Final Technical Report

# Development of an Expert System for Automated Forensic Mitochondrial DNA Data Analysis

**FY 2009 Forensic DNA Unit Efficiency Improvement**

**The University of North Texas Health Science Center**

**Department of Forensic & Investigative Genetics**

**Project Duration:** January 1, 2010 to June 30, 2012

**Award Number:** NIJ Cooperative Agreement 2009-DN-BX-K171

**Principal Investigator:** Rhonda K. Roby, PhD, MPH

**Scientific Team:** Nicole R. Phillips, MS; Jennifer L. Thomas, MS; and Marc L. Sprouse

**Support Staff:** Hector Saenz, MT (ASCP)

# *Abstract*

With continued advancements in the laboratory processing of samples for mitochondrial DNA (mtDNA), sequence data are generated at a faster rate resulting in bottlenecks of mtDNA sequence analysis. The **research goal** of this project was to reduce the review time of mtDNA sequence data by using advanced expert system tools, while also decreasing subjectivity and error in haplotype reporting. The development of expert system tools for the analysis of mtDNA can greatly enhances throughput capabilities, reduces error, and significantly impacts the amount of time required for data review by the analyst.

The UNT Health Science Center (UNTHSC) has identified several steps in the analysis of mtDNA that can significantly reduce labor in data review; thus, reducing the overall analytical time. A reduction in labor and processing time will improve efficiency and increase the overall capacity of mtDNA processing by the laboratory. The focus for the improvements addressed in this project included software development for two new programs (eFAST™ Software v2.0, UNTHSC, Ft. Worth, TX, and STATIS, UNTHSC) and enhancements to an existing advanced software data analysis package (MTexpert™, MitoTech, LLC, Santa Fe, NM). eFAST™ Software (eFAST) and STATIS Software for mtDNA data review and management were designed and programmed by UNTHSC personnel. MTexpert™, an expert system for mtDNA data analysis, was enhanced by the software vendor under contract to UNTHSC based on evaluations and improvements recommended by the highly experienced forensic DNA analysts at UNTHSC.

The **objective** of this project was to create an expert system process that would review and parse raw sequence data, including and improving an advanced sequence analysis program to better automate the routine and repetitive tasks in interpretation of mtDNA sequence analysis. By this means, the project would improve laboratory efficiency and the speed of mtDNA data analysis. Expert system software programs were developed and updated to fully automate the analysis of high quality mtDNA raw and annotated sequence data. When the expert system was unable to complete the automated analysis of "challenged data," the software was designed to direct the analyst to the specific area for

review and serve as a decision support tool to aid the analyst in resolving the data issues, so as to deliver a result or direct the sample to reanalysis.

This project is a close collaboration between forensic science experts at UNTHSC and experienced expert system developers at MitoTech, LLC (MitoTech). The UNTHSC, with support by the National Institute of Justice (NIJ), is a recognized national center for mtDNA analysis especially as it applies to the identification of missing persons and human remains. The UNTHSC has unique expertise in mtDNA sample analysis and is the largest contributor to the U.S. National Missing Persons Program. MitoTech is a small business located in Santa Fe, New Mexico that specializes in developing software systems for automated sequence data analysis. The **research design** was to have regular communication with the scientific staff and the software developers to improve the overall streamlining process for the expert system software programs and to rapidly test any new software versions to quickly identify bugs and make improvements.

Further, MTexpert Software has a standard set of base calling rules known as MitoTyper Rules integrated into its software; these rules and the new expert system tools were tested by UNTHSC with their sequence data. With MTexpert, scientists can edit the mtDNA sequence traces, review the assembly, and evaluate each polymorphism with dynamic links from the traces' base calls through the assembly to the type entries. The prototype MTexpert software was used as a platform and UNTHSC tested the software, made recommendations for improvement, and tested each new version for bugs. In **summary**, with the three software programs linked contiguously, mtDNA sequence data can be evaluated quicker, with less error, and generate a haplotype report.

# *Table of Contents*

# *Executive Summary*

Mitochondrial DNA (mtDNA) analysis has proven to be an invaluable tool for victim identification from mass disasters and missing persons programs to criminal casework (Isenberg, 2004). The UNT Health Science Center (UNTHSC) is primarily funded by the National Institute of Justice (NIJ) for the Missing Persons Program and uses advanced DNA technologies to process unidentified human remains and the family reference samples from biological relatives for both nuclear DNA (nDNA) and mtDNA. Missing persons cases rely heavily on mtDNA testing of the skeletal remains and maternal relatives for making associations. The resulting DNA profiles are uploaded to the Missing Persons Index of the CODIS database. In this database, mtDNA and nDNA profiles from the unidentified human remains can be searched against the biological family reference profiles and associations are recommended through kinship analysis testing. There are several hundred thousand missing persons cases reported each year and there are more than 14,000 unidentified human skeletal remains retained in medical examiners' and coroners' offices nationwide (Roby *et al*., 2007). These numbers alone demonstrate the throughput requirements needed for DNA processing.

Mitochondrial DNA testing is a laborious process which includes amplifying and sequencing two regions in the mtDNA genome (mtGenome) (Holland *et al*., 1995) resulting in over 1000 bases for data review and analysis. The rate of data generation exceeds that of data analysis, review, and reporting; hence, creating a bottleneck in the final review of data, reporting, and upload into the Missing Persons Index database. This bottleneck illustrates the need for an expert system. To address this bottleneck, an expert system and process for high throughput data management has been designed, developed, and evaluated. The mtDNA typing protocol of the Family Reference Section of the Missing Persons Program at UNTHSC requires a first and second read in which the analysts evaluate the traces manually, base-by-base. This process is tedious and can lead to transcription errors. Utilizing quality metrics and an expert system for mtDNA analysis allow the software to filter through data without human intervention. Expert systems have the potential to streamline data analysis and reduce the backlogs. An expert system for sequence analysis would reduce the amount of

time an analyst must spend reviewing sequence data and therefore increase the throughput of a laboratory. Expert systems may also reduce the potential for human error, as the process is automated, consistent, and accurate. Implementation of expert systems within a laboratory reduces analysis time; consequently, freeing the analyst for other duties.

The research goal of this project was to reduce the review time for mtDNA sequence data by using expert system tools, while also decreasing subjectivity and error in haplotype reporting. The development of expert system tools for the analysis of mtDNA greatly enhances throughput capabilities, reduces error, and significantly impacts the amount of time required for manual data review by the analyst.

The UNTHSC has identified several steps in the analysis of mtDNA that can significantly reduce labor in data review; thus, reducing the overall examination time. A reduction in labor and processing time will improve efficiency and increase the overall capacity for mtDNA processing by the laboratory. The focus of improvements addressed in this project includes software development for two new programs (eFAST™ Software v2.0, UNTHSC, and STATIS Software, UNTHSC) and enhancements to an already designed software data analysis package (MTexpert™, MitoTech, LLC, Sante Fe, NM). For software development, eFAST™ Software (eFAST) and STATIS Software (STATIS) for mtDNA data review and management were designed by UNTHSC. MTexpert™, an expert system for mtDNA data analysis, was enhanced by the software vendor under contract to UNTHSC based on evaluations and improvements recommended by the highly experienced forensic DNA analysts at UNTHSC.

The objective of this project was to create an integrated expert system process that will review and parse raw sequence data and to improve a sequence analysis program to better automate the routine and repetitive tasks in interpretation of mtDNA sequence analysis, thereby improving laboratory efficiency and the speed of data analysis. Expert system software programs were developed and updated to fully automate the analysis of high quality mtDNA raw and annotated sequence data. When the expert system is unable to complete the automated analysis of

"challenged data," the software is designed to direct the analyst to the specific area for review and serve as a decision support tool to aid the analyst in resolving data issues, so as to deliver a result or direct the sample to reanalysis.

This project is a close collaboration between forensic science experts at UNTHSC and experienced expert system developers at MitoTech, LLC (MitoTech). The UNTHSC, with support by the National Institute of Justice (NIJ), is a recognized national center for mtDNA analysis especially as it applies to the identification of missing persons and human remains. The UNTHSC has unique expertise in mtDNA sample analysis and is the largest contributor to the U.S. National Missing Persons Program. MitoTech is a small business located in Santa Fe, New Mexico that specializes in developing software systems for automated sequence data analysis. The research design required regular communication with the scientific staff and the software developers to improve the overall streamlining process for the integrated expert system software and to rapidly test any new software versions to quickly identify bugs and make further improvements.

Further, MTexpert™ Software has a standard set of base calling rules known as MitoTyper™ Rules integrated into its software; these rules and the new associated expert system tools were tested by UNTHSC with laboratory generated routine sequence data. With MTexpert™, scientists can edit the mtDNA sequence traces, review the assembly (sequence alignments), and evaluate each polymorphism with dynamic links from the traces' base calls through the assembly to the type entries. The prototype MTexpert™ software served as the basic platform from which UNTHSC made recommendations for improvement based on internal laboratory testing, and evaluated each new version for bugs. In summary, with the three software programs linked contiguously, mtDNA sequence data can be evaluated quicker, with less error, to generate a complete haplotype report.

Expert systems have helped reduce the sample backlog for STR (short tandem repeat) testing of single-source nuclear DNA samples in the nation's forensic laboratories. Expert systems improve the routine interpretation practice that experts carry out by reviewing data in an automated fashion. In forensic DNA analysis, expert systems permit scientists to obtain consistent, accurate results more quickly with less mental and physical stress, that allows these scientists to review and upload many DNA profiles into the national database more efficiently.

The NIJ Expert Systems Testbed (NEST) Project evaluated the ability of three commercially available expert systems, designed specifically for forensic DNA laboratories, to rapidly and accurately review convicted offender single-source nuclear DNA profiles for eventual upload into the national DNA database. NEST Project researchers — including the Principle Investigator (PI) for this proposal — found that the three programs effectively helped reduce the backlog and ensure timely submission of data into the national database. The outcome of loading substantially more profiles into the national database is that more investigative leads could be developed and thus ultimately more crimes solved (Roby and Jones, 2005).

Currently, most forensic laboratories use Sequencher™ software (Gene Codes Corporation, Ann Arbor, MI) to analyze mtDNA sequence data. Sequencher is a versatile general purpose research tool. Sequencher software has added specific functionality to the forensic version of this tool that facilitates comparing a sequence of an unknown sample to that of a reference sequence (known as the revised Cambridge Reference Sequence, or rCRS) in order to manually create a type description of the differences. Using this software, experienced mtDNA forensic scientists require several hours to assemble, trim, assess, align, and export the results for each sample. Much of this requires manual data review analysis that is routine and repetitive.

In fact, Sequencher software inadvertently contributes to some of the workload in mtDNA data analysis. Most of the errors in the current manual review process result from Sequencher software's alignment programming between the rCRS and the sample sequence; it does not address the nomenclature rules governing the way differences between the sequences are described and reported as a haplotype. Recently, scientists at the FBI DNA Analysis Unit II worked with MitoTech to revise the standard nomenclature rules that are used to describe the difference between a mtDNA sample sequence and the rCRS reference sequence and make the rules more compliant with historical practices (Budowle *et al.,* 2010). Dr. Budowle was the lead author on this paper and has extensive experience in nomenclature reported. MitoTech implemented these revised rules, called the MitoTyper™ Rules, in a software system that now guarantees absolute stability and consistency in the nomenclature and haplotype reported. It is likely that SWGDAM will

recommend that all future mtDNA typing results in CODIS be compliant with the MitoTyper™ Rules. Existing mtDNA sequences or other human DNA sequence data can be evaluated and re-characterized with the MitoTyper™ Rules software to meet the required standards. These rules also are compliant with phylogenetic nomenclature. Thus, a system exists that provides consistent nomenclature rules for use across the forensic science community. The value of consistent nomenclature is that when a mtDNA haplotype is searched against a reference database, the proper number of concordant types are identified and no underestimation of the rarity of a mtDNA haplotype will occur.

MitoTech has also developed a prototype sequence assembly and editing software package, MTexpert™ that is integrated with the MitoTyper™ Rules. This software replaces Sequencher software with a tool that is specifically designed for forensic mtDNA data analysis. MTexpert™ is used to manually assemble and edit the sequence data and as an interface to the new MitoTyper™ Rules. Although an improvement, even with this software, completing the data analysis and type generation still requires a great deal of manual editing and review of the data by highly-trained analysts. These manual steps include routine trace trimming to eliminate lower quality regions, review of the coverage over the sequenced region, review of each possible heteroplasmy position in the sample sequence, and manual comparison of the haplotype to the results of an independent analysis.

In the study, UNTHSC and MitoTech added expert system decision automation and additional workflow automation to MTexpert™ that eliminates the need for analysts to manually perform the routine repetitive steps in data analysis, type generation, and validation. Automating the routine and repetitive steps in mtDNA analysis substantially improves the efficiency of the laboratory. An additional objective achieved was to reduce potential variability in data analysis decision criteria and concomitant results within and among forensic laboratories.

The UNTHSC worked closely with MitoTech to define the user requirements - identifying the routine circumstances and procedures in the current process and defining the expert system rules and procedures that can recognize the specific sample sequencing phenomena and carry out the appropriate interpretation procedure. MitoTech applied these rules into prototype versions of the MTexpert™ Software that UNTHSC scientists then tested so

as to validate the rules programming with independent data. With guidance and feedback from UNTHSC, modifications were then made to the software.

eFAST™ Software v2.0 introduces an approach to sequence data quality assessment that is entirely novel. The expert system rules incorporated into eFAST™ Software v2.0 have been optimized and evaluated for performance and efficiency improvement. eFAST™ Software v2.0 provides advanced quality assessments for each sequence trace, providing better guidance to the analyst for troubleshooting or determining the need for retesting samples.

The second program in the series, STATIS Software, is a batch management program designed to be used in conjunction with eFAST™ Software and MTexpert™ Software. This integrated expert system design requires that each genetic analyzer instrument computer have eFAST™ Software v2.0, while the analyst's workstations should have both STATIS Software and MTexpert™ Software installed. The functionality of STATIS requires that the sequence trace files be evaluated by eFAST™ Software v2.0 prior to automatically loading into STATIS. STATIS centralizes the sequence data and information associated with a batch, eliminates numerous steps associated with data retrieval by an analyst, and decreases the time required to evaluate the status of a sample and batch.

The UNTHSC Field Testing Division (FTD) has presented several software advancements in the analysis of mtDNA for reference samples that significantly reduce labor in both the laboratory and in data analysis and reduce the overall processing time for these samples. In addition, these software programs perform a quality check of the data and apply MitoTyper™ Rules for consistent haplotype reporting. A reduction in labor and processing time will improve efficiency and increase the overall capacity of mtDNA processing by the laboratory. With increased efficiency and capacity, more reference samples can be processed and hence, identifications can be recommended earlier. Considerable savings in costs and time can be achieved by implementing these programs.

The FBI has funded several regional mtDNA laboratories to conduct mtDNA casework as an extension of its own operations. In addition, the FBI and the Jan Bashinski DNA Laboratory of the California Department of Justice conduct casework for the Missing Persons Index system. If the UNTHSC Field Testing Division could share their software

programs with all of these laboratories in a single setting or to provide the software to each of the laboratories for testing and evaluation, additional enhancements and acceptance of the expert system suite proposed here could be made. This project suggests that such acceptance by the wider community could materially improve the output quality of mtDNA sequence data and reduce missing persons case processing backlogs throughout the United States.

# *Introduction*

Expert systems have helped reduce the backlog of single-source nuclear DNA samples typing in the nation's forensic laboratories. Expert systems automate the routine interpretation practice that experts carry out when reviewing data in a manual fashion. Expert systems allow forensic scientists to obtain consistent, accurate results more quickly, with less stress, and can help these scientists review and upload higher numbers of DNA profiles to the national databases more efficiently.

The previously reported NIJ Expert Systems Testbed (NEST) Project evaluated the ability of three commercially available expert systems, designed specifically for forensic DNA laboratories, to rapidly and accurately review convicted offender single-source nuclear DNA profiles for eventual upload into the national DNA database. The outcome of loading substantially more profiles into the national database is the development of more investigative leads and thus ultimately solving more crimes (Roby and Jones, 2005). With an expert system for mtDNA sequence analysis of high quality samples such as family reference samples, the goal is very simply to rapidly and accurately review the family reference mtDNA sequences or haplotypes for eventual upload into the CODIS national DNA database index for Missing Persons. By analogy, the outcome of loading substantially more mtDNA profiles into the national missing persons DNA database would be the development of more associations and ultimately making more identifications. An expert system for mtDNA results would automate sample file parsing, mtDNA data analysis, haplotype reporting, and quality control analysis.

## *The State of the Art in mtDNA Data Analysis Technology*

Currently, most forensic laboratories use Sequencher™ software (Gene Codes Corporation, Ann Arbor, MI) to analyze mtDNA sequence data. Sequencher is a versatile general purpose research tool. Sequencher software has added specific functionality to the forensic version of this tool that facilitates comparing a sequence of an unknown sample to that of a reference sequence (known as the revised Cambridge Reference Sequence, or rCRS) in order to

manually create a type description of the differences. Using this software, experienced mtDNA forensic scientists require several hours to assemble, trim, assess, align, and export the results for each sample. Much of this manual analysis is routine and repetitive.

In fact, Sequencher software inadvertently contributes to some of the workload in mtDNA data analysis. Most of the errors in the current manual review process result from Sequencher software's alignment programming between the rCRS and the sample sequence; it does not address the nomenclature rules governing the way differences between the sequences that are described as a haplotype. Scientists at the FBI DNA Analysis Unit 2 worked with MitoTech, LLC, our software collaborator, to revise the standard nomenclature rules that are used to describe the differences between a mtDNA sample sequence and the rCRS reference sequence and make the rules compliant (Budowle *et al.*, 2010). MitoTech implemented these revised rules, called the MitoTyper™ Rules, in a software system that now guarantees absolute stability and consistency in the nomenclature and haplotype reported. It is likely that SWGDAM will recommend that all future mtDNA typing results in CODIS be compliant with the MitoTyper™ Rules. Existing mtDNA sequences or other human DNA sequence data can be evaluated and re-characterized with the MitoTyper™ Rules software to meet the required standards. These rules also are compliant with phylogenetic nomenclature. Thus, a system exists that provides consistent nomenclature rules across the forensic science community. The value of a consistent nomenclature is that when an mtDNA haplotype is searched against a reference database, the proper number of concordant types are identified and no underestimation of the rarity of a mtDNA haplotype will occur.

MitoTech has also developed a prototype sequence assembly and editing software package, MTexpert™ Software, which has integrated the MitoTyper™ Rules. This software is another option to using Sequencher software with a tool that is specifically designed for forensic mtDNA data analysis. MTexpert™ Software is used to manually assemble and edit the sequence data and as an interface to the new MitoTyper™ Rules. The ultimate goal of this project was to provide guidance to the software engineers of needs of the forensic analyst to tie in the MitoTyper™ Rules to routine sequence data review. These manual steps include routine trace trimming to eliminate lower quality regions,

review of the coverage over the sequenced region, review of each possible heteroplasmy position in the sample sequence, and manual comparison of the haplotype to the results of an independent analysis.

## *Developing an Expert System for mtDNA Data Analysis*

During this study, UNTHSC and MitoTech added expert system decision automation and additional workflow automation to MTexpert™ Software that eliminates the need for analysts to perform the routine repetitive steps in data analysis, type generation, and validation. Automating the routine and repetitive steps in mtDNA analysis substantially improves the efficiency of the laboratory. An additional objective for this proposal was the reduction of potential variability in data analysis decision criteria and concomitant results within and among forensic laboratories.

The UNTHSC defined the user requirements - identifying the routine circumstances and procedures in the current casework process and defining the expert system rules and procedures that can recognize the circumstances and carry out the appropriate procedure. MitoTech has incorporated these rules into prototype versions of the MTexpert™ Software and the UNTHSC scientists have tested and validated the rules with independent data. Under guidance and with the feedback from UNTHSC, modifications were made to the software.

## *Review of Relevant Literature*

Expert systems for single source nuclear DNA profiles in forensic laboratories are well-established (Roby, 2008a). These programs increase the speed and accuracy of nuclear DNA data analysis, increasing the throughput of the laboratories. One study recorded that the incorporation of an expert system into the work-flow reduced the data analysis time from 4.5 hours per 96 well plate to less than one hour (Frappier, 2005). The New York State Police report that use of the TrueAllele System 2 expert system (Cybergenetics, Pittsburgh, PA) reduced the time required to analyze a plate of data from five hours to five minutes (Kadash *et al.*, 2004). Recommended validation procedures for data analysis of expert systems have also been presented (Roby and Christen, 2007). These studies establish a foundation and guidelines for validating the proposed expert system for mtDNA applications.

# *Software*

The focus of this project was on the enhancement of eFAST™ Software v2.0 and the development of STATIS Software in concert with the improvements of MTexpert™ Automated mtDNA Data Analysis Software (MitoTech, LLC). The use of filter metrics to quickly assess mtDNA sequence data has been incorporated into these software programs for data screening using expert system rule firing features (Roby, 2008b; Curtis *et al.*, 2010; Roby *et al.*, 2011). These three software programs are designed to communicate with one another for an integrated workflow for mtDNA data analysis and haplotype report generation. To further streamline automation and reduce entry errors, barcoding, automated sample tracking, and auto-population of sample sheets have been put into use as described in the "Development of an Integrated Workflow from Laboratory Processing to Report Generation for mtDNA Haplotype Analysis" for increased efficiency (Phillips *et al.*, 2009; Roby *et al.*, 2009; Roby *et al*., 2011).

## *eFAST™ Software*

Sequence analysis is a time-consuming process, particularly due to the large amount of data that are required to obtain a complete profile. The standards for mtDNA sequencing for forensic casework require double coverage for all bases reported in an mtDNA profile. For one sample, a minimum of four traces must be generated, evaluated for quality, and, if the traces are of acceptable quality, assembled to the rCRS. The quality screening process used for casework is monotonous, subjective, and time-consuming. eFAST™ Software v1.1 was designed to replace the repetitive and subjective process of screening sequence data with an expert system approach based on optimized filter metrics (Roby *et al*., 2010). eFAST™ (electronic Filtering and Assessment of Sequence Traces) Software provides: 1) customizable trace name pattern analysis (Figure 1); 2) objective quality assessment of controls and traces (Figure 2); 3)

automated file distribution (Figure 3); 4) sample progress summaries to facilitate laboratory workflow (Figure 3); and, 5) electronic notification of run performance via email (Figure 4).

eFAST™ Software calculates a Contiguous Read Length (CRL) and Trace Score (TS) for each trace. CRL is calculated as the number of uninterrupted bases in the trace that have a quality value (QV) of greater than 20. TS is the average QV of the bases that remain in the trace (after trimming). These metrics are used to sort traces into three categories: high quality (HQ), review (REV), and low quality (LQ). These metrics are used to evaluate the trace quality of both controls and sample traces. The user can define the sample naming convention, set the thresholds in a primer-specific manner, and can define custom primers. Other customizable features make eFAST™ Software amenable for all dRhodamine and BigDye® sequencing applications.

During a plate run, eFAST™ Software evaluates controls as soon as the data collection is complete for each run. If a control fails early in the plate, an Early Warning email is generated and sent to alert the user of the problem. If the controls do not fail, eFAST™ Software creates a summary email for the user once data collection for a plate is complete. This email informs the user of the number of traces qualified as HQ, REV, and LQ. Additionally, the email summarizes the performance of the controls.

eFAST™ Software provides a color-coded interface which can be filtered to only display traces in need of review (REV). Once the analyst has assigned the quality of the REV traces manually, all of the sample trace files are automatically sorted into pre-defined directories. The traces categorized as LQ are archived in a directory titled *Low Quality*, and the HQ traces for each sample are grouped for analysis. After distribution, eFAST™ Software creates a Sample Report which indicates the status of all traces for each sample. The Sample Report can be sorted and exported; it is designed to facilitate subsequent laboratory processing to further increase efficiency.

