported file extension (including the dot) -
    /// ie ".txt"</returns>
    string retrieveExtensionSupport();
}
```

10.2 Sequence Alignment Plugins

New file parsing plugins may be developed in most languages, provided they are .NET compatible. The vast majority of popular languages already have their own .NET compilers, including (but not limited to) C, C++, C#, Basic, Python, and Perl. Many uncommon languages are available as well, including Prolog,

Haskell, LISP. See www.dotnetpowered.com/languages.aspx for an unofficial list of available compilers and their respective homepages. Wikipedia also has a list of .NET compatible languages for your perusal.

FLiPARS requires all new alignment plugins to support the libFLiPARS.IFliparsAlignmentPlugin interface, which consists of 2 methods: align and getDescription. That's all there is to it. Develop your .NET compatible plugin to implement the IFliparsAlignmentPlugin interface, build it into a .NET library (*.dll) and distribute/install it on the machines that will need to use it.

```
public interface IFliparsAlignmentPlugin
{
    /// <summary>
    /// Aligns the sequence contained in seq against the reference sequence in refSeq.
    /// The alignment information is saved directly into the seq object's
    /// EditTranscript property.
    /// </summary>
    /// <param name="seq">Sequence to be aligned.</param>
    /// <param name="refSeq">Reference sequence to be aligned against.</param>
    /// <returns>Success or failure of the alignment operation</returns>
    bool align(Sequence seq, string refSeq);

    /// <summary>
    /// Gets the description (if any) for this particular alignment algorithm.
    /// </summary>
    /// <returns>Description of this alignment function implementation.</returns>
    string getDescription();
}
```

11 Acknowledgements

FLiPARS

Development of this project was supported by the National Institute of Justice under Award 2009-DN-BX-K047.

Previous background research on linkage phase analysis was done by Richard Kristinsson and Phillip B. Danielson under NSF grant 0200484 and NIJ grants 2003-IJ-CX-K104 and 2006-DN-BX-K002.

Jesse Goeglein

I would like to thank Dr. Phil Danielson for the opportunity to work in his lab and on this project - without you I wouldn't be where I am professionally today.

I have to acknowledge my parents and family for all of their love and support - thank you for everything! You've meant more than you know.

I'd also like to extend a special thanks to The Boettcher Foundation, Tim Schultz, Molly Smith, Katy Craig, Katy Kramer and the rest of the Boettcher family and staff. Thank you from the bottom of my heart for everything that you've done for me. Your support genuinely made this project possible.

References

- [1] Budowle, B., et al., DNA Typing Protocols: Molecular Biology and Forensic Analysis. Forensic Science Series. 2000: Biotechniques Books. 36.
- [2] Holland, M.M. and T.J. Parsons, Mitochondrial DNA Sequence Analysis-Validation and use for forensic casework. Forensic Science Review, 1999. 11(1): p. 22-49.
- [3] Holland, M.M., et al., Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War. Journal of Forensic Sciences, 1993. 38(3): p. 542-53.
- [4] Ivanov, P.L., et al., Mitochondrial DNA sequence heteroplasmy in the Grand Duke of Russia Georgij Romanov establishes the authenticity of the remains of Tsar Nicholas II. Nat Genet, 1996. 12(4): p. 417-20.
- [5] Hagelberg, E., I.C. Gray, and A.J. Jeffreys, Identification of the skeletal remains of a murder victim by DNA analysis. Nature, 1991. 352(6334): p. 427-9.
- [6] Allen, M., et al., Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: sensitivity and matching probabilities. J Forensic Sci, 1998. 43(3): p. 453-64.
- [7] Budowle, B., et al., Mitochondrial DNA regions HVI and HVII population data. Forensic Science International, 1999. 103(1): p. 23-35.
- [8] Parsons, T.J., et al., A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet, 1997. 15(4): p. 363-8.
- [9] Fournery, R.M., Mitochondrial DNA and Forensic Analysis: A Primer for Law Enforcement. Can. Soc. Forens. Sci. J, 1998. 31(1): p. 45-53.
- [10] Wilson, M.R., et al., Validation of mitochondrial DNA sequencing for forensic casework analysis. Int J Legal Med, 1995. 108(2): p. 68-74.
- [11] Melton, T., et al., Forensic mitochondrial DNA analysis of 691 casework hairs. J Forensic Sci, 2005. 50(1): p. 73-80.