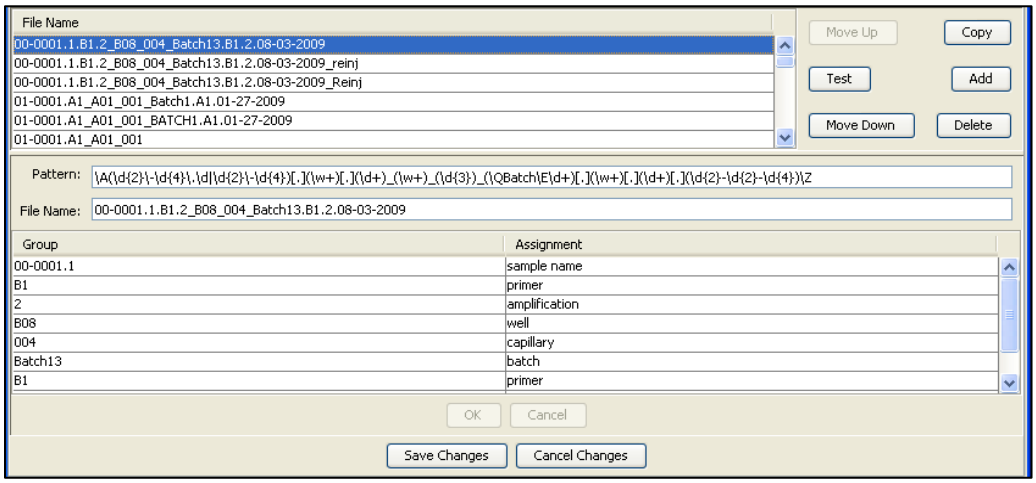
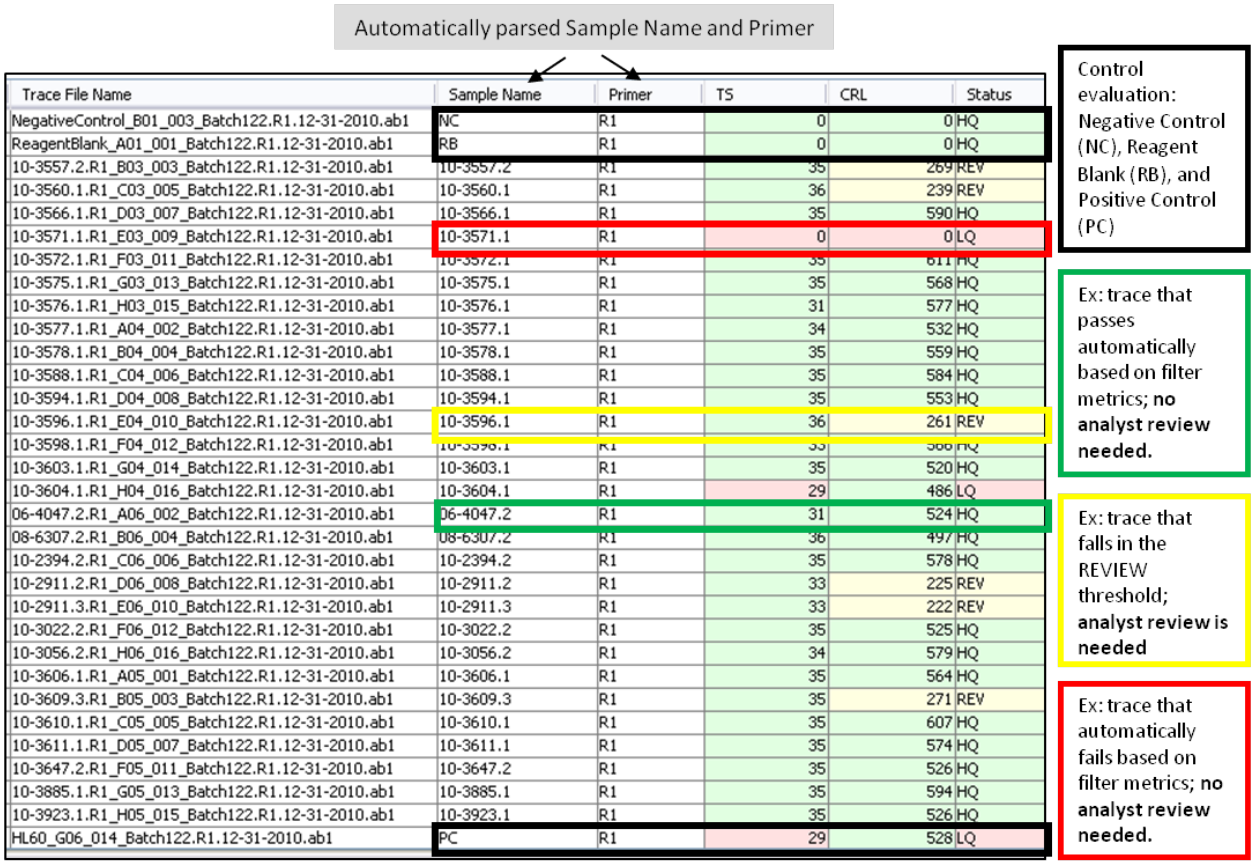
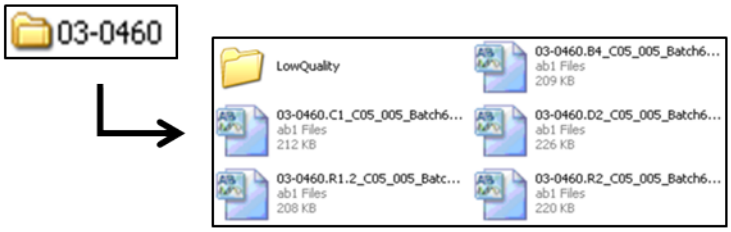



Figure 1. Pattern generator for trace names. The name pattern is used to define each handle of the trace name in order to automate sample grouping, control assessment, and primer-specific quality assessments.




| Trace File Name | Sample Name | Primer | TS | CRL | Status |
|---|---|---|---|---|---|
| NegativeControl_B01_003_Batch122.R1.12-31-2010.ab1 | NC | R1 | 0 | 0 | HQ |
| ReagentBlank_A01_001_Batch122.R1.12-31-2010.ab1 | RB | R1 | 0 | 0 | HQ |
| 10-3557.2.R1_B03_003_Batch122.R1.12-31-2010.ab1 | 10-3557.2 | R1 | 35 | 269 | REV |
| 10-3560.1.R1_C03_005_Batch122.R1.12-31-2010.ab1 | 10-3560.1 | R1 | 36 | 239 | REV |
| 10-3566.1.R1_D03_007_Batch122.R1.12-31-2010.ab1 | 10-3566.1 | R1 | 35 | 590 | HQ |
| 10-3571.1.R1_E03_009_Batch122.R1.12-31-2010.ab1 | 10-3571.1 | R1 | 0 | 0 | LQ |
| 10-3572.1.R1_F03_011_Batch122.R1.12-31-2010.ab1 | 10-3572.1 | R1 | 35 | 611 | HQ |
| 10-3575.1.R1_G03_013_Batch122.R1.12-31-2010.ab1 | 10-3575.1 | R1 | 35 | 568 | HQ |
| 10-3576.1.R1_H03_015_Batch122.R1.12-31-2010.ab1 | 10-3576.1 | R1 | 31 | 577 | HQ |
| 10-3577.1.R1_A04_002_Batch122.R1.12-31-2010.ab1 | 10-3577.1 | R1 | 34 | 532 | HQ |
| 10-3578.1.R1_B04_004_Batch122.R1.12-31-2010.ab1 | 10-3578.1 | R1 | 35 | 559 | HQ |
| 10-3588.1.R1_C04_006_Batch122.R1.12-31-2010.ab1 | 10-3588.1 | R1 | 35 | 584 | HQ |
| 10-3594.1.R1_D04_008_Batch122.R1.12-31-2010.ab1 | 10-3594.1 | R1 | 35 | 553 | HQ |
| 10-3596.1.R1_E04_010_Batch122.R1.12-31-2010.ab1 | 10-3596.1 | R1 | 36 | 261 | REV |
| 10-3598.1.R1_F04_012_Batch122.R1.12-31-2010.ab1 | 10-3598.1 | R1 | 35 | 588 | HQ |
| 10-3603.1.R1_G04_014_Batch122.R1.12-31-2010.ab1 | 10-3603.1 | R1 | 35 | 520 | HQ |
| 10-3604.1.R1_H04_016_Batch122.R1.12-31-2010.ab1 | 10-3604.1 | R1 | 29 | 486 | LQ |
| 06-4047.2.R1_A06_002_Batch122.R1.12-31-2010.ab1 | 06-4047.2 | R1 | 31 | 524 | HQ |
| 08-6307.2.R1_B06_004_Batch122.R1.12-31-2010.ab1 | 08-6307.2 | R1 | 36 | 497 | HQ |
| 10-2394.2.R1_C06_006_Batch122.R1.12-31-2010.ab1 | 10-2394.2 | R1 | 35 | 578 | HQ |
| 10-2911.2.R1_D06_008_Batch122.R1.12-31-2010.ab1 | 10-2911.2 | R1 | 33 | 225 | REV |
| 10-2911.3.R1_E06_010_Batch122.R1.12-31-2010.ab1 | 10-2911.3 | R1 | 33 | 222 | REV |
| 10-3022.2.R1_F06_012_Batch122.R1.12-31-2010.ab1 | 10-3022.2 | R1 | 35 | 525 | HQ |
| 10-3056.2.R1_H06_016_Batch122.R1.12-31-2010.ab1 | 10-3056.2 | R1 | 34 | 579 | HQ |
| 10-3606.1.R1_A05_001_Batch122.R1.12-31-2010.ab1 | 10-3606.1 | R1 | 35 | 564 | HQ |
| 10-3609.3.R1_B05_003_Batch122.R1.12-31-2010.ab1 | 10-3609.3 | R1 | 35 | 271 | REV |
| 10-3610.1.R1_C05_005_Batch122.R1.12-31-2010.ab1 | 10-3610.1 | R1 | 35 | 607 | HQ |
| 10-3611.1.R1_D05_007_Batch122.R1.12-31-2010.ab1 | 10-3611.1 | R1 | 35 | 574 | HQ |
| 10-3647.2.R1_F05_011_Batch122.R1.12-31-2010.ab1 | 10-3647.2 | R1 | 35 | 526 | HQ |
| 10-3885.1.R1_G05_013_Batch122.R1.12-31-2010.ab1 | 10-3885.1 | R1 | 35 | 594 | HQ |
| 10-3923.1.R1_H05_015_Batch122.R1.12-31-2010.ab1 | 10-3923.1 | R1 | 35 | 526 | HQ |
| HL60_G06_014_Batch122.R1.12-31-2010.ab1 | PC | R1 | 29 | 528 | LQ |

Figure 2. Automated and objective quality assessment of controls and traces. Controls are evaluated based on the expected results. For example, if the positive control has a low trace score (TS) and a low contiguous read length (CRL), based on the laboratory defined thresholds, the positive control will either receive a status of REV (review in yellow) or LQ (low quality in red, which fails). On the other hand, if a negative control or reagent blank has a low TS and low CRL, based on the laboratory defined thresholds, the controls will either receive a REV or HQ (high quality in green, which passes) status because the analyst expects these controls to not yield interpretable sequence. Sample traces are also assigned a status of LQ, REV, or HQ based on the laboratory defined thresholds for these filter metrics.




eFAST Sample Report

| Sample | Project? | B1 | B4 | C1 | D2 |
|---|---|---|---|---|---|
| 03-0431 | | | B4 | C1 | D2 |
| 03-0432 | | B1 | B4 | C1 | D2 |
| 03-0433 | | | | C1 | D2 |
| 03-0434 | | B1 | B4 | C1 | D2 |
| 03-0435 | | | | C1 | D2 |
| 03-0436 | | B1 | B4 | C1 | D2 |
| 03-0437 | | B1 | B4 | C1 | D2 |
| 03-0438 | | B1 | B4 | C1 | D2 |
| 03-0439 | | | B4 | C1 | D2 |
| 03-0440 | | B1 | B4 | C1 | D2 |
| 03-0441 | | B1 | | | |
| 03-0442 | | B1 | B4 | | D2 |
| 03-0443 | | | | C1 | |
| 03-0444 | | B1 | B4 | C1 | |

Figure 3. Automated file distribution and Sample Report. eFAST™ Software automatically creates a directory for every sample processed, based on the defined trace naming pattern. Within this directory, traces classified as HQ are grouped and traces that were classified as LQ are archived in a sub-directory. The sample directories are summarized in the Sample Report to facilitate subsequent sample processing. The red cells indicate that the sample does not have a high quality trace for this primer, while the green cell indicates that the primer sequenced successfully.

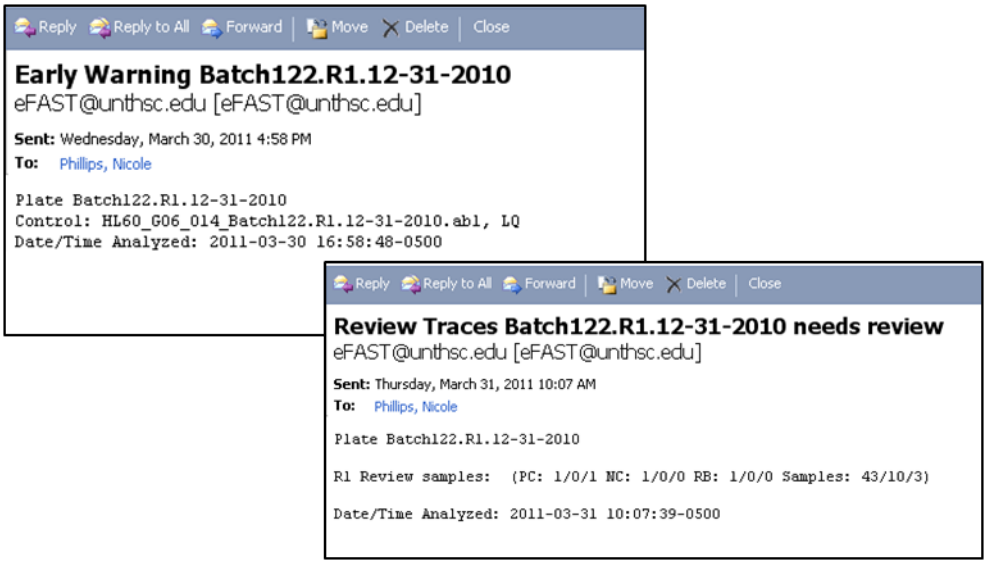



Figure 4. eFAST™ Software sends automated email notifications which include early warnings for a control failure as well as plate completion summaries. These plate completion summaries provide an overview of the run's performance.

Studies performed with eFAST Software v1.1 demonstrated that traces requiring an analyst's review decreases the efficiency of eFAST™ Software for sorting such traces. For these reasons, additional rules have been programmed into eFAST™ Software v.2.0 in order to reduce the error rate of trace evaluation and sorting and to reduce the time to review the data.

# *eFAST™ Software v2.0*

Version 2.0 of eFAST™ Software builds upon all of the functionality of eFAST™ Software v1.1. Most notably, seven new expert system rules (in addition to TS, Trace Score, and CRL, Contiguous Read Length) are featured in eFAST™ Software v2.0 in order to further enhance the efficiency and discriminatory power of the sorting algorithms (Table 4). They include High Baseline (HB), High Signal (HS), Low Signal (LS), Partial Read (PR), Mixture (Mix), Homopolymeric Stretch (HPS), and Length Heteroplasmy (LH). These additional rules decrease the error rate seen in eFAST™ Software v1.1 and provide valuable insight into trace nuances. The Trace Summary table has been expanded to incorporate the rules, where symbolic flags are used to indicate the status of each rule (Figure 5). A green check ( ✓ ) indicates that the trace passes the rule and does not exhibit the rule characteristic. A yellow exclamation point ( ❗ ) indicates that the trace may exhibit the characteristic being tested. A red X ( ✖ ) indicates that the trace does exhibit the rule characteristic. Certain rule conditions will not be detectable if another rule has previously fired. Such instances are indicated by "NC", not checked.

| Rule Name | Type of Rule | Description |
|---|---|---|
| **High Baseline** | Enforced | Nested minor peaks in the primary signal (user defined) |
| **High Signal** | Informative | Signal intensity saturates the CCD camera; potential pull up peaks |
| **Low Signal** | Enforced (if defined) | Average signal intensity below a threshold (user defined) |
| **Partial Read** | Informative | Peaks suddenly decrease in intensity and change in morphology; potentially fixed by reinjection |
| **Mixture** | Informative | An observed number of high quality mixed bases observed in the trimmed trace (user defined) |
| **Homopolymeric Stretch** | Enforced | A series of homogenous bases followed by an increase in baseline noise; creates a CRL exception |
| **Length Heteroplasmy** | Enforced | A heteroplasmic insertion/deletion causing out-of-phase minor species peaks; creates a CRL exception |

Table 1. Description of additional rules and functionality. Several rules have user defined thresholds. These include high baseline (HB), low signal (LS), and mixture (Mix). The rules are classified as either *informative*, which alerts the analyst of a condition and may act as a guide for further action to be performed by the analyst, or *enforced,* which affects the overall status of the trace.

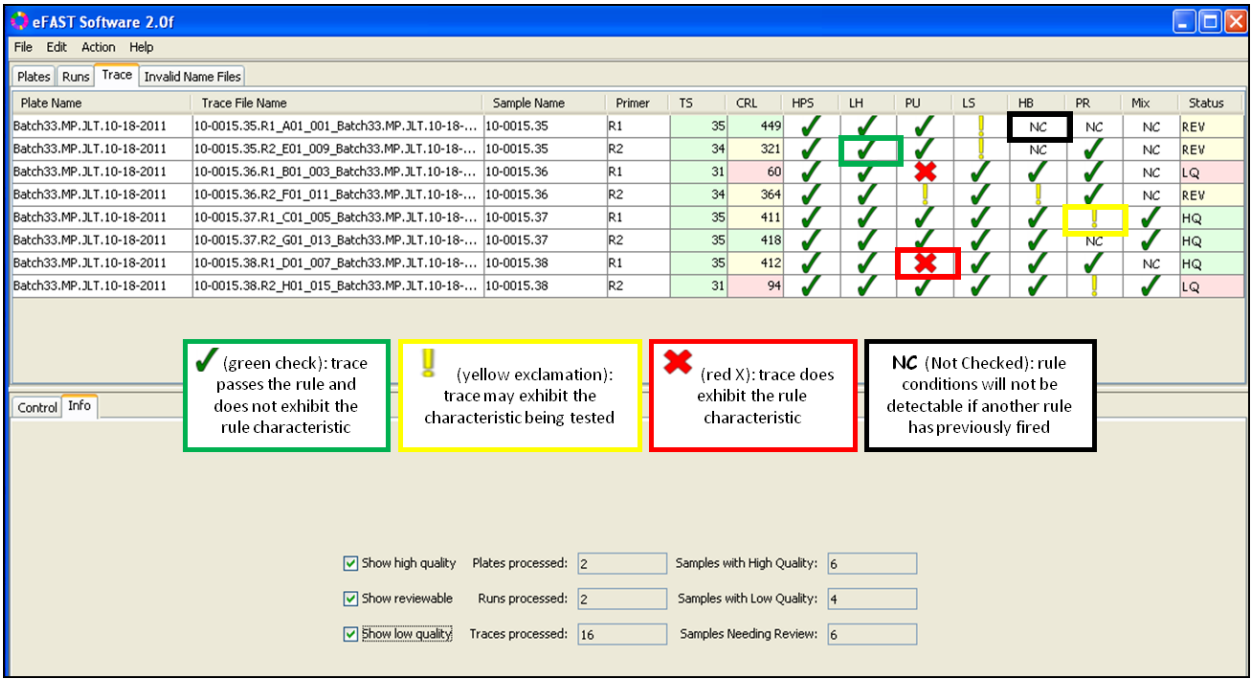



Figure 5. The new Trace tab interface in eFAST™ Software v2.0. The quality metrics and rule firings for each trace is displayed in the table with the overall status displayed in the far right column as low quality (LQ in red), review (REV in yellow), or high quality (HQ in green).

The rules are either *informative*, in which it simply alerts the analyst of a condition and does not pass/fail a trace, or *enforced,* in which the rule status affects the overall status of the trace and is based on user-defined thresholds. *Informative* rules guide the analyst in further action; for example, the PR rule indicates that an electrophoretic issue occurred, causing a sudden loss of signal (Figure 9). A PR (Partial Read) trace usually fails the CRL (Contiguous Read Length) and/or TS (Trace Score) rule(s), but since this anomaly is easily remedied by reinjection, it is very beneficial for the analyst to be informed of the condition. In contrast to the PR rule, the HPS (Homopolymeric Stretch) rule is an example of an *enforced* rule; it indicates that a homopolymeric stretch has been detected in the trace. When this occurs, as discussed previously with regard to R1 traces, the expected CRL is truncated. With HPS or LH (Length Heteroplasmy) detected, the CRL rule firing will be overridden and affects the overall status of the trace.

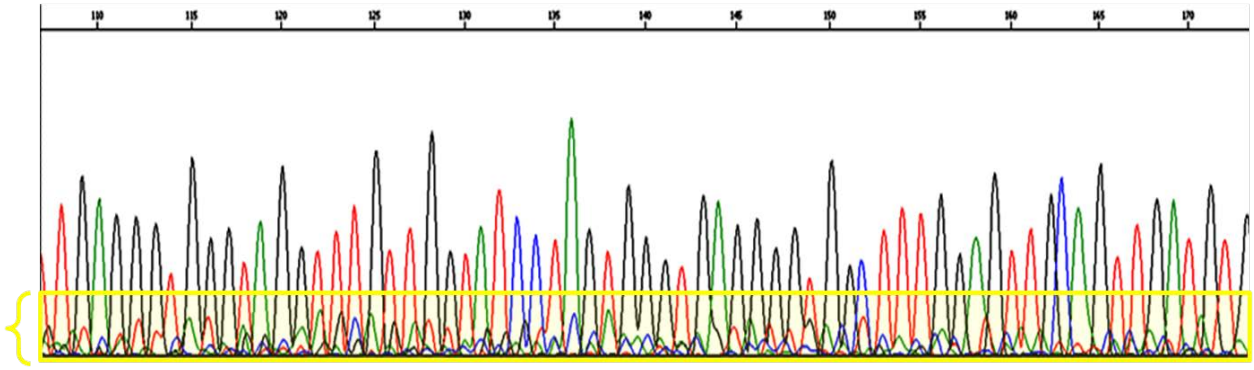


Figure 6. The HB rule firing. The High Baseline rule fires to indicate the level of *peak noise* within the called bases.

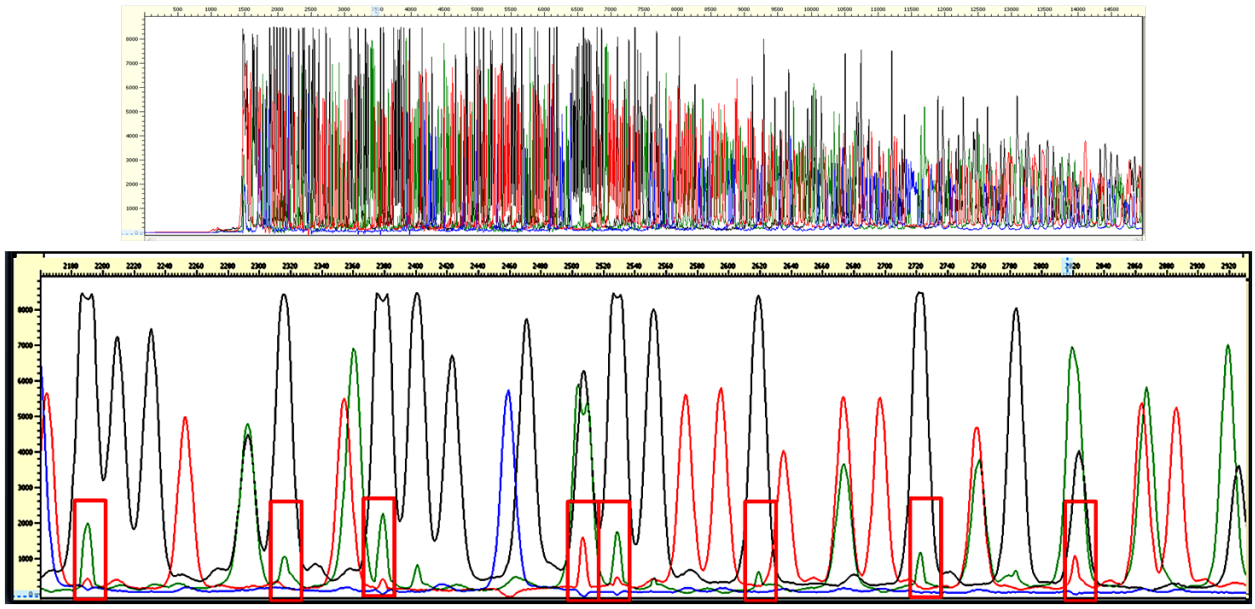


Figure 7. The HS rule firing. The High Signal rule fires when CCD camera saturation is suspected, which results in pull up peaks as indicated in the red boxes.

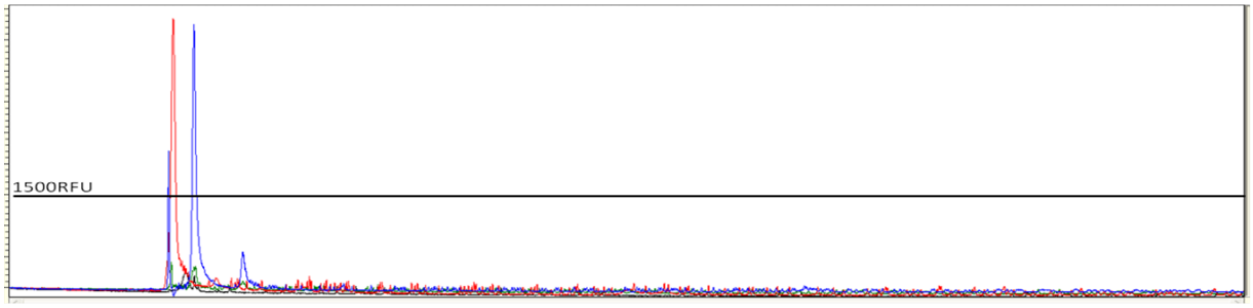



Figure 8. The LS rule firing. The Low Signal rule fires when the relative fluorescent units (RFU) are below a specified threshold which indicates that the signal intensity of the data may be too low for accurate interpretation.

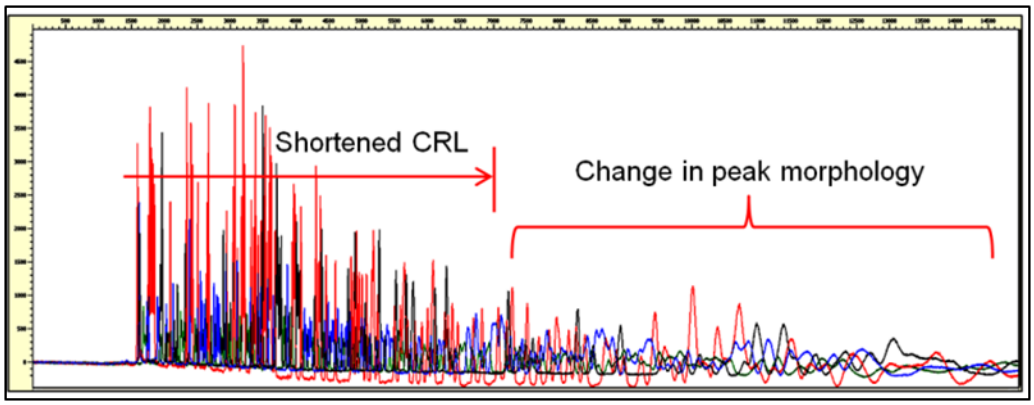



Figure 9. The PR (Partial Read) rule firing. This is an example of an informative rule. This trace would fail due to shortened CRL (Contiguous Read Length) and/or poor TS (Trace Score). However, since this condition is remedied by reinjection, the rule firing informs the analyst to consider reinjecting the trace.

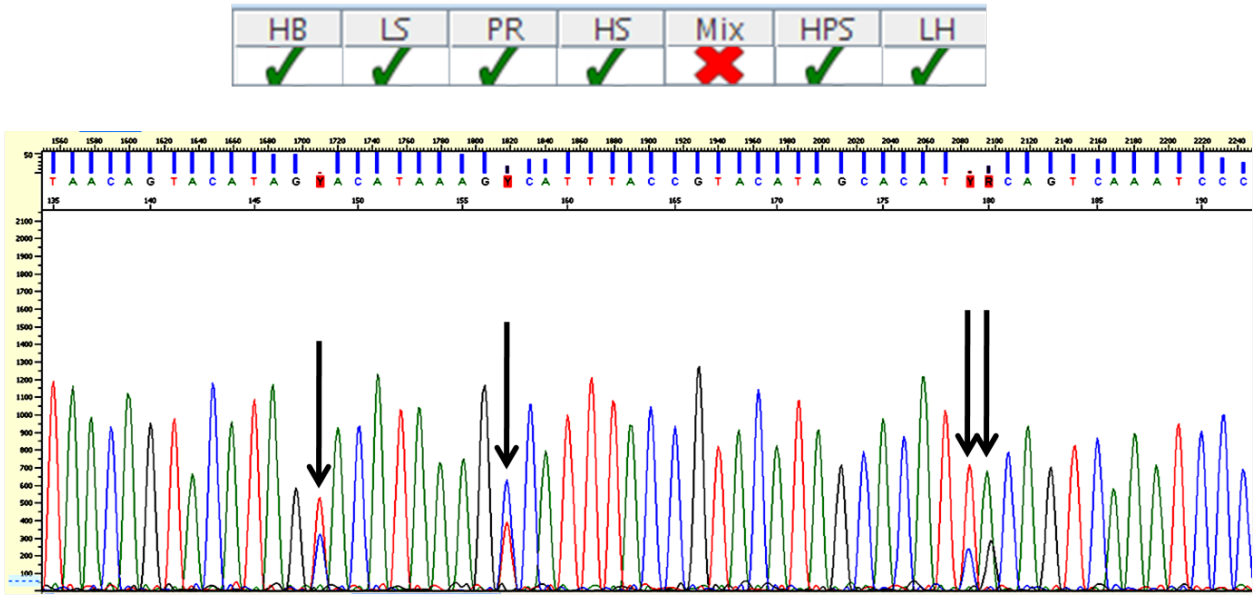



Figure 10. The Mixture rule firing. The Mix rule fires if more than a specified number of high quality mixed bases (indicated by black arrows) are observed in a trace; this rule is only checked if no previous rules fire.

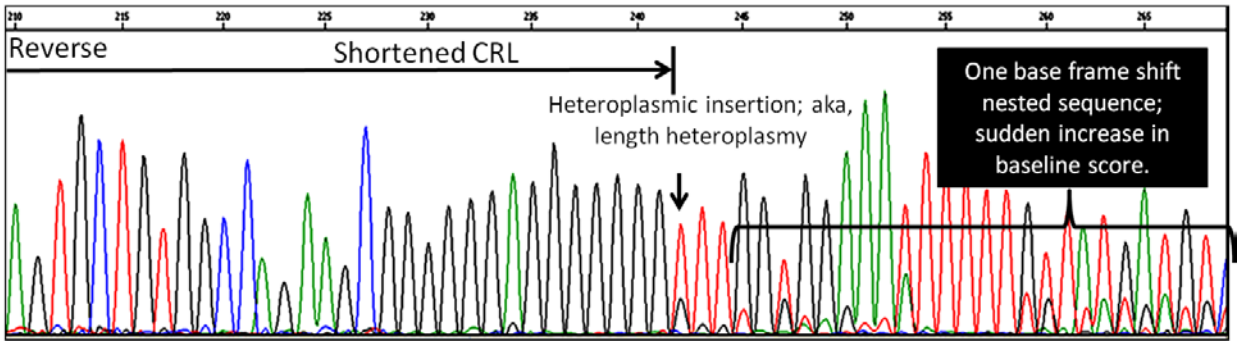



Figure 11. The LH (Length Heteroplasmy) rule firing. The Length Heteroplasmy rule fires if variable number of cytosines at base position 303 in HV2 is detected, resulting in frame shifted sequence nested down stream. This rule firing alone does not fail the trace, but overrides the CRL (Contiguous Read Length) requirement in order to pass these traces specifically.

eFAST™ Software automates control evaluation, trace quality assessment, and file management. The software program alerts the analyst of failed controls and overall plate quality via email. The quality metrics evaluated within the software include: trace score (TS), contiguous read length (CRL), homopolymeric stretch (HPS), length heteroplasmy (LH), pull-up (PU), low signal (LS), high baseline (HBL), partial read (PR), mixture (Mix). Based on these quality metrics, the software program assigns a status of low quality, review, or high quality to each sequence trace. The quality metrics follow a rule firing hierarchy. The software always checks for HPS, LH, PU, LS, and always assesses the TS and CRL. The HB rule is not checked when the LS rule fires. The PR rule is not checked when the HPS or LH rule fires or when the CRL threshold is met. The Mix rule is not checked if one or more of the following rules fire: HB, PU, LH, or HPS.

Using eFAST™ Software v1.1 significantly decreases the time required to assess sequence trace quality. Although there is an error rate associated with the trace sorting algorithm used in eFAST™ Software v1.1, this approach has great potential to increase automation and objectivity in the process of screening traces for quality.

While version 1.1 demonstrates significant efficiency improvement, there were opportunities for further development. eFAST™ Software v2.0 introduces an approach to sequence data quality assessment that is entirely novel. The expert system rules incorporated into eFAST™ Software v2.0 have been optimized and evaluated for performance and efficiency improvement. eFAST Software v2.0 provides advanced quality assessments for each sequence trace that provide the analyst with better guidance for troubleshooting or retesting samples. See Appendix A for the eFAST™ Software User Guide.

## *STATIS Software*

STATIS Software is a batch management program designed to be used in conjunction with eFAST™ Software and MTexpert™ Software. This integrated expert system design requires that each genetic analyzer instrument computer have eFAST™ Software v2.0, while the analyst's workstations should have both STATIS Software and MTexpert™ Software installed. The functionality of STATIS requires that the sequence trace files be evaluated by eFAST™ Software

v2.0 prior to automatically loading into STATIS. STATIS centralizes sequence data and information associated with a batch, eliminates numerous steps associated with data retrieval by an analyst, and decreases the time required to evaluate the status of a sample and batch.

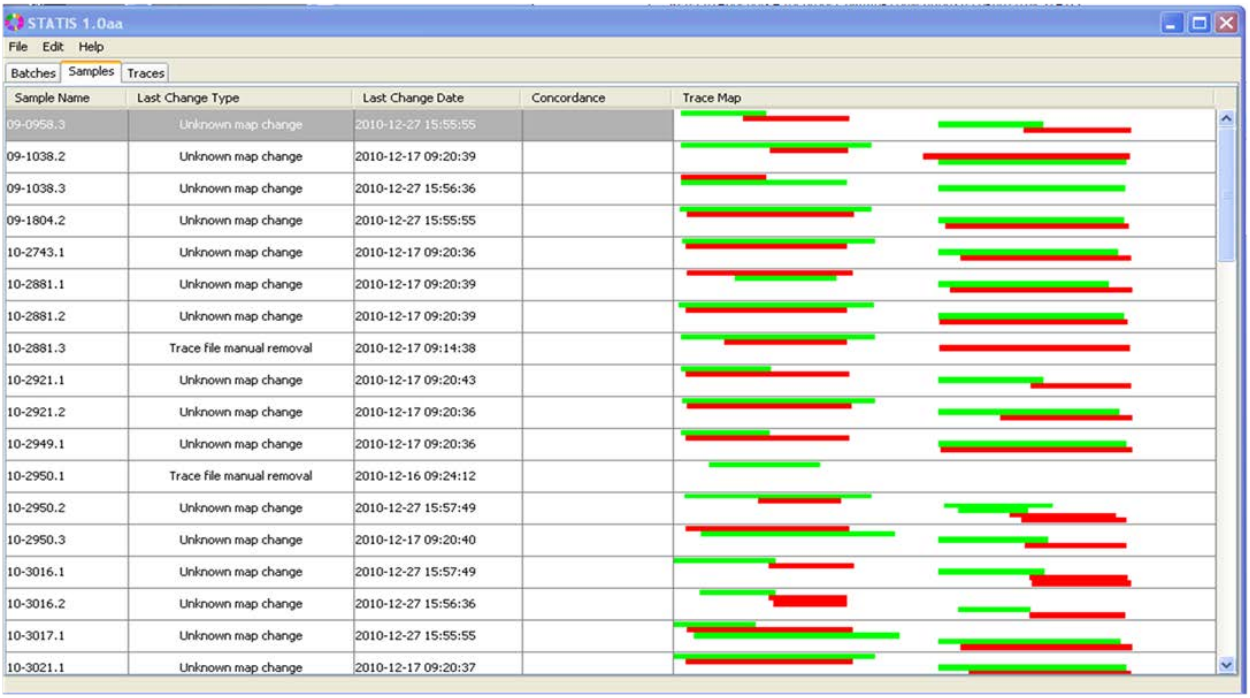



Figure 12. The samples tab displays the samples associated with the selected batch. This tab shows the most recent changes to the MTexpert™ sequence project as well as the Trace Map. The Last Change Type column displays the last edit made to the MTexpert™ project; the user can right click and select Edit to launch the MTexpert™ project. The Trace Map column provides a general overview of the sequence coverage of each sample. Forward primer coverage is illustrated by the green bar and reverse primer coverage is illustrated by the red bar.

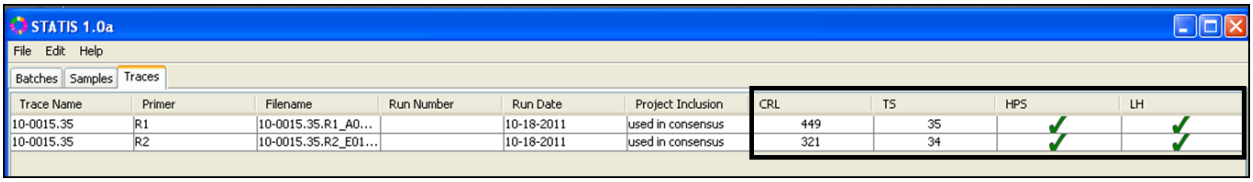



Figure 13. This menu displays the sequence traces associated with the selected sample. It also contains quality information such as the CRL, TS, HPS and LH information assigned by eFAST™ Software v2.0. The sequence trace can be launched into a viewer by double clicking the trace name. The order of the columns can be arranged by dragging according to the user's preference.

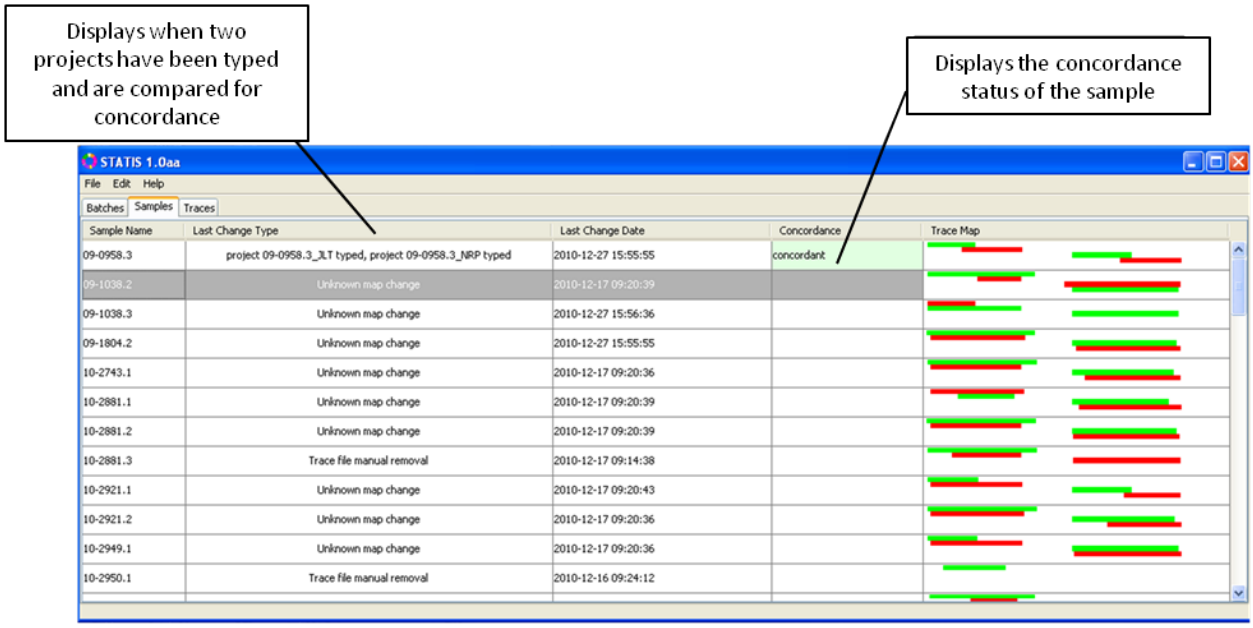



Figure 14. After two analysts have completed their reads, STATIS software loads the most recent changes and checks for agreement between the two projects. The status of agreement is shown in the concordance column. Because the two projects are concordant, the cell is highlighted green for a quick assessment of the concordance status.

## *MTexpert™ Software*

MTexpert™ Software is an automated mtDNA data analysis program which automatically trims and assembles traces, validates controls, highlights base positions that do not meet laboratory-defined criteria. The software program utilizes MitoTyper™ Rules to automatically produce historically concordant base calls and eliminates ambiguity

encountered when using other rule sets for mtDNA sequence data analysis (Den Hartog et al., 2009; Budowle et al., 2010). MItoTyper ™ is available as a standalone software program; however, MTexpert™ has the MitoTyper™ Rules integrated and uses them to automate haplotype or Signature designation. MTexpert™ and the MitoTyper™ Rules were originally developed by MitoTech, LLC under an agreement with the United States Federal Bureau of Investigation DNA Analysis Unit II. The UNTHSC has collaborated with MitoTech, LLC to further enhance this analysis software. Documentation of the issues and suggestions made, as well as the resulting enhancements, are included in Appendix B.

The MTexpert™ user interface is highly graphic (Figure 15). Most of the analysis is performed in the Assembly Map Window, which displays the traces from a birds-eye view; when reviewing individual flags, the analyst can view the traces that cover any particular region by clicking any base in the Assembly window. The individual trace panes launch and display to the left of the Assembly Map Window (Figure 15).

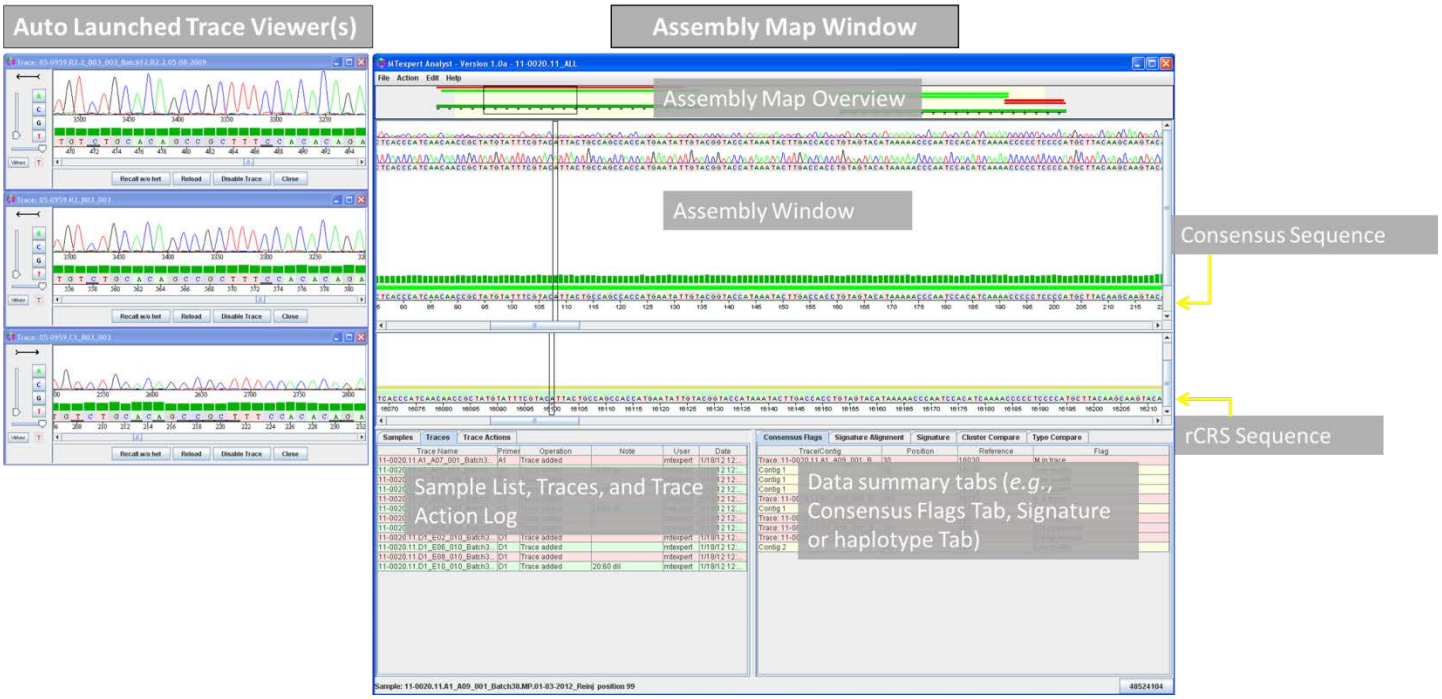



Figure 15. The MTexpert™ software graphic user interface (GUI) displays: 1) trace viewers; 2) assembly window; 3) assembly map; 4) sample list, trace list and trace action log; and 5) data summary tabs including Consensus Flags tab and Signature information. The Consensus Flags tab is automatically populated based on laboratory defined criteria for base quality, consensus quality, base disagreement, and heteroplasmy detection. This list enumerates the flags, or rule

firings, that require manual review. The Signature tab provides the mtDNA haplotype for the sample based on automated calling using the integrated MitoTyper™ Rules.

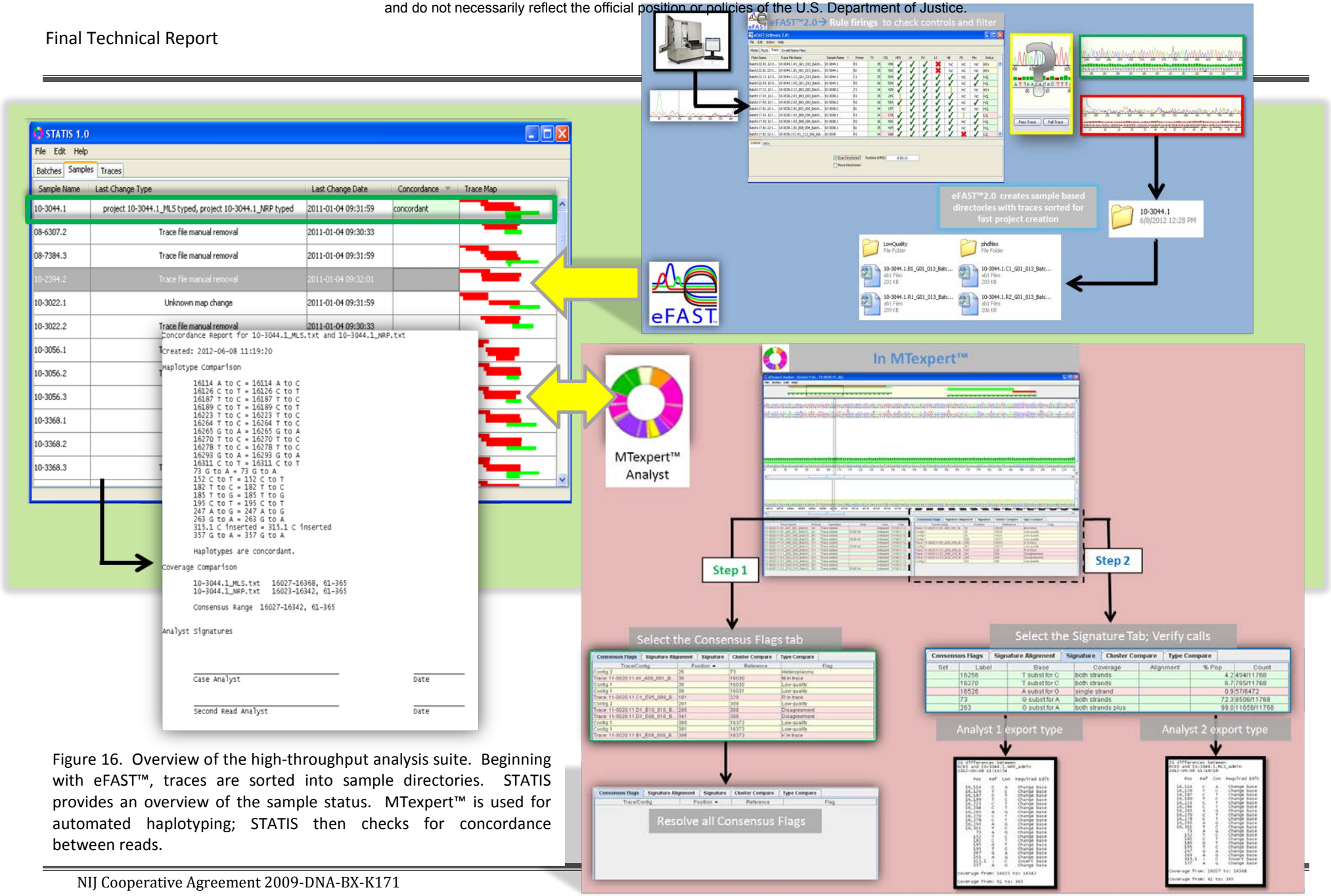



Figure 16. Overview of the high-throughput analysis suite. Beginning with eFAST™, traces are sorted into sample directories. STATIS provides an overview of the sample status. MTexpert™ is used for automated haplotyping; STATIS then checks for concordance between reads.

As a consequence of the extensive evaluation of the MTexpert™ Software performed as part of this project, additional issues and suggestions not implemented during this project were also provided to MitoTech programmers for further improvement of MTexpert™ Software. The results of these evaluations are discussed in more detail below.

**Note:** See Appendix B for a representative list of the enhancements.

## *Evaluation of MitoTyper™ Rules for Automated Type Calling*

One thousand and four samples were processed for mtDNA by UNTHSC to construct the Chilean population database. All samples were analyzed using the manual calling method established by the UNTHSC Field Testing Divisiony standard operating procedures. In order to evaluate the accuracy of the MitoTyper™ Rules, the 1004 samples were typed using MItoTyper™ Software and checked for concordance between the manual and automated calling methods. Nine hundred and eighty-five (985/1004 or 98.11%) of the samples were concordant between the calling methods; nineteen (19/1004 or 1.89%) of the samples did not produce concordant types. The discordant instances were further investigated and are summarized in Table 2 below.

| Reason for discordance | Number of instances | Explanation |
|---|---|---|
| **Human error** | 5 | MitoTyper called the correct haplotype in all of these instances. |
| **Ambiguity in call** | 12 | These are "challenging" samples, where there are multiple ways to make the call, and there is no clear correct haplotype. In 6 of the cases, the manual calls are concordant with AFDIL' s calling guide. In 4 instances, the MitoTyper call is concordant with AFDIL's calling guide. In two instances, the manual calls are consistent with AFDIL's calling guide, but the MitoTyper call has fewer differences. |
| **Discrepancy** | 2 | These discrepancies were due to an error in the MitoTyper program; the incidence of this error was very rare, and did not affect the calls in any other cases. MitoTech resolved the issue as soon as it was reported. |

Table 2. Summary of the Concordance Study between Manual Haplotype Calls and automated haplotype calls using MitoTyper™ Rules

## *Evaluation of Challenging Samples*

Twenty-nine challenging samples were evaluated for concordance between MTexpert™ Software using MitoTyper™ Rules (performed by a UNTHSC Field Testing Division or FTD mtDNA analyst) and Sequencher™ (performed by a UNT Center for Human Identification or CHI mtDNA casework analyst). The original sequence files were imported into MTexpert™ and analyzed following MitoTyper™ rules. The variance reports generated were then compared to the results reported by the CHI analyst using Sequencher™.

Twenty-five mtDNA haplotypes were concordant. In Samples 1 through 23, the differences reported in HV1 and HV2 were the same. For Sample 24, the FTD analyst obtained a length heteroplasmy in HV1 that was not reported by the CHI analyst so a direct comparison was not possible; however, the HV2 haplotype was concordant. Both FTD and CHI analysts reported a mixture for Sample 25; therefore, the results were considered concordant.

Only four samples were considered discordant. Three samples were reported correctly by the FTD analyst using MitoTyper™ Rules, but called incorrectly by the CHI analyst. The FTD analyst reported a length heteroplasmy for Sample 26 that was not reported by the CHI analyst. An insertion at base position 58 was called by the FTD analyst; however, the CHI analyst reported an insertion at base position 56, resulting in two difference between the two analysts. For sample 28, an insertion at base position 57.2 and a deletion at base position 66 were reported by the FTD analyst. For the same sample, the CHI analyst called an insertion at base position 60.1 and a deletion at base position 71, similarly resulting in two differences between the two analysts. When evaluating these latter differences, it was observed that the sample sequence data was not properly aligned by MTexpert™ Software due to a homopolymeric cytosine stretch in HV1. The programmer was notified and changes were introduced to the software program. All samples were re-evaluated with a revised version of MTexpert, and the latter discrepancy due to incorrect alignment was resolved.

**Note:** See Appendix C for a draft of the MTexpert™ User Manual.

## *Evaluation of MTexpert™ Software for Expert System Analysis of Reference Samples*

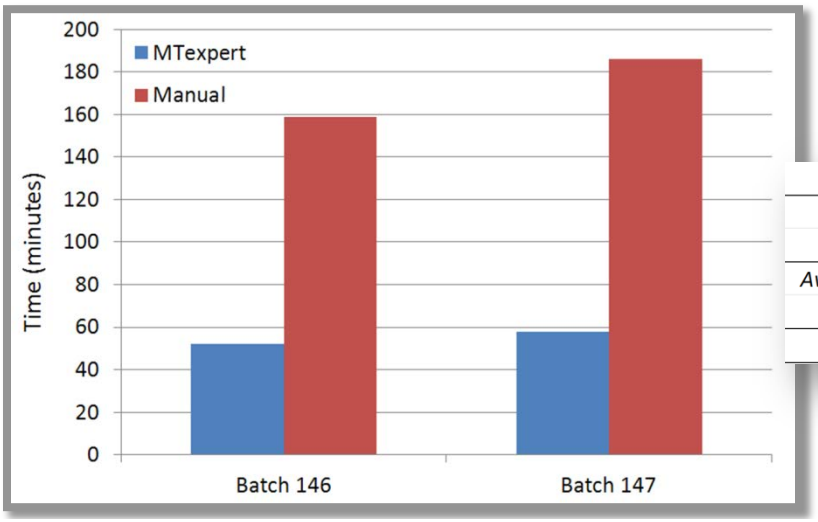
Two batches of family reference samples (FRS), containing a total of 47 samples, were randomly chosen to evaluate the use of MTexpert™ Software as an expert system for mtDNA analysis. Manual analysis using Sequencher™ Software was conducted by a CHI analyst and expert system analysis using MTexpert™ Software was conducted by a FTD analyst. Data were analyzed for (1) time-savings, and (2) accuracy of calls.

The CHI analyst is responsible for verifying every base position in the reported haplotype range. This is accomplished in Sequencher™ by manually inspecting every peak in the sequence trace. At a minimum, 610 bases must be verified for each sample, and every base must be validated in more than one trace; therefore, a minimum of 1220 peaks must be evaluated by the analyst manually for every mtDNA haplotype reported. Sequencher™ does have automated base calling and heteroplasmy detection. These tools greatly assist with mtDNA analysis, but ultimately, the analyst is responsible for checking every base call made by the software.