- [12] Hanekamp, J.S., W.G. Thilly, and M.A. Chaudhry, Screening for human mitochondrial DNA polymorphisms with denaturing gradient gel electrophoresis. Hum Genet, 1996. 98(2): p. 243-5.
- [13] Barros, F., et al., Rapid and enhanced detection of mitochondrial DNA variation using single-strand conformation analysis of superposed restriction enzyme fragments from polymerase chain reaction-amplified products. Electrophoresis, 1997. 18(1): p. 52-4.
- [14] Alonso, A., et al., Rapid detection of sequence polymorphisms in the human mitochondrial DNA control region by polymerase chain reaction and single-strand conformation analysis in mutation detection enhancement gels. Electrophoresis, 1996. 17(8): p. 1299-301.
- [15] Steighner, R.J., et al., Comparative identity and homogeneity testing of the mtDNA HV1 region using denaturing gradient gel electrophoresis. Journal of Forensic Sciences, 1999. 44(6): p. 1186-98.
- [16] Andreasson, H., et al., Forensic mitochondrial coding region analysis for increased discrimination using pyrosequencing technology. Forensic Sci Int Genet, 2007. 1(1): p. 35-43.
- [17] Andreasson, H., et al., Mitochondrial sequence analysis for forensic identification using pyrosequencing technology. Biotechniques, 2002. 32(1): p. 124-6, 128, 130-3.

- [18] Hall, T.A., et al., Base composition analysis of human mitochondrial DNA using electrospray ionization mass spectrometry: a novel tool for the identification and differentiation of humans. *Anal Biochem*, 2005. 344(1): p. 53-69.
- [19] Steven A. Hofstadler, P.D., Analysis of DNA Forensic Markers Using High Throughput Mass Spectrometry, in Final report: NIJ Award #2006-DN-BX-K011. 2008. p. 117.
- [20] Andreasson, H., et al., Quantification of mtDNA mixtures in forensic evidence material using pyrosequencing. *Int J Legal Med*, 2006. 120(6): p. 383-90.
- [21] Huber, C.G., P.J. Oefner, and G.K. Bonn, High-resolution liquid chromatography of oligonucleotides on nonporous alkylated styrene-divinylbenzene copolymers. *Anal Biochem*, 1993. 212(2): p. 351-8.
- [22] Huber, C.G., P.J. Oefner, and G.K. Bonn, Rapid and accurate sizing of DNA fragments by ion-pair chromatography on alkylated nonporous poly(styrene-divinylbenzene). *Analytical Chemistry*, 1995. 67: p. 578-585.
- [23] Emmerson, P., et al., Characterizing mutations in samples with low-level mosaicism by collection and analysis of DHPLC fractionated heteroduplexes. *Hum Mutat*, 2003. 21(2): p. 112-5.
- [24] Etokebe, G.E., et al., Physical separation of HLA-A alleles by denaturing high-performance liquid chromatography. *Tissue Antigens*, 2003. 61(6): p. 443-50.
- [25] Steighner, R.J., et al., Comparative identity and homogeneity testing of the mtDNA HV1 region using denaturing gradient gel electrophoresis. *J Forensic Sci*, 1999. 44(6): p. 1186-98.
- [26] Budowle, B., et al., Forensics and mitochondrial DNA: applications, debates, and foundations. *Annu Rev Genomics Hum Genet*, 2003. 4: p. 119-41.
- [27] Danielson, P.B., et al., Separating human DNA mixtures using denaturing high-performance liquid chromatography. *Expert Rev Mol Diagn*, 2005. 5(1): p. 53-63.
- [28] Danielson, P.B., et al., Resolving mtDNA mixtures by denaturing high-performance liquid chromatography and linkage phase determination. *Forensic Sci Int Gen*, 2007. 1(2): p. 148-53.
- [29] Danielson, P.B., Mitochondrial DNA Analysis by Denaturing Liquid Chromatography for the Separation of Mixtures in Forensic Samples. Final Technical Report for 2003-IJCX-K104 submitted to the National Institute of Justice, Forensic DNA Research and Development Program, 2008: p. 1-105.
- [30] Vial, P., et al., Software tool for portal dosimetry research. *Australas Phys Eng Sci Med*, 2008. 31(3): p. 216-22.
- [31] Budowle, B., J.A. DiZinno, M.R. Wilson, Interpretation guidelines for mitochondrial DNA sequencing. Proceedings of the Tenth International Symposium on Human Identification, Promega Corporation, Madison, WI. www.promega.com/ussvmp10proc/default.html, 1999.