Using MTexpert™ as an expert system reduces the amount of manual intervention required for haplotype assignment. User-defined flags bring the analyst's attention to potential issues within individual traces (such as low quality base calls, or heteroplasmic base calls) and within the consensus sequence (such as disagreements between base calls in the individual traces, or low quality consensus calls). The analyst is only responsible for inspecting the flagged regions of sequence data. Once these flags are all addressed, the analyst inspects the Signature (i.e., haplotype) and reports the final results. Bases that are not flagged do not require analyst review.

**Time savings using MTexpert™ Software.** Times reported here are for the second data read and include loading traces into the software, making base calls, reporting haplotypes, and checking for concordance with the first read. The time reported for manual analysis of Batch 146 (26 samples) was 159 minutes. The time reported for manual analysis of Batch 147 (21 samples) was 186 minutes. Using MTexpert™ Software (and STATIS Software for checking concordance), the time reported for expert system analysis of Batch 146 was 52 minutes, and the time reported for Batch 147 was 58 minutes (Figure 17). Although this is a very small trial, a significant time savings of 68.05% was observed (two sample paired t-test for means, P-value = 0.028369359).

**Accuracy of results using MTexpert™ Software.** The resulting haplotypes for all 47 samples analyzed were 100% concordant between analysis methods.




| | *MTexpert* | *Manual* |
|---|---|---|
| *Batch 146* | 52 | 159 |
| *Batch 147* | 58 | 186 |
| *Average (time/batch)* | **55** | **172.5** |
| *Variance* | 18 | 364.5 |
| *P(T<=t) one-tail* | 0.028369359 | |

**Figure 17.** Results of time trial. Analysis of two batches, Batch 146 and 147, was performed manually using Sequencher™ Software and current CHI protocols, and with an automated, expert system approach using MTexpert™ Software. Significant time savings was observed. The resulting haplotypes were 100% concordant between analysis methods.

## *Conclusion- MTexpert™, STATIS and eFAST™ as an Integrated Expert System*

Using MTexpert™ Software in conjunction with eFAST™ Software and STATIS Software significantly decreases the time required for routine haplotype reporting *while producing accurate haplotype calls*. The MitoTyper™ Rules, which are integrated into MTexpert™, reduce ambiguity in base calling particularly when samples have variants in challenging regions. The reduced ambiguity will result in fewer missed hits or familial associations due to jumping alignments. This three part software suite has the potential to reduce backlogs while improving the accuracy of database searching by decreasing ambiguity in haplotype calling, both within and between mtDNA testing laboratories.

# *Barcoding*

Barcoding software was developed and has been enhanced to allow the user to electronically catalogue and track large numbers of samples in real-time. The system provides a means for optimized inventory and sample processing management whilst providing highly detailed electronic documentation. Adding a new sample to the inventory can be accomplished through two convenient methods; by scanning the submitting barcode or manual entry. After a sample is added to the database a new, unique barcode is printed and affixed to the sample package or container. All information associated with a sample is stored in a secure, limited access electronic database that utilizes the data management functionality of Microsoft® Access® (Figure 18).

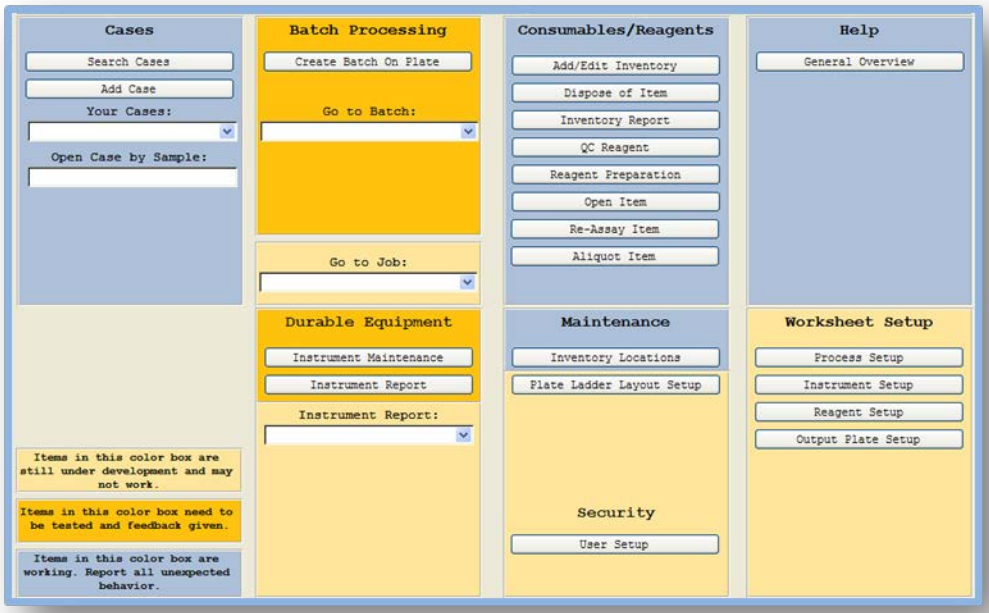



Figure 18. Barcoding software main screen and user interface. The design strategy for the system focused on delivering a valuable user experience that combines simplicity and visible functionality. The intuitive layout of the system's main screen and use of drop-down menus and buttons make navigation an easy task. Operation of different processes follows a logical progression making sample tracking and management more stream-lined.

The batch processing capability of the barcoding software is one of its most compelling features as it facilitates efficient and accurate sample processing management. Large batch sizes, such as for family reference samples, can be easily organized into a batch layout that includes the identification number and barcode, for each sample, in a familiar 96-well plate format (Figure 19). The system applies information from the batch layout to effortlessly create protocol worksheets for each of the different processes in the high throughput analysis work-flow procedure, *e.g.*, quantification, amplification, and cycle sequencing (Figure 20).

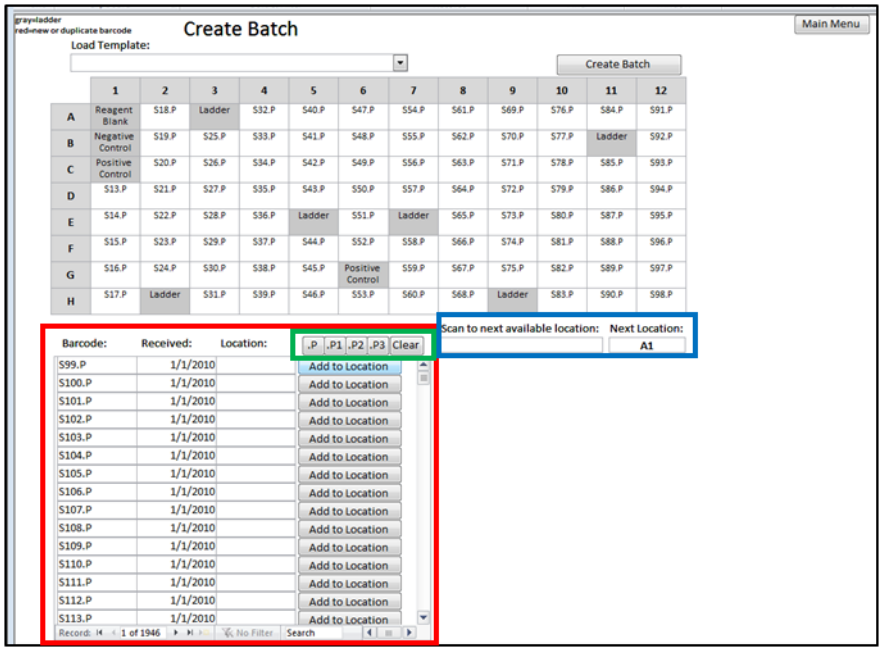



Figure 19. Batch creation menu. After the analyst is notified that enough samples are present to process a batch, a sample list is generated (red box). The analyst can also filter the samples based on sample type (green box). There are two approaches to creating a batch. The analyst can choose the appropriate plate template and select "add to location" and the samples will auto fill into the next available location of the 96-well plate or the analyst can scan the samples (blue box) and the samples will auto fill into the next available well.

In order to appropriately track a sample batch, the system generates a separate plate barcode for each process performed which allows for a more streamlined work-flow and reduces transcriptional errors (Figure 21).

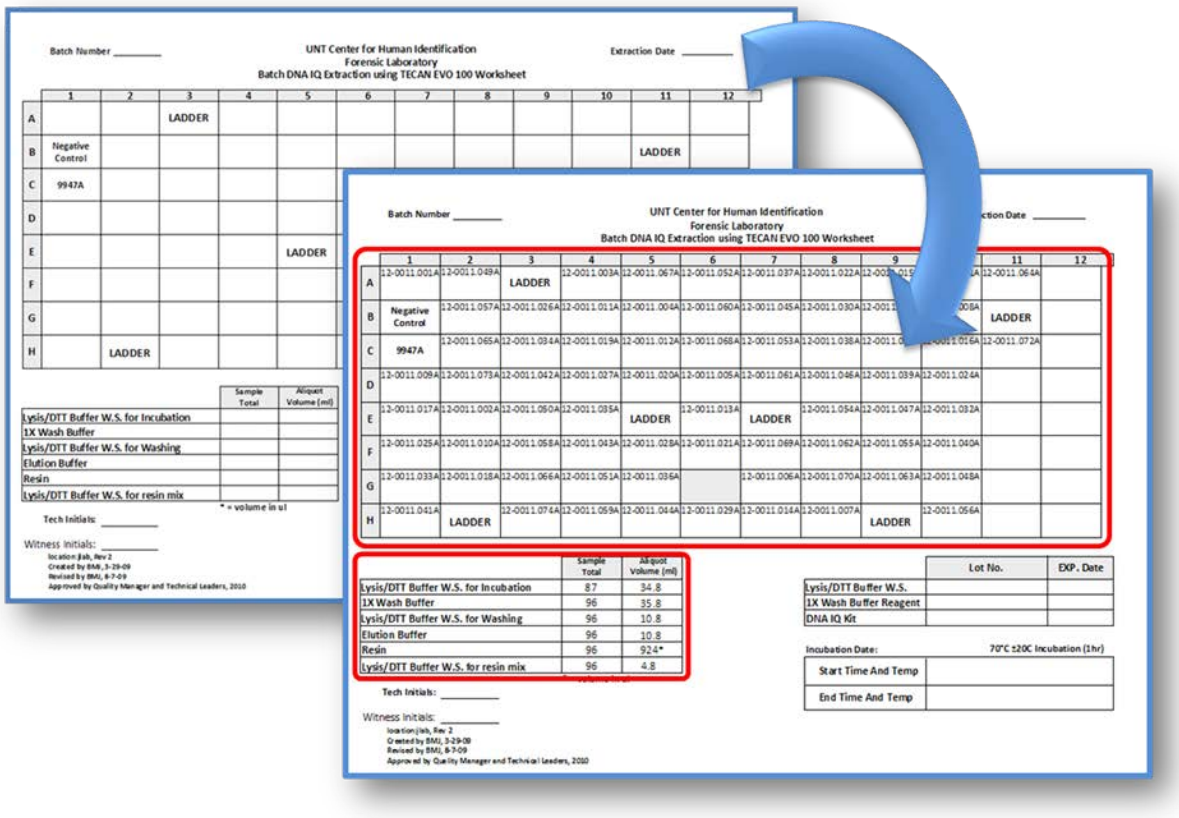



Figure 20. Automated protocol worksheet creation. From the batch view window, the system can automatically create a protocol worksheet which incorporates individual sample information and sample number at the touch of a button. Each worksheet becomes a digitized document that provides two significant time saving benefits. First, each sample from the batch layout is instantly added to a well position that is appropriate for the type of analysis performed and second, it calculates the volume of reagents needed (red boxes).

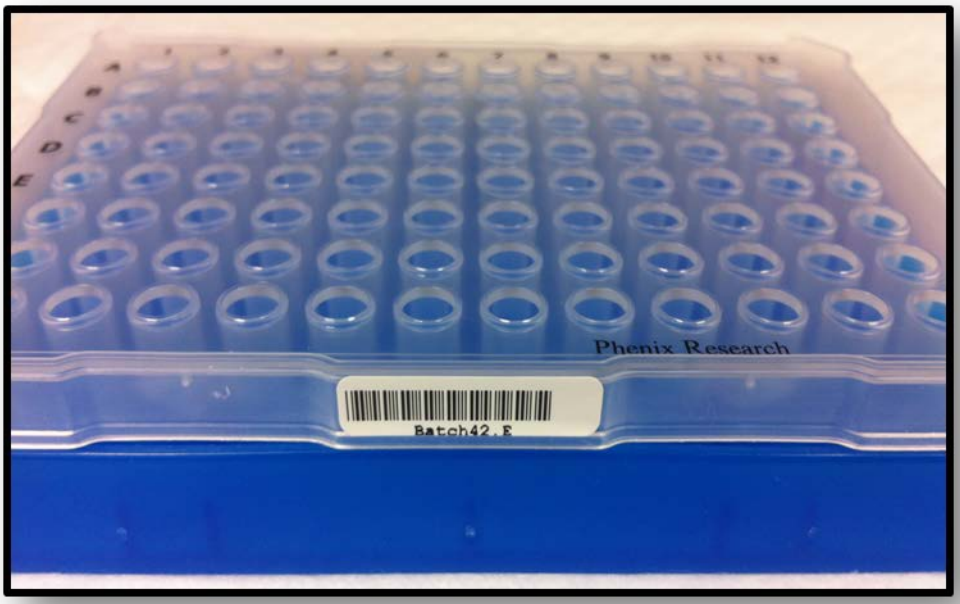



Figure 21. Plate barcode. Each processing plate for a single batch receives a unique barcode. The label displays the batch number and an abbreviation describing the plate type: E for extracted DNA, Q for quantification plate, N for normalized DNA, A for amplification plate, etc.

To further simplify laboratory processing, the software automatically calculates the volume of reagents needed for each step (Figure 20). During each process the necessary reagents and required instruments are scanned. The appropriate information is then retrieved from the sample tracking database and imported to the worksheet. Barcoding these steps decreases the chance for human error (*e.g.*, transcriptional errors), provide faster entry and are easier to read than handwriting.

The system's framework of menus and commands has been expanded to apply the same inventory tracking and management capabilities for instruments and reagents. This allows notification of important quality control measures for reagents (*e.g.*, notification of reagent in-service and whether or not the reagent is expired).

# *Conclusion*

The UNTHSC Field Testing Division has developed several software advancements in the analysis of mtDNA for reference samples that significantly reduces labor in both the laboratory and in data analysis, reducing the overall sample processing time. In addition, these software programs perform a quality check of the data and apply MitoTyper™ Rules for consistent haplotype reporting. A reduction in labor and processing time will improve efficiency and increase the overall capacity of mtDNA processing by the laboratory. With increased efficiency and capacity, more reference samples can be processed and hence, identifications can be recommended earlier. Considerable savings in costs and time can be achieved by implementing these programs.

The FBI has funded several regional mtDNA laboratories to conduct mtDNA casework as an extension of its own operations. In addition, the FBI and the Jan Bashinski DNA Laboratory of the California Department of Justice conduct casework for the Missing Persons Index System. If he UNTHSC Field Testing Division could share their software programs with all of these laboratories in a single setting or to provide the software to each of the laboratories for testing and evaluation, then additional enhancements and acceptance of the expert system suite proposed here could be made. This project suggests that such acceptance by the wider community could materially improve the output quality of mtDNA sequence data and reduce missing persons case processing backlogs throughout the United States.

# *References*

Budowle, B, Polanskey, D, Fisher, CL, Den Hartog, BK, Kepler, RB, and Elling, JW. (2010) Automated alignment and nomenclature for consistent treatment of polymorphisms in the human mitochondrial DNA control region. *Journal of Forensic Sciences*; 55(5):1190-1195.

Curtis, PC, Thomas, JL, Phillips, NR, and Roby, RK. (2010) Optimization of primer-specific filter metrics for the assessment of mitochondrial DNA sequence data. *Mitochondrial DNA,* 21(6), 191-191-197. doi:10.3109/19401736.2010.528756

Den Hartog BK, Elling JW, Kepler RB. MTexpert, an automated software system for forensic mitochondrial DNA data analysis. *Forensic Science International: Genetics Supplement Series*. 2009;2(1):102-3.

Frappier, R. (2005) GeneMapper® ID Software: Why did we switch? Applied Biosystems Forensic News Volume 3.

Holland, MM, Fisher, DL, Roby, RK, Ruderman, J, Bryson, C, and Weedn, VW. (1995) Mitochondrial DNA sequence analysis of human remains. *Crime Laboratory Digest, 22*(4), 108-115.

Isenberg, AR. (2004) Forensic mitochondrial DNA analysis. In R. Saferstein (Ed.), *Forensic science handbook, Volume II* (Second Edition ed., pp. 297-327). Upper Saddle River, NJ: Prentice-Hall.

Kadash, K, Kozlowski, BE, Beiga, LA, and Duceman, BW. (2004) Validation Study of the TrueAllele Automated Data Review System. *Journal of Forensic Sciences;* 49(4):660-667.

Phillips, N R, Thomas, JL, Pantoja, JA, Gonzalez, SD, Planz, JV, Eisenberg, AJ, and Roby, RK. (2009) The Chilean Population Database Project: A High Throughput Approach for DNA Profiling of 1000 Chilean Population Samples. Poster Presentation at the University of North Texas Health Science Center Seventeenth Annual Research Appreciation Day: Fort Worth, TX.

Roby, RK and Jones, JP. (2005) Evaluating Expert Systems for Forensic DNA Laboratories. Applied Biosystems Forensic News Volume 3.

Roby, RK and Christen, A. (2007) Validating Expert Systems: Examples with the FSS-i$^3$ Expert Systems Software. Profiles in DNA. 10(2), pp. 13-15.

Roby, RK, Williamson, P, Josserand, M, Planz, JV, Lorente, JA, and Eisenberg, AJ. (2007) Streamlining Mitochondrial DNA Sequencing of Reference Samples. *Eighteenth International Symposium on Human Identification*, Hollywood, CA.

Roby, RK. (2008) Expert Systems Help Labs Process DNA Samples. NIJ Journal, No. 260, pp. 16-19.

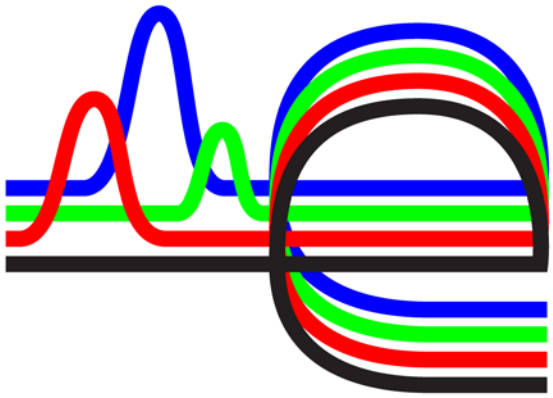
Roby, RK, Thomas, JL, Phillips, NR, Gonzalez, SD, Planz, JV, and Eisenberg, AJ. (2009) High Throughput Processing and Increased Efficiency for Mitochondrial DNA Testing: Robotics, Automated Sample Tracking and Filter Metrics. Poster Presentation at the Fifteenth American Academy of Forensic Science: Denver, CO.

Roby, RK, Phillips, NR, Thomas, JL, Keppler, R, and Eisenberg, AJ. (2010) Quality Assessment and Alert Messaging Software for Raw Mitochondrial DNA Sequence Data. Conference Proceedings of the Sixteenth American Academy of Forensic Sciences: Seattle, WA.

Roby, RK, Phillips NR, Thomas, JL, Sprouse, ML, and Curtis PC. (2011) Development of an Integrated Workflow from Laboratory Processing to Report Generation for mtDNA Haplotype Analysis. NCJ 234292 Grant Report.

# *Appendix A: eFAST™ Software 2.0 User Guide*

This page intentionally left blank.



# eFAST™ Software v2.0

## User Guide

University of North Texas
Health Science Center

UNT HEALTH SCIENCE CENTER

**For Research, Forensic, or Paternity Use Only. Not intended for any animal or human therapeutic or diagnostic use.**

Information in this document is subject to change without notice.

UNIVERSITY OF NORTH TEXAS HEALTH SCIENCE CENTER DISCLAIMS ALL WARRANTIES WITH RESPECT TO THIS DOCUMENT, EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO THOSE OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. TO THE FULLEST EXTENT ALLOWED BY LAW, IN NO EVENT SHALL UNIVERSITY OF NORTH TEXAS HEALTH SCIENCE CENTER BE LIABLE, WHATHER IN CONTRACT, TORT, WARRANTY, OR UNDER ANY STATUTE OR ANY OTHER BASIS FOR SPECIAL, INCIDENTAL, INDIRECT, PUNITIVE, MULTIPLE OR CONSEQUENTIAL DAMAGES IN CONNECTION WITH OR ARISING FROM THIS DOCUMENT, INCLUDING BUT NOT LIMITED TO THE USE THEROF, WHETHER OR NOT FORESEEABLE AND WHETHER OR NOT UNIVERSITY OF NORTH TEXAS HEALTH SCIENCE CENTER IS ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

**NOTICE TO PURCHASER:**

Not of resale.

**TRADEMARKS:**

Trademarks of the University of North Texas Health Science Center and MitoTech LLC include: eFAST™.

Windows is a registered trademark of Microsoft Corporation in the U.S. and other countries.

All other trademarks are the sole property of their respective owners.

Part Number XXXXXXX DRAFT
3/2012

# Contents

# Preface

## How to use this guide

### *Purpose of this guide*

The eFAST™ Software v2.0 User Guide provides information about the software.

### *Text conventions*

This guide uses the following conventions:

- **Bold** text indicates user action. For example:

  Type **20**, and then press **Enter** for each of the remaining fields.

- *Italic* text indicates new or important words and is also used for emphasis. For example:

  Before analyzing, *always* prepare fresh matrix.

- A right arrow symbol (▶) separates successive commands you select from a drop-down or shortcut menu or text indicates user action. For example:

  Select **File ▶ Open ▶ Spot Set.**

### *User attention words*

Two user attention words appear in this guide. Each word implies a particular level of observation or action as described below:

**NOTE:** - Provides information that may be of interest or help but is not critical to the user of the product.

**IMPORTANT:** - Provides information that is necessary for proper operation of the software.

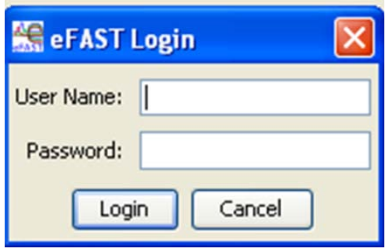
## How to obtain support

# eFAST™ Software v2.0 Operation

## Setup eFAST™ Software v2.0 for Trace Quality Assessment

1. **Launch** eFAST™ Software v2.0.


2. **Login** to eFAST™ Software v2.0.




**IMPORTANT:** User Name and Passwords are case-sensitive.

**NOTE:** To add an additional user to eFAST Software, an administrator must login. The administrator must select Edit ▶ Users ▶ Add User. To remove a user, an administrator must login and select Edit ▶ Users. Highlight the appropriate user and select Delete User.

3. **Refer** to Appendix A for proper naming conventions for sequence data to be recognized by eFAST™ Software.

**IMPORTANT:** Specific naming conventions are required for eFAST™ Software to recognize sequence data.

4. **Refer** to Appendix B in order to modify thresholds for quality assessment.

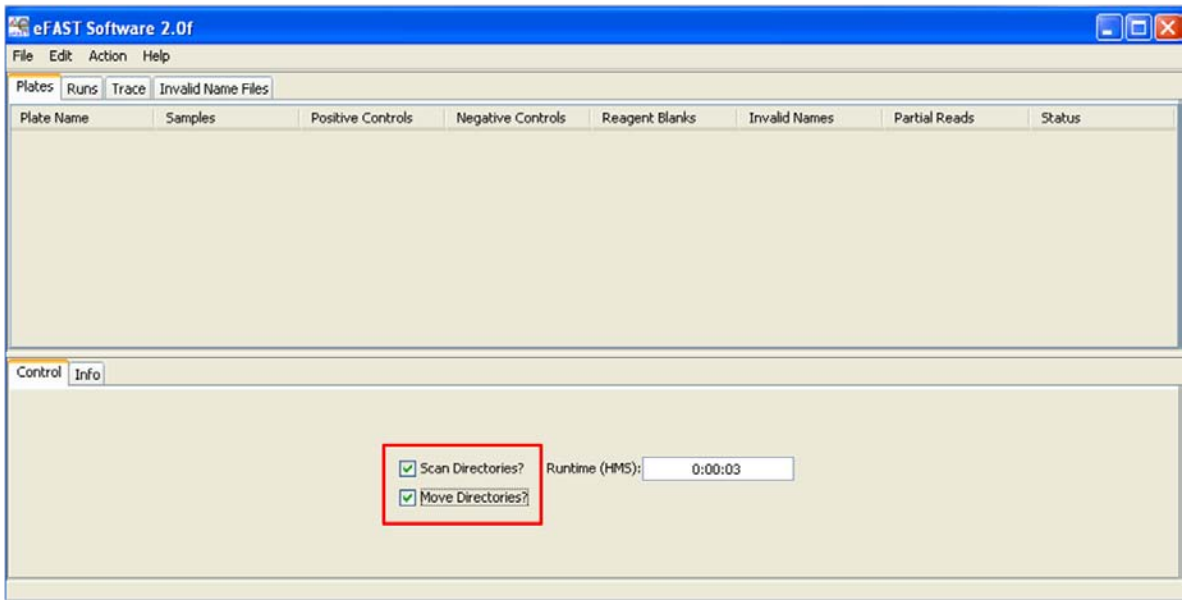
**NOTE:** To establish the individual thresholds, internal validation studies should be conducted.

5. **Verify** the email and directory information in the preferences menu. **Go** to Edit ▶ Preferences ▶ Plate/Run. **Select** the appropriate run name pattern and **verify** the directory information listed at the bottom of the menu. **See** Appendix C for additional information.

6. **Verify** the information in the Instrument tab. **Go** to Edit ▶ Preferences ▶ Instrument. **See** Appendix D for additional information.

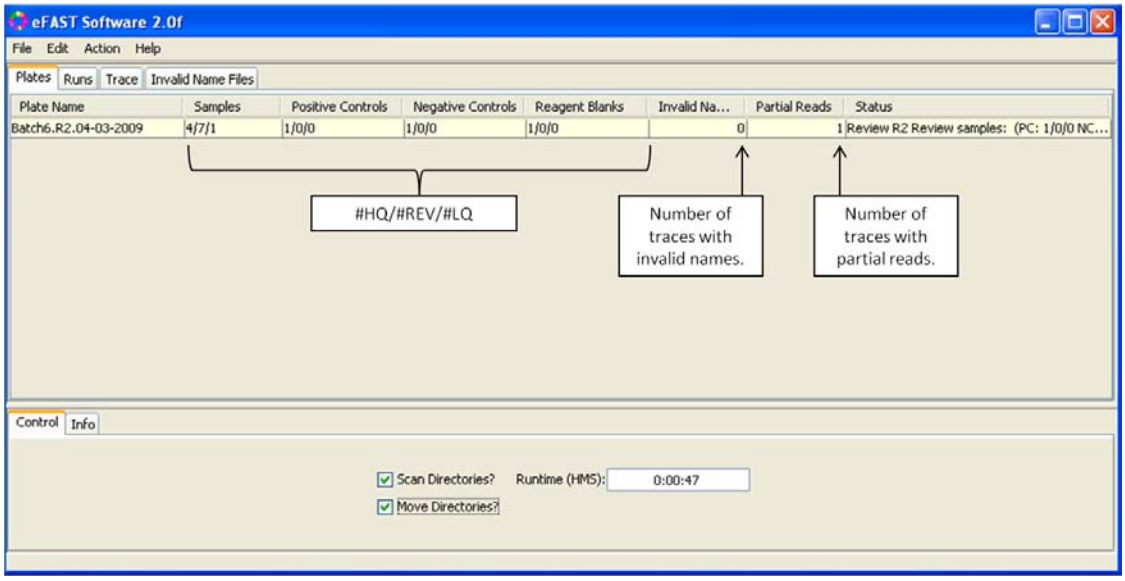
**7.** In the *Plates* tab, **check** the *Scan Directories?* and *Move Directories?* boxes. *Scan Directories?* initiates automatic sample name parsing and quality assessment. *Move Directories?* initiates trace distribution into the destination directories upon plate completion, as specified in the paths tab of the preferences menu.

> **NOTE:** If a plate is complete but eFAST™ Software has not yet detected the completion, the user can **right click** the plate name and **select** "Mark Complete".



## Trace Quality Assessment

1. **Select** the plate to analyze.




2. The *Plates* tab displays the plate name, the quality of the controls, the number of samples with invalid file names and the number of samples with partial reads, as well as, the status of the plate. The quality summaries are in the following format: number of high quality (#HQ)/number of reviewable samples (#REV)/number of low quality samples (#LQ).

3. **Select** the *Runs* tab to display only selected run information, or **select** the *Trace* tab to view all traces associated with the plate.

> **NOTE:** Any traces that do not follow an established naming convention will be listed in the *Invalid Name Files* tab. **Select** the *Invalid Name Files* tab to display the sample(s) with invalid trace names. **Right click** the trace name to change the file name or remove the file from the eFAST™ Software run. **See** Appendix A for additional information regarding sample naming.

4. The *Trace* tab displays the following information:

> **NOTE:** To establish the individual thresholds, internal validation studies should be conducted.

   a. Plate Name: the name assigned to the plate processed on the sequencing instrument.

   b. Trace File Name: the name of the file as designated by the user and the analysis protocol's naming convention.

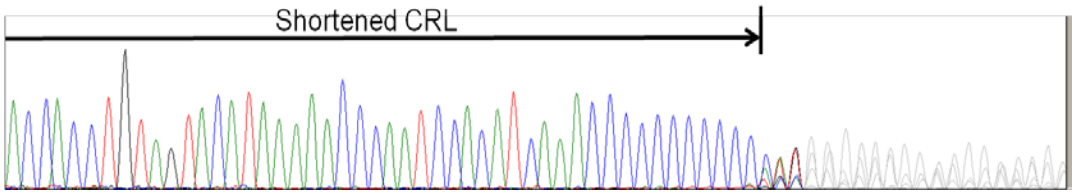
c. Sample Name: the sample name assigned by the user.

d. Primer: the primer used when cycle sequencing.

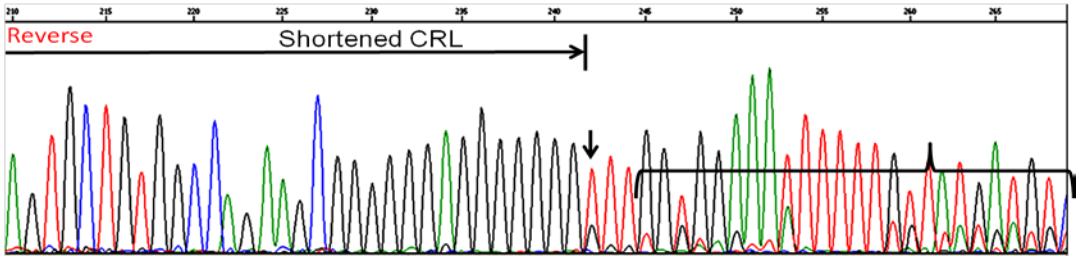
e. Trace Score (TS): the average quality value (QV) of the post-trim sequence.

f. Contiguous Read Length (CRL): the longest, uninterrupted stretch of bases with a QV of a defined value.

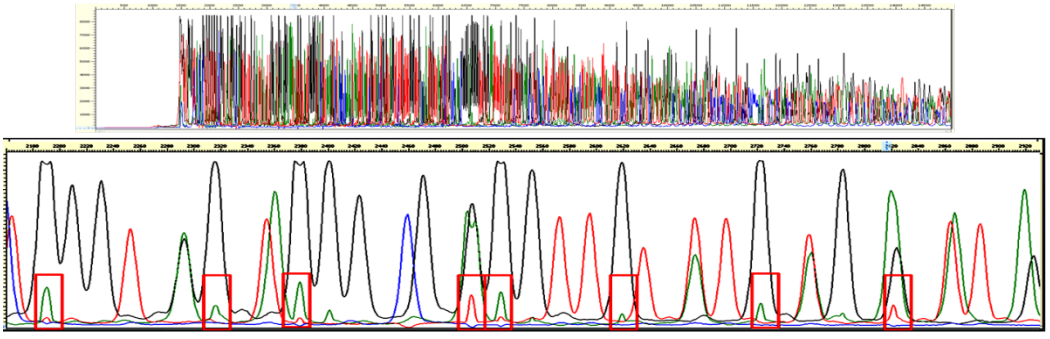
g. *Homopolymeric Stretch (HPS): alerts the user if a homopolymeric C-stretch is observed in HV1. An example of this is shown below.



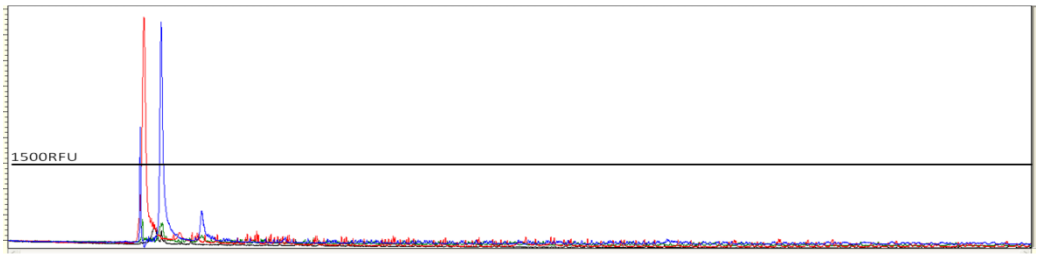

h. *Length Heteroplasmy (LH): alerts the user if length heteroplasmy is observed in HV2. This is indicated by the presence of a variable number of Cs at position 303 resulting in frame-shift sequence nested downstream. An example of this is shown below.



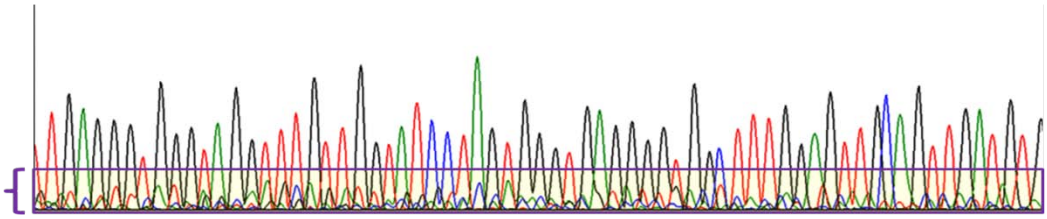

i. *Pull-Up (PU): alerts the user if pull-up is observed in the sequence, typically caused by saturation of the CCD camera. An example of this occurring is shown below in the raw data.



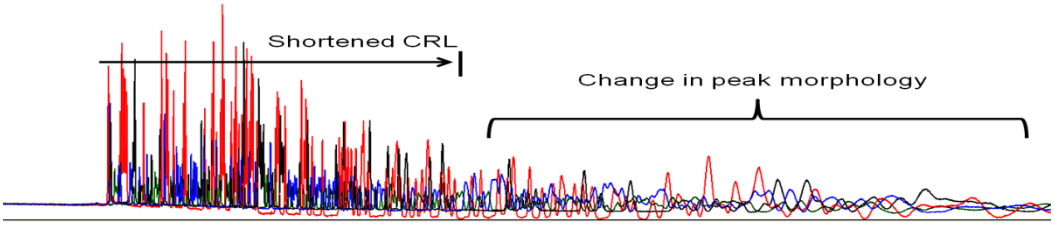
j. *Low Signal (LS): alerts the user if the signal of the sample is below a specified threshold. An example of this is shown below in the raw data.



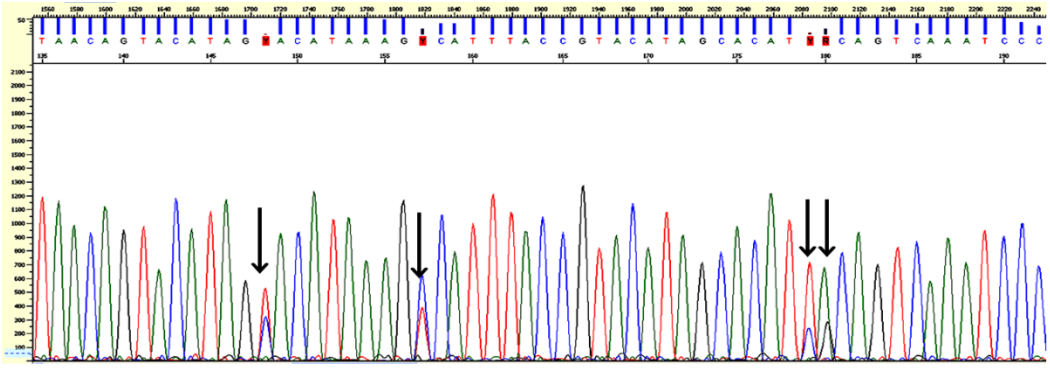

k. *High Baseline (HB): alerts the user if the baseline signal of a sample exceeds a specified threshold indicative of the level of peak noise within the called bases. An example of this is shown below.



l. *Partial Read (PR): alerts the user if the CRL is not reached but quality sequence was initially obtained, indicating an electrophoretic mobility result that can often be repeated with longer read lengths by reinjecting the sample. An example of this is shown below in the raw data.

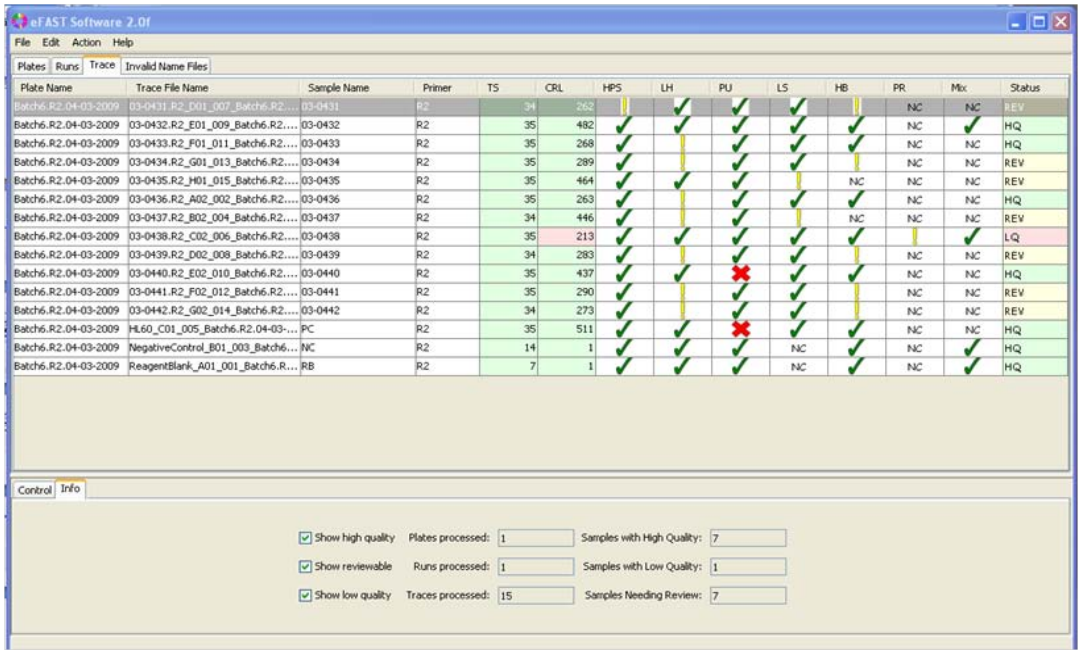



m. *Mixture (Mix): alerts the user when the number of high quality mixed bases detected in a sample exceeds a defined threshold. An example of this is shown below.



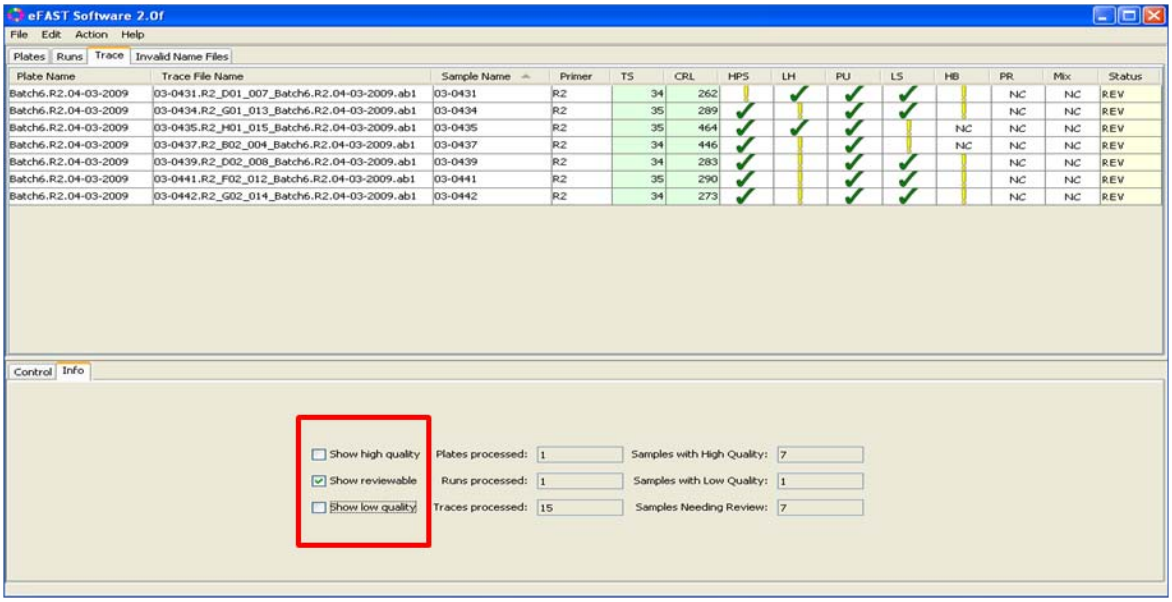
*These features alert the user of sequence quality. The status of each is deemed pass (✓), review (!) and fail (✖) and not checked (NC).

n. Status: displays the overall status of the sample. Status is defined as high quality (HQ), review (REV) and low quality (LQ).

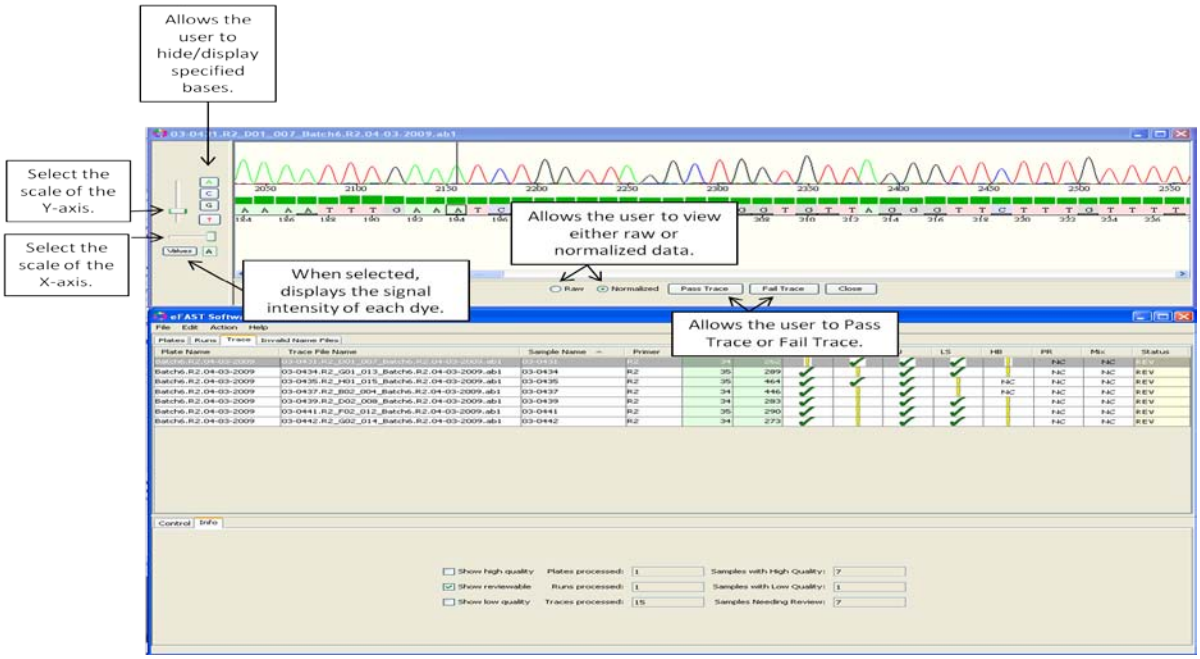



## Review Traces

1. To display only traces marked “REV”, **select** the *Info* tab at the bottom of the menu and **check** “Show reviewable”.




2. **Double-clicking** the Trace File Name will launch the trace viewer.




3. **Right click** the trace name to pass or fail a trace, as well as, view trace property information.

4. To view a sample report, **select** Action ▶ Sample Report.

5. If more than one sample directory is present, the analyst will be prompted to select the appropriate sample directory.

6. Once all traces have been designated as pass or fail, they will be distributed accordingly. Samples can now be evaluated using MTexpert™ Software.

# Appendix A

## General eFAST™ Software Naming Conventions

*Plate Name:*

BatchN.PP.INT.MM-DD-YYYY

N = Batch Number (any number)
PP = Primer (any combination of numbers and characters)
INT = Initials of the analyst (any 3 letters)
MM = Two digit month (01-12; numeric only)
DD = Two digit day (01-31; numeric only)
YYYY = Four digit year (0000-9999; numeric only)

*Sample Names:*

YY-CCCC.SX.PP.A

YY = Two digit year (00-99; numeric only)
CCCC = Case/project number (0000-9999; numeric only)
S = Sample number (any number of digits; numeric only)
X = Additional sample information (must begin with any character, then any combination of characters and numbers may be used; *optional*)
PP = Primer (any combination of numbers and characters)
A = Amplification number (0-9; *optional*)

*Control Names:*

C.PP.A

C = Control type (HL60, NegativeControl, ReagentBlank)
PP = Primer (any combination of numbers and characters)
A = Amplification number (0-9; optional)

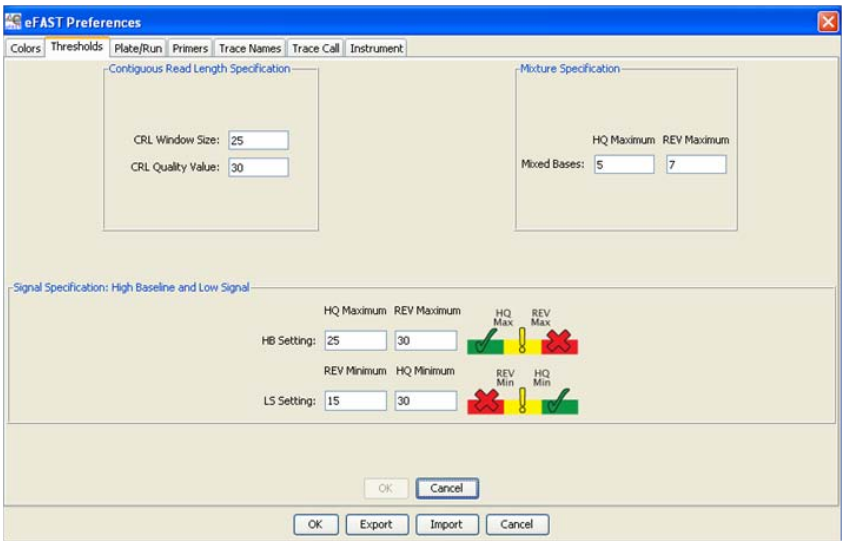
**NOTE:** Additional name conventions are possible.

# Appendix B

## Trim Threshold and Primer Specific Criteria

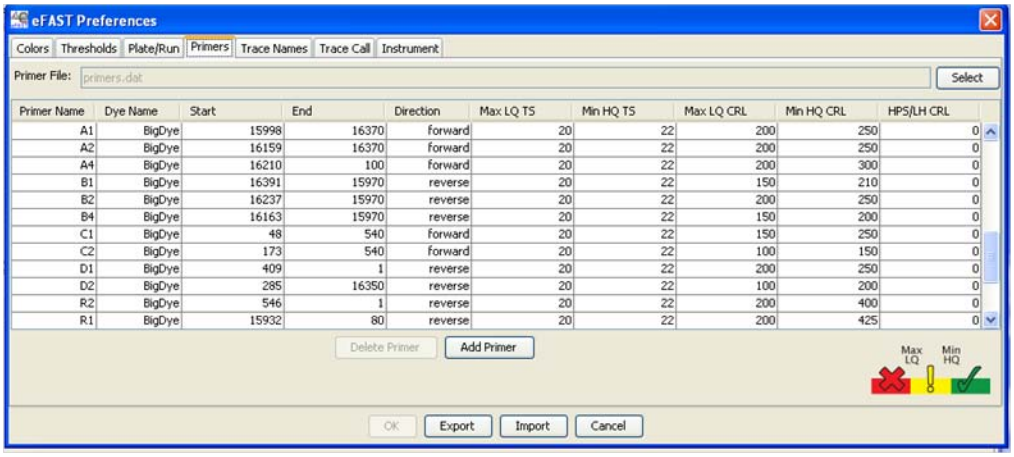
*Thresholds* tab:

- Contiguous Read Length Specification
- Mixture Specification
- Signal Specification:  High Baseline and Low Signal




*Primers* tab

Primer specific thresholds can be specified for trace score and contiguous read length.

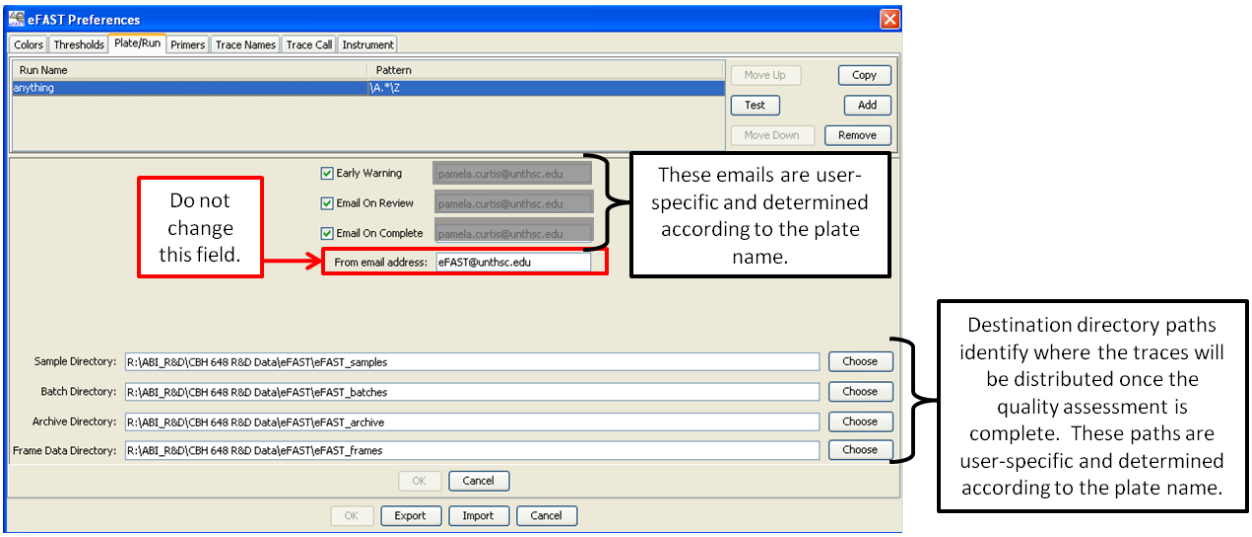


# Appendix C

## Destination Directories and Email Preferences

*Plate/Run* tab:

**Identify** the user-specific destination directories and email addresses for the notifications of the distribution of traces.

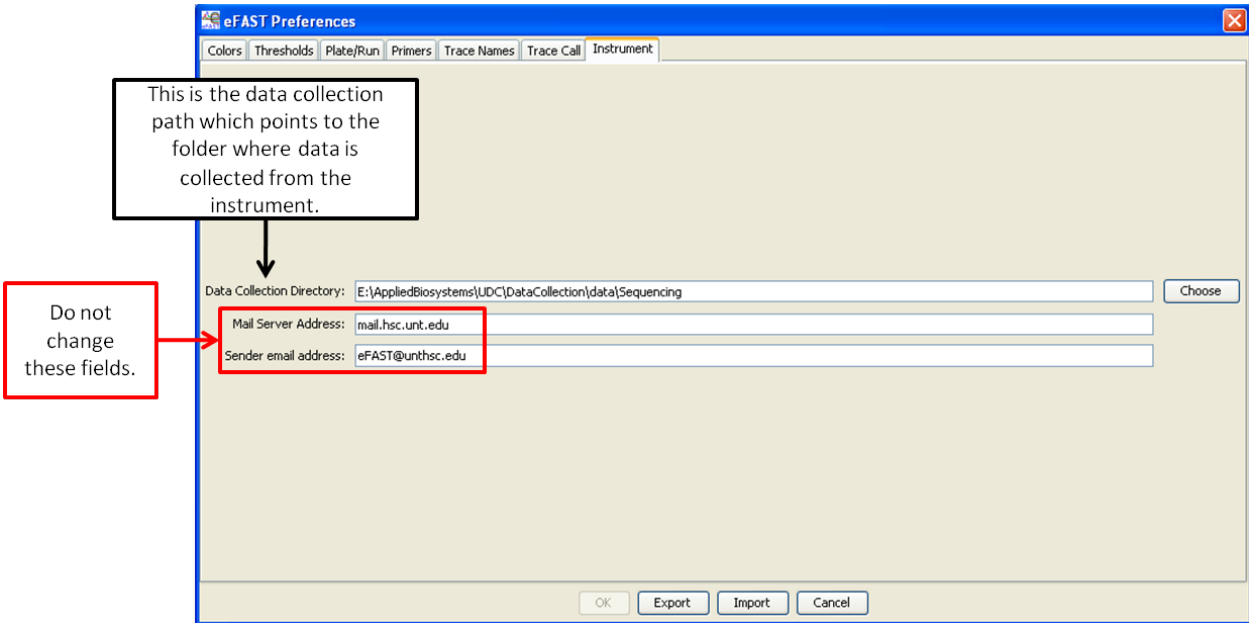


# Appendix D

## Data Collection Directories and Instrument Email

*Instrument* tab:

**Insert** the address for the Data Collection Directory and instrument email address information.



# Bibliography

1. Sprouse M., Kavlick M., Phillips N., and Roby R. Internal Validation of a Real-Time qPCR Assay for Human mtDNA. Conference Proceedings of the Eighteenth American Academy of Forensic Science 2012: Atlanta, GA.

2. Phillips N. UNTHSC Mitochondrial Working Group: Fort Worth, TX; June 24, 2011.

3. Sprouse M. UNTHSC Mitochondrial Working Group: Fort Worth, TX; June 24, 2011.

4. Phillips N. and Roby, R. Expert System Rules and Software Advancements for Mitochondrial DNA Analysis. Conference Proceedings of the Seventeenth American Academy of Forensic Sciences 2011: Chicago, IL.

5. Phillips N. eFAST© Software: Automated Quality Assessment, Alert Messaging, File Distribution and Sample Tracking of Mitochondrial DNA Sequence Data. Proceedings of the University of North Texas Health Science Center Eighteenth Annual Research Appreciation Day 2010: Fort Worth, TX.

6. Roby R., Phillips N., Thomas J., Kepler R., Elling J., and Eisenberg A. Quality Assessment and Alert Messaging Software for Raw Mitochondrial DNA Sequence Data. Conference Proceedings of the Sixteenth American Academy of Forensic Sciences 2010: Seattle, WA.

7. Sprouse M, Phillips N, and Roby R. Internal Validation of a Real-Time Quantitative Polymerase Chain Reaction Assay for Human Mitochondrial DNA. Poster Presentation at the University of North Texas Health Science Center Nineteenth Annual Research Appreciation Day: Fort Worth, TX; April 1, 2011.

# *Appendix B: Issues and Suggestions Report for MTexpert*

This page intentionally left blank.

# MTexpert Review

Issues and Suggestions Report

| Item # | Suggestion or Issue | Analyst's Action | Description of the Issue or Suggestion | MitoTech Response |
|---|---|---|---|---|
| UNT-01 | Issue | Viewing Trace Windows of a sequence variation by double-clicking the row in the Signature Table. | The traces open and minimize immediately; user must individually click each Trace Window to restore the windows. Usually the traces launch correctly when opened from the Consensus Issues table, although the same minimization occurs sporadically. | The software is not designed for windows to open on top of each other. The main analysis window should be sized to a portion of the screen and the trace windows designated to open in the unoccupied space. This adjustment resolved the problem. |
| UNT-02 | Issue | Trimming the forward and reverse sequences at the 303 length heteroplasmy position to remove the resulting mixed bases. | After trimming, the sequences do not align properly with the reference, and an incorrect signature is generated (see Figure 1). | Russ demonstrated that untrimming the data a few bases beyond the start of the heteroplasmic data gives the aligner the coverage required to align the data into one contig (see Issue UNT-24). |
| UNT-03 | Issue | Viewing a consensus issue in the Trace Windows. | The trace position in the Trace Window view does not correspond to the position provided in the table; it appears that the position in the table is the post-trimmed position, whereas the position in the trace window is based on the pre-trimmed sequence (see Figure 2). A more dramatic marker of the implicated base would be beneficial. | This issue was resolved. |
| UNT-04 | Issue | File→Create New Project. | The name of the project does not change when a sample directory chosen, then is changed (see Figure 3). | This issue was resolved. |
| UNT-05 | Suggestion | Importing sample project directories. | With the configuration of sample directories, the user can only import one sample directory into a project at one time. Multiple selections are not allowed when importing directories. | MTexpert is not configured for analysis of multiple samples if the appropriate control data is not included. The procedure utilized by UNTCHI R&D for processing batched samples (*i.e.,* one primer per plate) does not correlate with the control checks in place in MTexpert. MitoTech has mentioned working on adjusting the control check code for this type of high-throughput processing. |

| Item # | Suggestion or Issue | Analyst's Action | Description of the Issue or Suggestion | MitoTech Response |
|---|---|---|---|---|
| UNT-06 | Suggestion | Editing data in an unsaved program. | An auto-save option might be helpful to prevent loss of progress in the event of unexpected program termination. Some programs auto-save every 5 minutes. | Due to the large size of the projects assembled in MTexpert, saving takes a considerable amount of time. For this reason, auto-saving would be distracting while an analyst is working in a project. This suggestion will not be pursued. |
| UNT-07 | Suggestion | Analyzing the electropherograms in the Trace Windows. | A greater zoom on the data in the Trace Window would is needed for analyzing low level sequence. | The zoom will be increased from 4X to 8X on a linear scale. A 10X zoom was incorporated in v.4.3. |
| UNT-08 | Suggestion | Including a trace file in the contig that extends beyond the reference. | An option to automate trace trimming to the reference sequence would be very beneficial (see Figure 5). | This option will be incorporated as an Auto-Edit feature. |
| UNT-09 | Issue/ Suggestion | Analyzing the electropherograms in the Trace Windows. | When the horizontal slide is used in one Trace Window, the other windows do not sync accordingly. Likewise, if you scroll horizontally through the sequence in one window, the other windows do not follow. It is difficult to realign the trace windows once one of the windows is out of sync. Additionally, the analyst would benefit from having the y-axis slide operate independently in the Trace Windows to analyze data of different intensities simultaneously. | The initial design did allow for the Trace Windows to scroll in sync; however, it challenged the GUI, creating a choppy flow. MitoTech will not pursue this recommendation.<br><br>MitoTech will unlink the y-axis scale slide between the windows and will add a right-click option to bring all Trace Windows to the same scale. This allows the analyst to quickly adjust the scale uniformly as desired. |
| UNT-10 | Suggestion | Analyzing the sequence variations in the Trace Windows. | In the Trace Window view, include a yellow bar at positions that differ from the reference. | This recommendation will be considered. |
| UNT-11 | Issue | Reporting the signature and the coverage range for a sample. | The ranges of coverage/analysis for each sample are provided in the reports, but they do not correlate to the double coverage positions (see Figure 6). | This issue was resolved. |

| Item # | Suggestion or Issue | Analyst's Action | Description of the Issue or Suggestion | MitoTech Response |
|---|---|---|---|---|
| UNT-12 | Suggestion | Reporting the concordance check between first and second reads. | Compare the coverage ranges from each analyst, and report the overlapping coverage range between the two readers. For example: Reader 1 has coverage from 15950-379, and Reader 2 has coverage from 15962-530; the conservative range is 15962-379. Variations outside of this range are considered not confirmed. | MitoTech realizes the value of this incorporation; however, the coverage range is not reported in a constant position in the text exports from other programs. For concordance checks between two reads in MTexpert, the conservative range can be included in the report. Further consideration must still be given to how the Signature will be modified given the conservative range (*i.e.*, which project will be modified, and how will the first and second readers' original calls be preserved?). |
| UNT-13 | Expert Rule Suggestion | Assessing quality of the imported traces. | Evaluated each traces for the possibility of a mixture using some threshold for the number of heteroplasmies (n) above some percentage (x) of the major peak. For example, if a single trace has more than 2 point heteroplasmies called, both above 40% of the major peak, then a rule can fire indicating the possibility of a mixture. The interpretation guidelines for the UNTCHI R&D Laboratory allow for 2 point heteroplasmies in a signature. More than that is suspect. Of course, the values for n and x can be customizable/validated. | The programming for this rule could get complicated; MitoTech will consider this and other trace level assessments/rules. |
| UNT-14 | Expert Rule Suggestion | Trimming data after a C-stretch or length heteroplasmy. | Automate the required trimming around C-stretch and length heteroplasmies. The UNTCHI R&D Laboratory trims the sequence after the last "clean" C in the HV1 C-stretch (both in the forward and the reverse stands). For HV2 length heteroplasmy, sequence is trimmed from the major 310T→5' in the forward strands, and from the last clean C (the 303C)→3' in the reverse. | Version 4.2 includes a HV1 C-stretch trim that can be selected in the Auto Edits menu. The HV2 trim will be much more difficult to program, but this feature is still being considered. |
| UNT-15 | Expert Rule Suggestion | Assessing appropriate coverage for double confirmation. | If a sample directory contains duplicate data (*i.e.,* sequence from the same primer and the same amplification), a rule firing would be good to alert the analyst that only one of the two sequences can be used for coverage/confirmation. This rule would be heavily dependent on the user's naming convention. A more simplistic approach would be a rule firing if there are more than two sequences from any one primer in a project (see Figure 7). | This suggestion will be considered as a trace level rule firing (as mentioned in UNT-13). |

| Item # | Suggestion or Issue | Analyst's Action | Description of the Issue or Suggestion | MitoTech Response |
|---|---|---|---|---|
| UNT-16 | Suggestion | Reporting Signatures. | Include the logged in analyst's name in the reports generated. | This recommendation will be incorporated (not included as of v.5.1). |
| UNT-17 | Suggestion | Viewing coverage of HV1 and HV2 in the Assembly Map. | Add an indicator for where HV1 and HV2 lie in the Assembly Map. A quick visual of the coverage through the required typing regions would be useful. | This recommendation will be considered. Added in versions after 4.2. Rhonda suggested that these range markers reflect the user defined signature range. Russ indicated that the programming of this is difficult, but he is still considering a solution. |
| UNT-18 | Issue | Using the heteroplasmy-based auto trim feature. | Many traces are not trimming based on the designated trim criteria (Figure 8). | MTexpert begins "walking" the trace from the end of the trace, and when it reaches a window that passes the criteria, it stops trimming there. The reason HV2 trimming is presenting problems is because herteroplamies are plentiful only in the first 200 bases after the LH. When the peaks broaden and the signal weakens, not as many are called. The window scan for heteroplasmies is arrested too early. MitoTech has mentioned investigating some other trimming mechanisms for this region specifically. |
| UNT-19 | Issue | Importing sequences for a sample and creating signature | MTexpert could not generate an aligned consensus sequence for sample data 00-0001 (staff profile from Batch 13). | Russ forwarded this sample to Bobi who is investigating the problem. They believe that one of the problems is in the code for primer name specification. The primer name .lut files were sent with the 4.3 .jar. |
| UNT-20 | Issue | Using the Complete Alignment feature to force a single contig. | The Complete Alignment featue generates an error "The signature generatioin failed: Untrapped Exception in > the signature generation, message is: fromIndex(5)>toIndex(-1)". | Russ reported that this error is due to a bug on their end (v.4.3). Communication through email. |
| UNT-21 | Issue | Referring to the Signature/Consensus window to determine the position in the D-Loop. | The Consensus window scroll bar would not extend to the bottom of the display, cutting off the position numbers | This issue was resolved (v.4.3). |
| UNT-22 | Suggestion | Accidentally override a base call incorrectly. | MTexpert automatically assigns all over-ride calls a high quality, even if the call disagrees with the consensus call. A downstream rule should fire if any conflicting base calls exist. | This issue was only seen in v.4.0. |
| UNT-23 | Issue | Depending on high quality base calls made automatically with MTexpert. | Often the basecaller assigns an incorrect base call with a high degree of certainty (high quality score). These variants are reported in the signature table and are never called to the attention of the analyst (see Figure 9). | Russ was able to resolve this issue by removing the FBI's .lut files from the program directory. The exact reason that these file were affecting the base caller is not known. |

| Item # | Suggestion or Issue | Analyst's Action | Description of the Issue or Suggestion | MitoTech Response |
|---|---|---|---|---|
| UNT-24 | Issue | Analyzing samples with HV1 and/or HV2 C-stretches/length heteroplasmies. | The HV1 and HV2 joiners are not assembling the multiple consensus sequences into one contigs. | This issue was resolved in v.4.3. The primer names for C1 and R2 were not correctly entered into the joiner. |
| UNT-25 | Issue | Analyzing data for a length of time (approximately 10-20 minutes). | The typer begins to time out and the program freezes up. The user can close and reopen the project and the typer calls the sample fine. On occasion, the user must ctrl+alt+del to exit the program. | There was a significant memory leak. The problem was resolved through versions 4.4a, 4.4b, 5.0, and 5.1 |
| UNT-26 | Issue | Using the yellow HV1 and HV2 indicator boxes to determine coverage. | The indicator boxes seem to be stationary/independent of the actual consensus sequence. When excess sequence is present either 5' or 3' of the HV regions, the yellow boxes no longer accurately correspond to the HV1 and HV2 regions (see Figure 10). | This issue was resolved in v. 5.1. |
| UNT-27 | Issue | Using the Auto Edit feature of MTexpert. | The problem in Auto Edit appears when there were 2 or more edits in the same trace in the same pass. | This issue was resolved. |
| UNT-28 | Issue | Using the quality based trimming option. | In version 5.0, this user is able to specify the window and quality level of bases for auto trimming. Regardless to the values entered, the traces were not trimmed more/less stringently (*i.e.*, the feature was not activated). | This issue was resolved in v.5.1. |
| UNT-29 | Issue | Using the HV2 length heteroplasmy caller. | This feature is available to select; however, the multiple species are not being called in samples with the HV2 length heteroplasmy. | This issue was reported and acknowledged. The programmers are working on this feature. |

**Note:** Items UNT-01 – UNT-29 were assessed using older versions of MTexpert™ Software.

| Item # | MTexpert version | Date Reported | Suggestion or Issue | Analyst's Action | Description of the Issue or Suggestion | MitoTech Response |
|---|---|---|---|---|---|---|
| UNT-30 | 6.0 | 020711 | Issue | Project save, 86 samples | When saving the project, the program window went blank and was frozen for approximately 2 minutes. It might be due to the size of the project; if so, a progress indicator is needed. FYI, the memory counter was at 164627088. | 021111- Russ in Friday call. Corrected; added a save progress bar for when saving a large project to a network. |
| UNT-31 | 6.0 | 020711 | Issue/Suggestion | Viewing the sample in the Assembly Map window | MTexpert does not give a consistently clear overview of coverage. The Assembly Map window has no points of reference to indicate where the coverage is occurring and is inconsistently sized based on coverage, which can be confusing [See S1]. | 021111- GTM Russ and Nicole. Discussed making alterations in the Contig window. Possible right click→ zoom out option to give a better picture of coverage. |
| UNT-32 | 6.0 | 020711 | Issue | Viewing Sample Quality in Batch mode | All samples here appear Red, indicating a control problem. All issues with the PC were resolved and the sample was successfully typed. However, the Samples tab still says "Sequence value M at mte.C1". [See S2] | GTM Russ and Nicole, 021111. The error was not reproduced on Nicole's PC. However, we will be operating in sample mode so this problem should not be of real consequence in the future. He is considering a mechanism for validating our PCs though. |
| UNT-33 | 6.0 | 020711 | Suggestion | Preferences menu→Reference tab | "Enable HV2 C-stretch" would more accurately be titled "Enable HV2 Length Heteroplasmy" | |
| UNT-34 | 6.0 | 020711 | Suggestion | Preferences menu → Reference tab | Many laboratories use two reporting ranges: HV1 and HV2. Polymorphisms between the two regions are not usually reported; therefore, two ranges for Signature reporting would be beneficial. | 021111- GTM Russ and Nicole. Up until now, MitoTech has not been asked to allow for split signature range. They are finding a way to incorporate the HV1 and HV2 coverage range separately. |
| UNT-35 | 6.0 | 020711 | Issue | Inserted base | When a base is inserted, the base call is not visible in the GUI. [See S3] | 021111- GTM Russ and Nicole. Demonstrated the problem. Russ identified why the error in display was occurring, and is contemplating a solution. |

| Item # | MTexpert version | Date Reported | Suggestion or Issue | Analyst's Action | Description of the Issue or Suggestion | MitoTech Response |
|---|---|---|---|---|---|---|
| UNT-36 | 6.0 | 020711 | Suggestion | Using the Trace Actions tab | It is somewhat confusing to see the trace edits logged with the trace names. It would be nice to have another tab called "Traces" which lists the names and statuses of the traces separately from the "Trace Edits (or action)" tab. | |
| UNT-36 | 6.0 | | Issue | Project Save | When saving a project, MTexpert often displays an error "Cannot Save Project". When this occurs, the only way to save progress is to Save As, which creates a new project. | |
| UNT-37 | 6.0 | 030411 | Suggestion | Viewing Assembly Map | The green and red trace indicators imply quality (i.e., good or bad) rather than direction. It would be beneficial to allow the user to customize the colors in the assembly map, or simply change the forward and reverse indicators to another color (not red yellow or green). | 030411- This was discussed in a phone conference; John suggested that Russ enable the user to select color preference for the assembly map. |

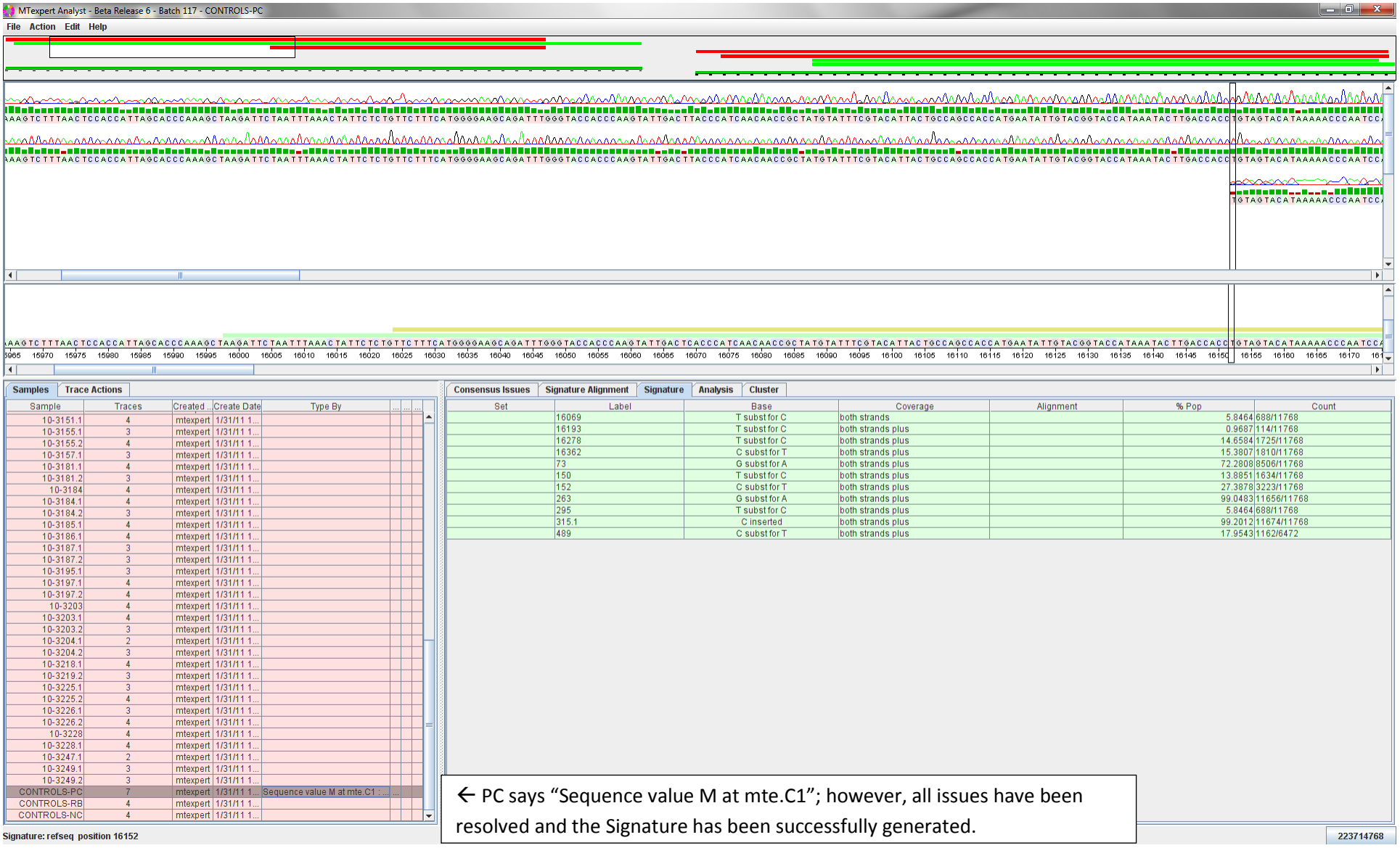
**Screenshot Appendix**

**S1**

←Here, there is only full coverage for HV2; the Assembly Map appears only on the right side of the screen, which sort of implies that the coverage range is only in HV2.

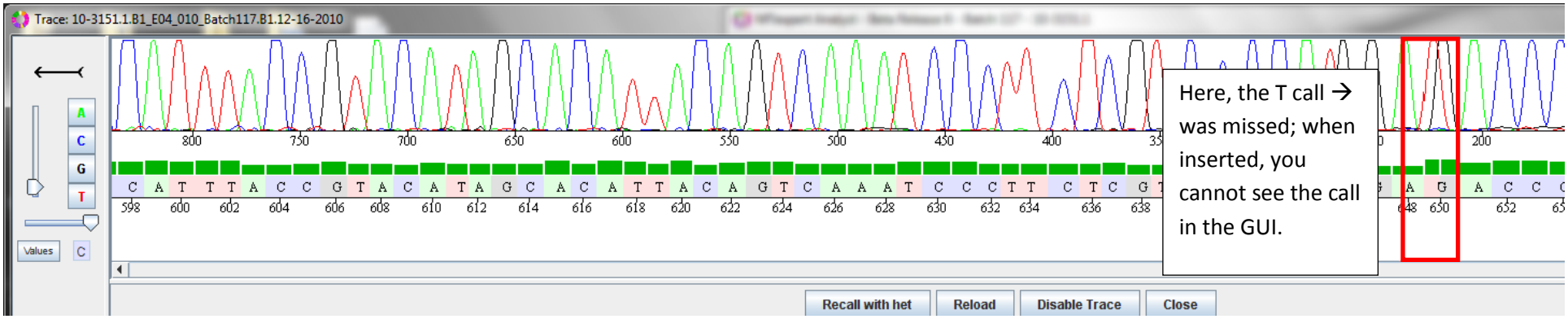
←When there is only full coverage for HV1, the bars are large. It appears as if you have full range coverage.

**S2**




← PC says “Sequence value M at mte.C1”; however, all issues have been resolved and the Signature has been successfully generated.

**S3**



# *Appendix C: MTexpert User Manual*

This page intentionally left blank.

# MTexpert™

## An Automated Software System for mtDNA Data Analysis and Type Generation

Version 2.0
July 2012

The Types generated by this software have been tested and validated.

## Table of Contents

# 1 Running MTexpert Analyst

When a project is created or opened in MTexpert, the software attempts a completely automated analysis of the sample or samples in the project directory. Any problems with control traces and any consensus issues that cannot be automatically resolved will cause the software to stop the automated process for that sample with a message. The MTexpert Analysis user interface allows scientists to review the automatically-generated projects and to correct any issues that prevented the automated software from producing an mtDNA type.

## *1.1 Instrument and Computer System Requirements*

### 1.1.1 Instrument and Chemistry

MTexpert analyzes .ab1 files generated from ABI Genetic Analyzer instruments using either ABI PRISM® dRhodamine Terminator Cycle Sequencing or BigDye® Terminator v1.1 or v3.1 Cycle Sequencing Kit chemistries.

### 1.1.2 Computer

Hardware: MTexpert should be installed on a 2.8G, or faster, dual core processor with a minimum of 2GB of ram. Since the user interface is multi-pane and highly visual, a large format display is recommended (one 1920x1200 display or two displays with at least 1080 resolution).

Software: MTexpert is compatible with Microsoft© Windows XP, Windows Vista and Windows 7; the latter two must be installed in the user data space. MTexpert also runs on Linux and Unix systems having a Java runtime.

## *1.2 Overview of the Process*

Using the software, the routine steps used to analyze a directory of trace files are:

1. Start MTexpert and login.
2. Create or open an MTexpert project in a directory of `.ab1` sequence data files for that sample. When a project is created or opened, MTexpert automatically tries to generate an assembly and produce an mtDNA type for the sample(s) in the directory.
3. Resolve issues with the contig assembly.
4. For each sample, review the Consensus Flags and edit the sequences as necessary to resolve the Consensus Flags.
5. For each completed sample, import a second type generated by an independent analyst
6. Reconcile any differences in the MTexpert type and the independent type and then export the mtDNA type as a CODIS 4.1 XML file.
7. Save and close the project

These steps are discussed in more detail the following sections.

## 1.3 Starting the Software

### 1.3.1 User Login

When the software starts, you are prompted for a user login (Figure 1). The software takes the identification of the user logged into Windows as the default User Name but you can change this to anything. User Names and Passwords are case sensitive


**Figure 1: User login window**

### 1.3.2 User Accounts

When the software is first run after installation, only the "admin" user exists with the password "password". Logging in as the "admin" user opens a User Login Administration window shown in Figure 2. New user names and passwords can be entered in the table in this window and each new user must be assigned one of three permission levels: Administrator, Analyst, and Viewer. This is the sole function of the "admin" user. Closing the window starts the login dialog again.


**Figure 2: User table in the User Login Administration window.**

The first time the software runs with the "admin" user, please change the administrator password. The "admin" user name and Level cannot be changed. If you change it accidentally, it reverts when the software is restarted. The User Login Administration is only accessible immediately after logging into the software as the "admin" user.

### 1.3.3 User Permission Levels

A user can be assigned one of three Permission Levels – Administrator, Analyst, and Viewer. A user's Permission Level determines what functions the user is able to access. For example, only Administrator-level users are allowed to change the Reference Preferences and the Primer Preferences and the paths to the executables in User Paths because these can fundamentally change the behavior of the software and the Mitotypes that are produced. Other than these, all of the software functions are available to a user with Analyst privileges. Viewer-level users are only allowed to review existing projects – they may not make changes and save them or save results from projects.

Access to some of the functions in the MTexpert software that are controlled by user Permission Levels are listed (Table 1). These functions are discussed in more detail in later sections.

**Table 1: Functions controlled by Permission Level**

| | Setting User Preferences | Saving a Type | Setting Reference Preferences | Creating or Saving a Project | Setting Primer Preferences | Defining a User |
|---|---|---|---|---|---|---|
| Administrator | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Analyst | ✓ | ✓ | | ✓ | | |
| Viewer | ✓ | | | | | |

**Note:** For clarity, this manual describes MTexpert as used by an Administrator. In practice, a specific User will only be able to select menu items and operations that are allowed for that Permission Level.

## *1.4 Creating or Opening a Project*

After logging in, the initial window in the MTexpert screen is blank and the only actions possible are to create a new project or open an existing project. MTexpert projects are directory/folder specific - all the files in a directory belong to that directory's project. When a project is opened or created as described below, MTexpert attempts to process all of the .ab1 files in a directory, using the data file name to determine what to do with each file.

**Important Note:** MTexpert has been designed with the assumption that a project directory either contains all the data files for a single sample or all the data files for an entire plate.

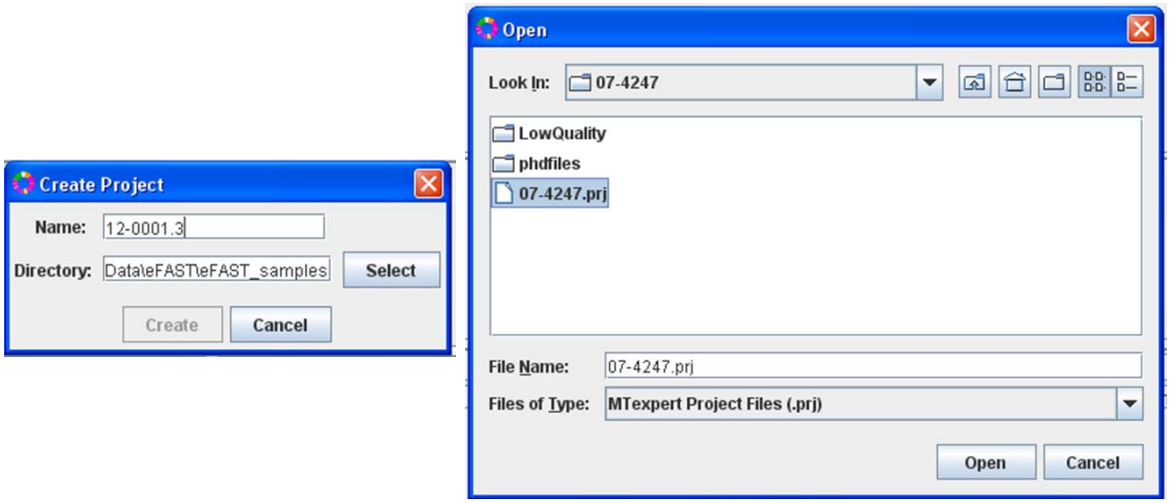



**Figure 3: Project Creation and Project open dialog boxes.** To open or create a new project, navigate to the directory where the plate/sample traces are stored (see Section 1.4.1.1 and 1.4.1.2 for Plate versus Sample mode). Select the directory in order to create a new project; open the directory and select the .prj file to open an existing project. Detailed instructions on this

### 1.4.1 Creating a new project

1.4.1.1 Plate Mode

If the directory contains any control data files at all (see Section 1.4.3, How MTexpert Reads a Trace File Name), the software assumes that it is processing a plate of data and analyzes and enforces all of the control requirements. If any primer's controls fail in a plate project, any trace file using that primer and run on the same plate is not automatically included in any sample's assembly in the same project. The flow diagram in Section 9.1 describes the "plate mode" processing.

To create a new project, select **File→Create New Project**. Navigate to the directory which contains all sample and control traces for the entire plate. Figure 3 shows an example of the dialog box that will appear.
When you select **Create** to close this dialog, the MTexpert software automatically processes the data files in the directory. The data file processing depends on the data file name (see Section 1.4.3, below). In general, the MTexpert process is as follows:

1. Copies the reference sequence information into the project (see Section 6.5);
2. Finds every `.ab1` data file in the directory;
3. If the bases have not been called for a data file, then the TraceTuner base caller software is run and the `.ab1`. The `phd.1` base call file is placed in the "phdfiles" subdirectory of the sample directory;
4. Reads the file name to determine if the trace belongs to a control or a sample:
    a. If the data file is from a negative or reagent blank plate control, process the control;
    b. If the data file is from a sample (or positive control DNA):
        i. Load the base call and sequence data into the project,
        ii. Align the sequences into a consensus sequence,
        iii. Compare the consensus sequence to the rCRS reference sequence
        iv. Generate a list of items in for "Consensus Review", and
        v. Generate a "Signature" mtDNA type.

Section 9, Workflow Diagrams, shows this process in a flow chart. This process may take a few seconds depending on the number of primer files, the ease of assembly, and the speed of your computer.

☐ A new project file will be created in that data directory with a .prj extension only after the new project is saved by the user.

### 1.4.1.2 Sample Mode

If no control data files are found in the project directory, MTexpert simply does its best to load and assemble all the .ab1 files in the directory into a single consensus sequence, and create a signature type for the consensus. In a sample project so every data file is assumed to have good controls. *Thus, when no controls are present, MTexpert assumes that the user is operating in "sample mode", that is, all traces in the folder belong to a single sample.* The MTexpert project data is contained in a .prj file that is saved into the same directory where the project traces are stored. Multiple projects can be stored in a single directory, given that they have unique names.

To create a new project, select **File→Create New Project**. Navigate to the directory that contains all traces for a single sample. Figure 3 shows an example of the dialog box that will appear.

Use the Select button to open a browser to the directory with the .ab1 data files. Type a name for the project, and select the Create button (you may need to hit the "enter" key after typing the name for the project in order to enable the Create button). To make browsing easier, the default root data directory in the Directory box can be set in the User Preferences accessed from Edit → Preferences (Section 6.3, Paths).

When you select **Create** to close this dialog, the MTexpert software automatically processes the data files in the directory. The data file processing depends on the data file name (see Section 1.4.3, below). In general, the MTexpert process is as follows:

5. Copies the reference sequence information into the project (see Section 6.5);
6. Finds every `.ab1` data file in the directory;
7. If the bases have not been called for a data file, then the TraceTuner base caller software is run and the `.ab1`. The `phd.1` base call file is placed in the "phdfiles" subdirectory of the sample directory;
8. For the sample, MTexpert will:
    i. Load the base call and sequence data into the project,
    ii. Align the sequences into a consensus sequence,
    iii. Compare the consensus sequence to the rCRS reference sequence
    iv. Generate a list of items in for "Consensus Review", and
    v. Generate a "Signature" mtDNA type.

Section 9, Workflow Diagrams, shows this process in a flow chart. This process may take a few seconds depending on the number of primer files, the ease of assembly, and the speed of your computer.

☐ A new project file will be created in that data directory with a .prj extension only after the new project is saved by the user.

### 1.4.2 Opening a Project

Once a project has been created in a sample data directory, it can be reopened using **File→Project Open**. A dialog box lets you browse to the .prj MTexpert Analyst project file.

A project file contains all the previous actions, automatic and manual, that have been taken in the project when it was saved.

When the project is opened, MTexpert looks for any new `.ab1` data files in the directory and, if found, calls the bases for the sequence and aligns the new sequence to the consensus sequence in the project.

**Important Note:** When a project is opened, if MTexpert does not find an `.ab1` data file matching a trace that was previously part of the project, that sequence is completely deleted from the project. If you rename a file in the directory that was previously part of a project, then the old trace identified by the old file name, and any edits on it, are removed from the project and the new file name is treated as a new trace. This will discard any edits that were made on that trace. If a missing file is a control then the entire control processing is redone in the project.

### 1.4.3 How MTexpert Reads a Trace File Name

When a project is created or opened, MTexpert looks for loads any `.ab1` trace data files in the sample directory into the project. The software uses the trace file name to determine what to do with each trace. The entire trace file name is shown in the first column of that trace's row in the Trace Action Window, "Trace Name". See section 6.10 on the file Name Parse tool for detailed instructions on how to establish your laboratory's sample and control naming convention.

### 1.4.4 Errors Reading a Data File

If the software has a problem reading the `.ab1` file it displays an error message listing the file name as shown in Figure 4. There is no recovery from this error – the trace file is simply not used in the project. The best way to try to solve the problem is to get another copy of the file from the instrument and try again. This error is displayed each time the project is opened because the MTexpert software detects that an `.ab1` file in the directory is not included and so tries to load it.


**Figure 4: Error message from reading a bad trace file**

### 1.4.5 Multiple Runs and Instruments

The Applied Biosystems Data Collection Software embeds 2 values in each trace file - the "Plate ID" and the "Plate Name". The Plate ID comes from the barcode on the plate or, if it's missing, the ABI software sets it to be the Plate Name that the user inputs when the first load a plate. The MTexpert software reads the Plate ID and Instrument Name from each .ab1 data file. Control Sets (see below) must be from the same Plate

ID and Instrument Name in order to validate the primer from that Plate ID and Instrument Name. Once validated, all the traces for a sample or positive control are assembled into the consensus sequence if possible, regardless of the Plate ID and Instrument Name.

If the user does not enter a Plate Name when loading a plate into the sequencer, a blank "Plate ID" will be used as the identifier for that plate. This would cause some confusion, of course, if the data from two plates were combined into a single directory and neither plate had a Plate Name.

## *1.5 Processing Control Files*

If there are control `.ab1` data files in the project directory, they are processed by MTexpert according to the type of control as described here.

### 1.5.1 Control Sets

In "plate mode," when any control trace is found in the directory, the software analyzes the reagent blank and negative control for each primer as described below. The software only assembles the sample (or control) traces from each primer if the controls for that primer passed the required checks.

Reagent blanks and negative and positive control traces for each primer must be from the same Plate Name in order to validate the primer's sample traces from that run.

### 1.5.2 Reagent Blanks and Negative Controls

MTexpert looks at the base caller results from the `.ab1` file run from the reagent blank. If there are more than one hundred base calls in the untrimmed region of the sequence and the intensity of the signal over the noise is greater than 2.0, the reagent blank is failed. See Section 1.5 for a discussion of the automated trimming processes.

The value for the signal is calculated to be the intensity of the electropherogram at the peak location that was used to make the base call at the position. The value for the noise is calculated to be the sum of the electropherogram values at that base position in the three electropherograms that were not used to call the base.

In some cases you might want to approve a control even though MTexpert failed it - for example if the base sequence in this control is not similar to any sample sequence. In this case, a right click on the red line in the Trace Action Window that signals a failed negative control allows you to "Mark Control Good," overriding the MTexpert failure. You can also override a passing negative or blank control with the "Mark Control Bad". If you have made a change automated acceptance of a control like this, you will need to execute the Process Traces Action in order to propagate the changes to the other samples in the project.

MTexpert also considers each set of RB and NC traces as a unique sample and attempts to assemble all the traces into a consensus sequence and calculate a type. The RB and NC "sample" is indicated by a CONTROLS-RB or CONTROLS-NC line on the Samples table (2.1.5, Samples Window). Usually, of course, there should never be a consensus assembly for the RB or NC "samples" unless there is contamination so the Assembly Window, the Consensus Window, and the Signature Window tabs will usually be empty. Generally, if there is not at least 2 traces with enough signal that overlaps, nothing can be assembled.

### 1.5.3 Control Flags Messages

If the analysis of the control data sets fails, messages appear in dialog boxes. For example the “Project Positive Control Flags” dialog box is shown in Figure 6.

MTexpert expects that each primer used in a sample will have a matching positive and negative control and reagent blank. Messages appear in the Project Trace Warnings dialog box when this is not true. Table 2 describes the cause and result of each possible Project Trace Warning.

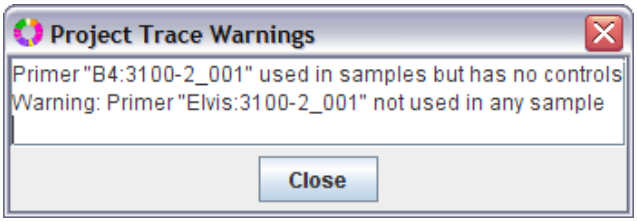



**Figure 5: Project Trace Warnings Dialog Box**

**Table 2 Project Trace Warning Messages**

| Project Trace Warning Message | Cause | Result |
|---|---|---|
| Control "filename" failed for S/N: # | If a Negative Control or Reagent Blank has a signal to noise ratio over 2.0 and more than 100 base calls in the untrimmed region, then the signal is too strong. If the Positive Control has a signal to noise ratio less than 2.0 then the signal is too weak. | The primer controls fails. All sample traces from that primer are marked as unusable in a sample assembly and the control traces are not used in the positive control assembly |
| Trace "filename" has an unknown control type | The control type in the file name is not exactly “RB”, “HL60”, “NC” or “BLANK” | The trace file is loaded into the CONTROLS sample but not used in any evaluation or assembly |
| No loadable trace files found | Only control files were found in the project directory | Software does nothing. Should it assemble the positive controls? |
| Warning: Primer "primer + instrument + run" not used in any sample | Control files were found for a primer but there were no sample files in the project that use the primer | The primer is assembled into the HL60 consensus assembly. |
| Unknown control type in “filename” | The second field of the `.ab1` file name is not exactly “BLANK”, “HL60” or "PC", “NC”, or “RB” so the software does not | The file is loaded as a Control sample trace but not used in any blank analysis or positive control assembly. It is marked in blue |

| Project Trace Warning Message | Cause | Result |
|---|---|---|
| | know what to do with the file. | in the Trace Action table. |
| Primer "primer + instrument + run" used in samples but has no controls | No control files were found for a primer that was used in some sample files | The trace generated with this primer is marked as unusable in any sample assembly since the primer has not been validated by controls |
| Primer "primer + instrument + run" failed because No Reagent Blank and/or No Positive control and/or No Negative control found | One or more of the set of three control files needed for each primer was not found. | The primer controls fail. All sample traces from that primer are marked as unusable in a sample assembly and the control traces are not used in the positive control assembly |
| Primer "primer + instrument + run" failed because Reagent Blank failed and/or Positive Control failed and/or Negative Control failed. | One or more of the set of three control files needed for each primer failed | The primer controls fail. All sample traces from that primer are marked as unusable in a sample assembly and the control traces are not used in the positive control assembly |

### 1.5.4 Positive Controls

MTexpert first checks the signal level in each of the base-called Positive Control data files. If the signal to noise is less than 2.0 in the untrimmed region of any trace, that positive control trace is failed without further processing because the signal is too weak.

Next MTexpert assembles all the Positive Control traces into the consensus sequence, verifies forward and reverse strand coverage, and calculates a type. If the type is not the type expected for HL60, then the positive control fails. If the positive control consensus sequence does not span the reference region with at least forward and reverse trace double coverage, then an error message is generated.


**Figure 6 Message box reporting problems with the positive control**

Each time a project is created or opened, an error message box is generated reporting any positive control issues. Figure 6 shows a message box reporting a problem with the length of the consensus sequence. All of the possible Project Control Issues messages are listed in Table 3 below with a description of the error condition that leads to it and what the software does as a result. As appropriate, the file name or the primer + instrument + run are identified

**Table 3: Project Positive Control Issues Messages**

| Project Positive Control Issues Message | Cause | Result |
|---|---|---|
| Positive controls failed to assemble | Could not assemble the positive controls into a consensus sequence. This often occurs if enough individual control primers fail that there are not enough left for a successful assembly | No Positive Control assembly is present in the view of the Controls sample. The other samples in the project are processed |
| Positive controls signature length failed to match HL60 | The Control consensus sequence results in a type with more or less than 8 elements so it cannot match the HL60 type | The Positive Control assembly is present and editable. The user should try to edit any trace problems that may create an erroneous type. The other samples in the project are processed |
| Positive controls signature failed to match HL60 | The type for the Control consensus sequence does not match the HL60 type | The Positive Control assembly is present and editable. The user should try to edit any trace problems that may create an erroneous type. The other samples in the project are processed |
| Assembly does not include entire reference region location :[x,y] is uncovered | The positive control consensus sequence does not span the reference range | The Positive Control assembly is present and editable. *The software takes no action on samples in the project directory despite the failure of the positive control.* The user may be able to edit the traces (for example by overriding some automated trimming) to fix the problem |
| Low coverage (single strand) from x to y | The consensus sequence is not assembled from at least two traces in the specified base range | The Positive Control assembly is present and editable. *The software takes no action on samples in the project directory despite the failure of the positive control.* The user may be able to edit the traces to fix the problem |

## *1.6 Trimming and Assembling the Traces*

### 1.6.1 Automated Trimming

Trimming removes undesirable data at either end of a trace. A trimmed region must include at least one end of the trace. Automated trimming occurs at several points in

the process of loading and processing a project: first by the basecalling TraceTuner software, then by the assembler, and then by MTexpert in an automated procedure.

In the first trimming step, the TraceTuner basecaller software automatically trims the base calls at the ends of the electropherogram until the running average of the quality of the base calls is above 20. The TraceTuner trimmed regions are shown in the Trace Window with almost colorless base call backgrounds and with a double red bar drawn through the base. Figure 7 below shows the display in a Trace Window of the edge of a TraceTuner trimmed region. (The double red line indicates the trimmed region of the trace that is not sent to PGA for assembly, which could be a result of either manual trimming or TraceTuner trimming).

After base calling and trimming in TraceTuner, the remaining trace is passed to the Paracel Genome Assembler (PGA) for assembly.

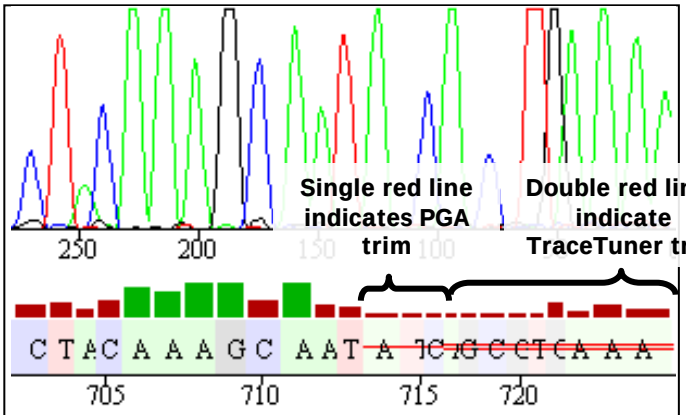



**Figure 7. Trace and base calls showing TraceTuner or manual trimming (double red bar) and PGA trimming (single red bar) over the light-color base calls**

PGA also can also automatically trim each trace in the process of assembling it into the consensus sequence. PGA uses quality values and alignment-disagreement to determine where to trim poor data from the start and end of each trace. The PGA trimmed regions are shown in the Trace Window with a single red bar drawn through the light-background base windows. Figure 7 shows the display in the trace window with three PGA trimmed bases to the left of a string of bases trimmed by TraceTuner.

Finally, MTexpert uses an automated trimming expert rule to further trim problematic data from the start and end of the trace using information from the heteroplasmic base calls. TraceTuner's sensitive detection threshold (1.6, Automated Point-Heteroplasmy Base Calls) results in a lot of spurious heteroplasmy calls in lower-quality areas of the trace. When this trimming rule is enabled, as each trace is loaded it is trimmed from each end until the window of 20 bases in front of each base being evaluated contains fewer than 6 heteroplasmic base calls. This rule can be enabled or disabled, and the parameters adjusted, in the MTexpert Preferences Settings (Section 6.8). Regions trimmed by this expert rule are also shown by the double line in the Trace Window as shown in Figure 7.

☐ MTexpert only runs this expert trimming rule when a trace is loaded. Enabling or disabling the rule or changing the parameters requires you to create a new project in the directory in order to reload all the traces.

Only the untrimmed bases in each trace are shown in the MTexpert Assembly Window.
.

### 1.6.2 Assembly

The PGA software uses the base quality values generated by Trace Tuner to optimize the assembly and to assign a quality to each base in the assembly. When scoring an alignment, matching bases with high quality values makes a large positive contribution to the score and differences at bases with high quality values makes a large negative contribution to the score. Matches and differences at bases with low quality values contribute smaller positive and negative scores. In this way, using base quality values allows the alignment to weight the better quality regions in each trace while still considering supporting evidence from lower quality regions.

The quality-driven automated trimming process often retains some relatively poor looking data at the ends of a trace when this data is making a positive contribution to the alignment process. Parameters in the assembler provide some control over the degree to which low-quality bases are considered to be contributing to the alignment. However the nature of the quality driven alignment means that there is no harm in keeping low-quality data that contributes to the alignment score even if it looks pretty bad.

### 1.6.3 Manual Trimming

Section 3 describes editing a trace with the Trace Window. Manual trimming is performed in this window. A range of bases can be selected either by dragging with the mouse or using the ctrl-[ (to select all bases to the left of the highlighted base)or ctrl-] (to select all bases to the right of the highlighted base) from an active base and then trimmed with ctrl-t. The traces in the Trace Window show all trimmed regions. Regions marked with a double red line are never sent to the assembler. Regions marked by a single red line are the result of trimming during assembly. Note: using delete or backspace does not trim the bases; this removes them entirely from the trace viewer.

Manual trimming overrides any automated trimming in that end of the trace that has been performed on the trace by both TraceTuner and, initially, by PGA. When a region of the trace is trimmed manually in the Trace Window with the ctrl-t key, the remaining untrimmed region is sent to PGA for reassembly into the consensus. PGA may make additional automated trims on the trace in the reassembly process which will be shown as single red lines in the Trace Window after reassembly is complete.

## *1.7 Automated Point-Heteroplasmy Base Calls*

By default, MTexpert runs the TraceTuner base calling software with heteroplasmy calling enabled. TraceTuner calls a base heteroplasmy if there is a second electropherogram peak at a base location that is 10% or more as high as the primary peak at the location. The standard IUPAC symbols are used for the heteroplasmies and displayed in the trace base calls and assembled into the consensus sequence.

The 10% threshold is a very sensitive test for heteroplasmies and as a result noise peaks in the trace or low levels of contamination in the sample can result in spurious heteroplasmy calls. Spurious heteroplasmy calls can also occur on long C stretches. The Trace Properties information window (Figure 17) indicates how many heteroplasmy calls appear in the trace or otherwise indicates that base calling on this trace was performed with heteroplasmy detection disabled.

Heteroplasmy calling can be disabled with the button in the Trace Window "Recall w/o het". Section 2.3 Trace Windows describes this display. Note that using this button to switch between calling bases with or without heteroplasmies removes any of the manual edits that may have been performed on the trace.

A string of heteroplasmies can also indicate a heteroplasmic insertion or deletion in the sequence. In this case the trace looks like a mixture of mtDNAs on one side or the other of the location of the indel.

When a heteroplasmy base call in a trace is not included in the consensus assembly, that base call is listed in the Consensus Flags Window (Figure 20) as an "Unused het call in trace". Manual review of each unassembled heteroplasmy call allows the analyst to rename the base as needed to eliminate the Consensus Issue. A heteroplasmy call in the sample assembly also prevents the project from completing the type automatically, see Section 1.10.

Any multiple calls at a base location are passed to the assembly program. The current parameters for the assembly program allow it create a mixed (heteroplasmic) base in the consensus sequence when 2 or more bases have a mixed call at that location that has a lower probability than the highest probability base call as long as both have a "reasonable" quality (at least 50% of the highest). So, if there are 2 traces aligned and each call an "A" with a lower quality "M" that is at least 50% of the A quality the assembly will contain the "M".

## *1.8 Closing and Saving a Project*

Projects can be closed, saved, or saved with a different name with dialogs accessed under the File pull down menu. Figure 8 shows these project management options.

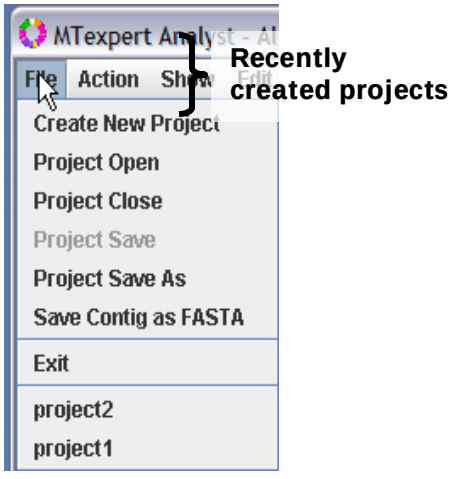



**Figure 8: Options under the File pull down menu**

A project is also closed when the MTexpert Analyst software is closed, either by using File→Exit or by closing the window. Using the Exit command saves the project in the "recent projects" list that will appear in this pull-down menu below the Exit command.

If you attempt to use the Windows to exit the software or File→Exit to close a project or load a new project and the current project contains unsaved changes, a warning

dialog box will appear providing the option to save the changes before proceeding.

## *1.9 Saving the Consensus Sequence*

File → Save Contig as FASTA allows you to save the assembled consensus sequence in a text file in the FAST-A format. Selecting this option opens a browser window that lets you enter the new file name, etc.

## *1.10 Automated Edits*

The MTexpert software attempts a complete analysis of the data for each sample in the project when a project is created or opened.

Some of the routine resolution of consensus issues can be automated with expert rules that run after an assembly has been created but before the type is generated. These rules are run automatically when a project is opened or created unless a problem with the controls or assembly halts the automated attempt to create a type. When an automated rule edits a trace, a line in the Trace Action result records the action with the user "AutoEdit".

In order to review and validate the expert rules, they can be run from the MTexpert Analyst software. The Action → AutoEdit Assembly command runs all of the expert rules in sequence. Individual rules can be disabled for a user with the user preferences, see Section 6.7.

Running the AutoEdit rules can take some computation time since MTexpert reassembles the consensus sequence each time a rule is executed.

### 1.10.1 Assembly Gap Removal

The first expert editing rule automates the process of removing insertions created by the assembler when the base caller creates a spurious extra base. These insertions result from an extra low-quality base call in a fragment. The rule is run on every insertion - which would be marked by a "-" in the Consensus Flags list for the contig being displayed. The logic for the rule run on each insertion gap in the consensus is:

If there a single base in one fragment that is causing the insertion in the consensus then:

If the extra base is in a string of bases of the same type
Then delete the lowest quality base in the string
Otherwise delete the single base causing the insertion

We have found that this rule resolves most of the insertions in the consensus created by extra base calls from TraceTuner. However this is a little simplistic so the user should perform additional checks prior to automatically deleting a base in a trace.

### 1.10.2 Automated Primer Edits

In many cases the traces and resulting base calls in the first 5 to 10 bases after specific primers are consistently distorted. MTexpert has incorporated automated edits that look for specific, known distortions and replaces them with the appropriate experimental sequence. These automated edits will only be attempted on the base calls from traces identified as being from the appropriate primers so the file names need to be set up correctly (see Section How MTexpert Reads a Trace File Name 1.3.5). The distortions are also dependent on the dye set used so the software checks the AB1 file for BigDye in the appropriate field entered by the user when setting up the sample list.

### 1.10.3 Other Rules

Any consensus issue that stops the automated processing and that is routinely solved with a simple procedure might be automatable. MitoTech is open to discussion of such issues; contact John Elling for more information.

## *1.11 Sample Assembly Issues*

The MTexpert attempts a completely automated analysis of all of the samples in a directory. If successful, the type is generated by the MTexpert software automatically. If there is an issue with a trace or with the sample sequence assembly that is not normal, it can be trapped with a message and automated analysis stopped. This will require manual review, potentially editing, and validation before the type can be generated. The situations that are currently trapped are described in Table 4.

**Table 4: Sample Assembly Issues Messages**

| Project Sample Assembly Issues Message | Cause | Result |
|---|---|---|
| Sample *name* did not complete for the following reasons: Sequence Value *X* at *trace:position* | If there is a heteroplasmic base call in a trace being assembled, the processing requires manual intervention. | Either correct the base call or decide that the heteroplasmic call is appropriate. The software will never type a sample sequences with a heteroplasmic automatically because we decided these cases should require manual validation |
| Sample *name* did not complete for the following reasons: Sequence Value *N* at *trace:position* 0 | If there is an ambiguous base call, like "N", the system will not generate an automated type | Correct the base call. The software will never type a sample sequences with an ambiguous base automatically because we decided these cases should require manual validation |
| ? | Suggest other situations in trace or sample sequences that should require manual review, interrupting the automated analysis | |

If a sample fails automated analysis, the cause of the failure is saved instead of the signature, and the row in the sample table is shaded red. If you hover the pointer over the "Saved By" column in the Samples window, the text of the failure will appear in a tooltip popup.

After the samples have been processed a dialog with the samples that failed to complete automatically will appear with the sample names and causes (Figure 9). If the project is saved and restarted, the message will be repeated to warn the user that there are unresolved sample problems in the project.

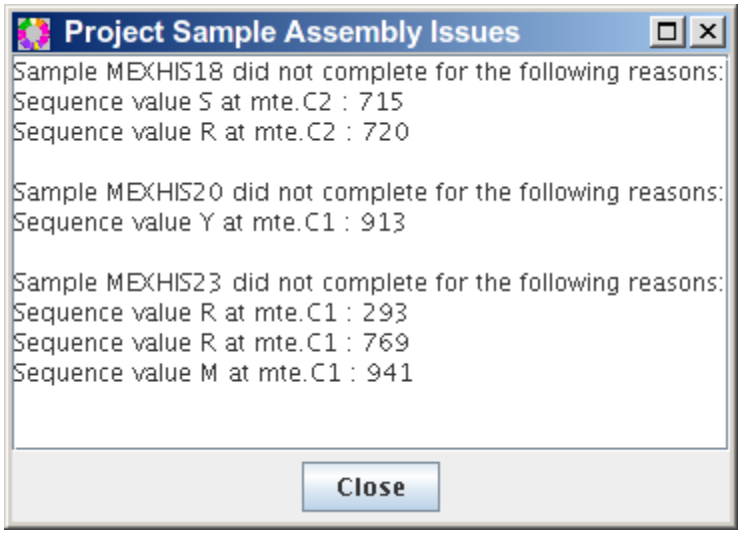



**Figure 9. Sample Assembly Message Box example**

In order to address these problems, the user can select the appropriate sample by selecting the sample's row in the sample table and then exploring the Consensus Flags window. After applying appropriate edits the Action->Process Traces can be performed and any unsaved samples will be assembled and evaluated for any remaining problems.

## *1.12 Signature Generation Issues*

Occasionally the MitoTyper type generation software needs to pass a message to the user about a condition that occurs when creating the signature. These messages are only produced the first time the condition occurs in the project (usually when the signature is generated for the first time for the assembled consensus sequence) and the first time the condition disappears. The condition might disappear, for example, if a base is edited. The Signature Generation Issues messages are repeated if the project is opened or if a project archive is opened.

### 1.12.1 Multiple Types

The MitoTyper Rules for describing the sample mtDNA sequence may treat polymorphisms in the HV2 300 to 315 region differently than the rest of the sequence. In this region the software has a preference for aligning the 310 Ts, the 300 to 302 poly-A region and the 303 to 309 poly- C region in the rCRS with matching base and regions in the sample sequence. Sometimes following these preferences results in a type

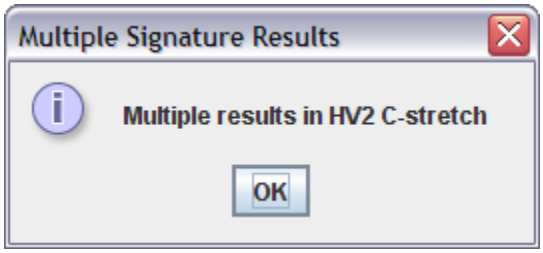



**Figure 10 Message box generated when special handing of the HV2 C-stretch generates a unique type**

for the region that is different than the type that would be generated with the regular Mitotyper rules that are applied to the rest of the sequence.

If these preferences result in a type for the 300 to 315 region that is significantly different than the type generated by a normal treatment for the sequence in this region, the software informs the user with a message shown in Figure 10. The reporting threshold for a significant difference is set in the Reference user preferences (see Section 6.5).

The MTexpert software allows you to review the different types. If the HV2 preferences in the MitoTyper rules are used to create a preferred type for the region and the message shown in Figure 10 is generated, those type entries in the Signature table that are different between the regular and preferred types are marked with a number in the first "Set" column indicating if that type was generated by the normal MitoTyper rules (number 1) or the alternate HV2 C-Stretch typing rules (set 2). You can toggle between the two types with the "Select next set" option accessed by right clicking on the numbered line in the Signature table.

When multiple signatures are generated for a sample using the HV2 preferences instead of the regular MitoTyper rules, it is worth a review by an analyst to confirm that the special handling is based on a valid consensus and the preferred type is, in fact, preferred.

### 1.12.2 Untyped Region

Occasionally the MitoTyper software cannot generate a type description for a polymorphic region. When this occurs a message box shown in Figure 11 is displayed. The Signature Window displays a solid navy blue bar between the consensus sequence and the reference sequence in the untyped region.


**Figure 11 Warning of an untyped region of the consensus**

Often the MitoTyper software can successfully return a type for an untyped region if it runs longer. Increasing the Signature Run Time in the Reference Preferences (Section 6.5) and rerunning the MitoTyper code with the Process Traces action (Section 5) might fix the problem.

### 1.12.3 Aligning Polymorphisms at the End of Contigs

A Mitotyper-compliant alignment cannot be calculated for polymorphism at or near the ends of a consensus sequence. Normally MTexpert defaults to an arbitrary alignment. The arbitrary alignment and resulting type description is obtained by forcing a one-to-one sequential base match between the reference and sample

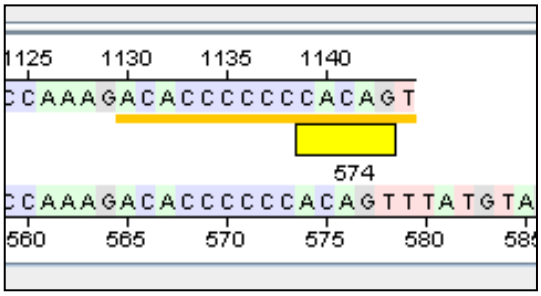



**Figure 12.** An example display in the Signature Window of the end of the sample consensus in which polymorphism are arbitrarily aligned base-by-base with the reference sequence

sequence bases at the end of the sample's sequence. In the Signature Window, the region of arbitrary, base-by-base alignment is indicated by a yellow bar below the end consensus sequence as shown in Figure 12. Figure 12 also shows that in this example, five substitutions reported from position 574 to 578 are used to describe the difference from reference at this end of the sample sequence.

Specifying a reference base as the Complete Start or Complete End forces the software to align the last base in the consensus to this position on the reference.

In the example shown in Figure 12, the Arbitrary alignment can be improved by aligning the last base in the sequence to the reference base at position 578 rather than at 579. To force this as a Complete Alignment the analyst would enter 578 for Complete End (a number must also be entered for the Complete Start). Selecting the Complete Alignment option and then selecting the Recall Type button will rerun the trace processing and force a MitoTyper alignment all the way to the ends of the sample sequence. The result, shown in Figure 13, shows that the MitoTyper rules place an insertion after position 573 to achieve the forensic alignment.

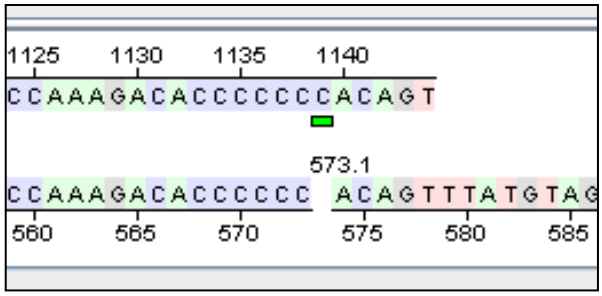



**Figure 13.** An example display in the Signature Window of the end of the sample sequence aligned according to the MitoTyper rules to the region specified by the analyst.

The arbitrary alignment of the ends of the sample sequence may not be an issue in generating the Mitotype if, for example, the sample sequence is outside the region defined in the user preferences in which the software will report a signature. This problem can also be solved by trimming if the ending bases are uninteresting or are of poor quality.

The Rough Alignment is the third end treatment option on the Signature Alignment tab of the Signature. Before creating a forensic alignment with the MitoTyper Rules, the MTexpert software uses a simplified Smith-Waterman algorithm to find a rough alignment of the sample sequence with the reference. The Smith Waterman algorithm looks for the longest subsequences in the sample and reference that have the greatest similarity and will leave off polymorphisms near the ends if adding them reduces the overall similarity score. When the Rough Alignment option is selected (and the traces are reprocessed with the Action -> Process Traces command), the software performs a a complete forensic alignment on whatever subsequences were rough aligned and ignores the end sequences that the preceding Smith-Waterman algorithm left off. Selecting the Rough Alignment uses the rough alignment as the cutoff for the MitoTyper Rules. This option is effectively equivalent to the Arbitrary alignment but typically ignores polymorphisms at the end rather than arbitrarily aligning them.

## 1.13 Reviewing, Comparing, and Approving a Type and Generating a Report

When a project is created or opened, MTexpert automatically runs the entire process described at top level in Section 9.1. If the automated process halts before a type is generated, the analyst can fix the problem with the controls and/or edit the traces to resolve the issues and then restart the automated workflow with the Action → Process Traces function.

When the automated process is successfully completed, the analyst should review (and resolve if necessary) any remaining Consensus Flags, validating the type that is produced.

Once MTexpert has generated a good signature for the sample, the analyst can compare this type to a type for the sample that has been generated by a second analyst. To do this, the analyst must load the independently-produced type using the Action → Load Compare Type. Use the Select button to open a file browser and Load and Save a type file into the project. When loading an independent type for comparison, MTexpert can parse a text-format type file generated by Sequencher or by MitoTyper. Simply use the file browser to find and open the correct .TXT file.

Use the Action → Compare Results with Saved to generate the Signature Difference Report shown in Figure 28. This function compares the loaded compare type with the current signature. In the Signature Difference Report, a < or > indicates an insertion in the saved signature or an insertion in the current signature respectively. An "=" indicates that the positions of the elements is the same in both signatures and a "|" indicates that the positions are different.

If there are no differences in the type signatures generated by the two analysts (or if the differences are not important), the analyst can create a report Action → Create Report, generating either an output text file or a CODIS 4.1 XML format file. Selecting this option opens a file browser dialog box that is used to name and save the output report in either the `.txt` or `.xml` format. The text file contains every line in the Trace Action Window and every line in the Signature Window. The XML report creates a file in the CODIS 4.1 format. When generating a CODIS report, a dialog box appears that allows the user to enter additional text for various report fields that cannot otherwise be determined by the software.

Once a report is generated for a sample the sample table row is shaded green.

The only part of this two-analyst process that is currently enforced by the software is the requirement that a second type be loaded before the software enables the Compare Results with Saved and the Create Report actions. Note that a report can be created even if there are differences with the imported and saved type.

# 2 Navigating MTexpert

## 2.1 The MTexpert Display

After creating or opening a project, the MTexpert display gives an overview of all the information in the project.

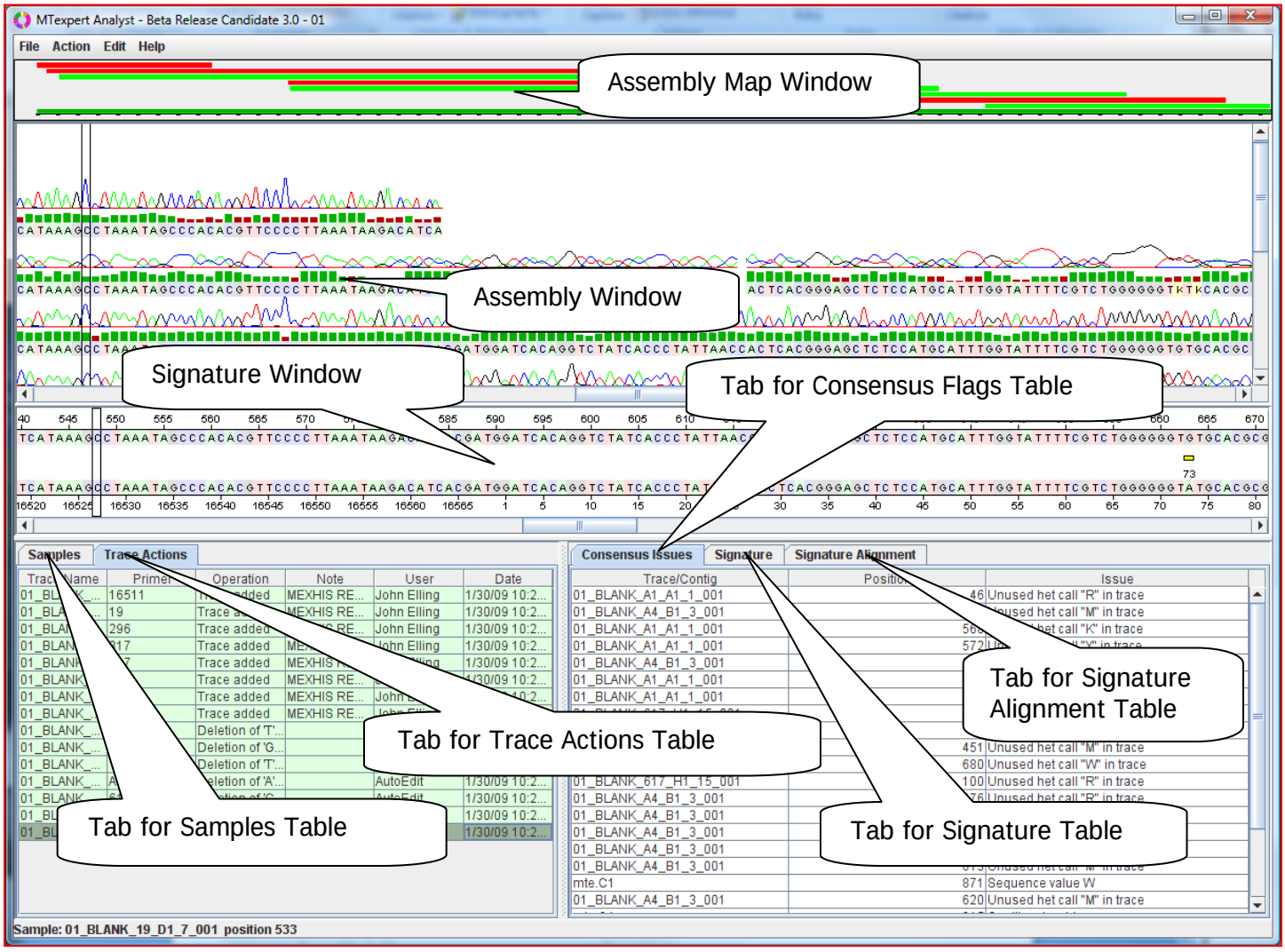



**Figure 14 Example initial display of MTexpert after creating a project, with windows labeled**

An example is shown in Figure 14. There are five windows in the MTexpert display.

- The Assembly Map Window provides a graphical display of all the traces assembled into the displayed contig as well as a “you are here” box showing which region of the traces and contig is currently displayed in the Assembly Window below.
- The Assembly Window displays trace, basecalls, assembly and consensus information for the subregion of the consensus.

- The Signature Window hat three tabs that contain information on the consensus assembly and the MitoTyper signature that has been generated.
- The Trace Action Tab displays a list containing the history of actions performs and provides undo capabilities for some of those actions.
- The Samples Tab displays a list of the samples and controls in the project and allows each to be selected for display and editing the other windows
- Insert description of new tab

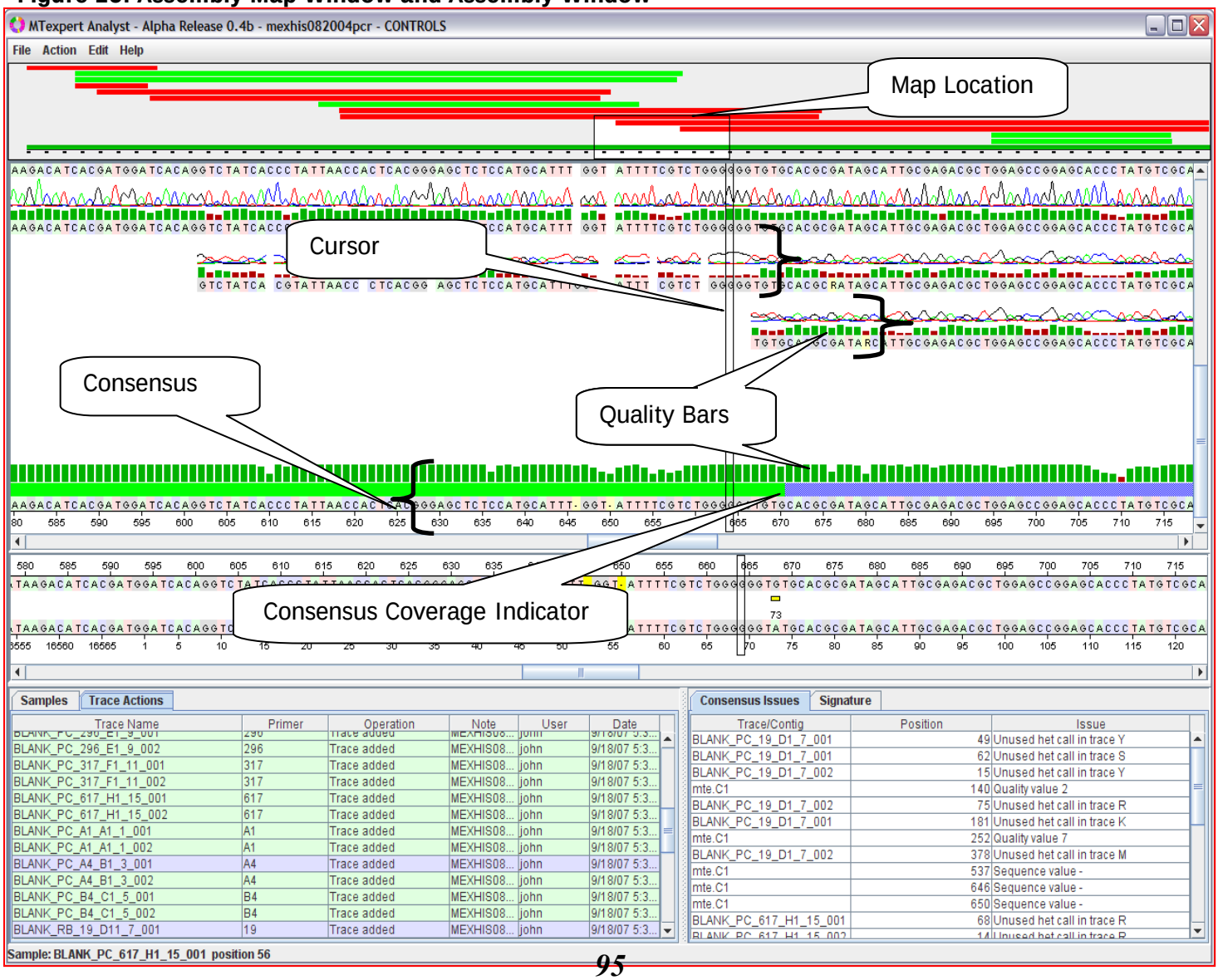
All the windows except the Assembly Map Window can be resized by dragging the window border. The cursor changes to the appropriate double headed arrow, ↕ or ↔, when it is positioned over a draggable window border.

The following sections describe the contents and operation of the windows in the MTexpert display in more detail.

### 2.1.1 Assembly Map Window

Figure 15 shows the Assembly Map Window and its relationship to the Assembly Window.

**Figure 15: Assembly Map Window and Assembly Window**



The white outlined box in the narrow top Assembly Map Window shows the area of the entire map that is currently displayed in the Assembly Window below. By default, the reference range is 15961 to 780 in the mitochondrial genome circle (the reference range is a preference on the Reference tab in the MTexpert Preferences Settings).

The assembled traces in the project are shown as horizontal lines in the Assembly Map, with color indicating direction. Forward traces are green, reverse traces are red.

Clicking in the Assembly Map Window centers the Assembly Window on that location in reference range. In addition the white outline box may be 'dragged' to a region in the map – press and hold the left mouse button on the box and 'drag' to change the assembly view.

Quality bars can be displayed for each trace and for the consensus assembly the Assembly Window. The quality values run from 0 to 40 for traces and from 0 to 100 for the consensus assembly.

## 2.1.2 Assembly Window

The Assembly Window shows the traces loaded in the project as they are assembled into the consensus sequence, and also shows the consensus sequence. Figure 15 labels some of the components of the Assembly Window display. There is a scroll bar at the bottom that can be used to position the view of this window in the assembly reference region.

The consensus sequence is located at the bottom of the Assembly Window and Assembly Map Window. In some cases it may not be visible unless the Assembly Window is showing the bottom of the Assembly Map. Bases are numbered below the consensus in order of their position in the consensus sequence.

The Coverage color bar over the consensus sequence in the Assembly Window provides a rapid visual indication of the coverage in the consensus (Figure 16).

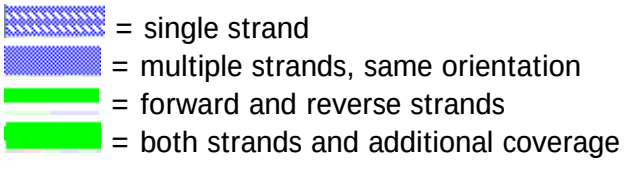



**Figure 16: Key to the consensus sequence visual cues**

### *2.1.2.1 Changing the Display in the Assembly Window*

The display in this Assembly Window can be changed to include the electropherogram traces and the quality value bar graphs. The User Preferences tab accessed from Edit → Preferences pull down menu has buttons to toggle the display of the electropherogram trace and the quality value bar graph. The User Preferences also has

an option to change the font size of the base characters in the Assembly and Signature Windows. The base character font size is the fundamental limitation on how much of the contig can be fit in the Assembly Window, lowering the font size gets a larger portion of the control region into the Assembly Map Window. See Section 6.1 for additional information.

#### *2.1.2.2 Cursors in the Assembly Window*

Moving the pointer in the Assembly Window and the Signature Window changes the highlighted bases in the traces, consensus sequence, and reference sequence. The highlighted base position is indicated by the vertical box around the bases. Double clicking opens all the Trace Edit Windows (see below) that cover that base location in the consensus sequence.

### 2.1.3 Signature Window

The Signature Window below the Assembly Window compares the consensus sequence and the reference sequence. The consensus sequence is displayed on top and the reference sequence is displayed on the bottom. Any bases that differ between the consensus sequence and the reference sequence are highlighted between the two sequences in this window.

The reference sequence is read from a file specified on the Reference tab of the Preference Settings.

### 2.1.4 Trace Actions Window

When a Trace is edited, the action is recorded in a line in the Trace Action Window. The Trace Name, the Operation performed, an optional Note, the Date, and the User that performed the action are recorded in each line in the table. In the Trace added operation, the Note field is set to be the sample id that is taken from the trace. The user can add text to the Note field of any trace edit operation in the table by double-clicking the box to get a typing prompt. This field could be used to manually record an explanation for a trace edit, for example.

Automated trace edits which resolve routine consensus issues are recorded here as well, with AutoEdit noted as the user performing the actions.

#### 2.1.4.1 Color Coding of Trace Action Records

Records in the Trace Action Table are color coded.

- A green line in the table indicates an action on that trace indicates that the trace has been successfully included in the displayed assembly.
- A blue line indicates a control trace that is not included in the displayed consensus sequence assembly. The trace can be opened in a Trace Window. When MTexpert is showing the positive control assembly, all the negative control and reagent blank data in the project directory are also shown in the Trace Window as blue lines in the table.

- A yellow line means that the trace is not included in the displayed assembly even though the software thinks it should be according to the file name. The most common reason for a trace to be excluded from the assembly is if there are not enough bases left after quality-based automated trimming.
- A red line indicates that the trace has been "Marked as Unusable" by the user or the software and so is not included in the assembly. The most common reason that MTexpert automatically makes a trace unusable is if the controls for that primer failed the control checks in the active project.

### *2.1.4.2 Sorting the Trace Action Records*

By default, the Trace Actions table is sorted by date and time, so the last Trace Actions in the project appears at the bottom of the list. Clicking the header of any column of the Trace Actions table sorts the list alphabetically in that column – except for the Date column which sorts by ascending or descending date and time. You can use this feature to sort by Trace Name so that all the edits to a trace are grouped together in the table.

### *2.1.4.3 Displaying and Undoing the Trace Actions*

Right clicking any record in the Trace Action Window opens a menu of the operations that can be performed on that record through the Trace Action Window:

- Display Trace displays the trace in a Trace Window with the cursor positioned at (or near) the location in the trace at which the action occurred.
- Mark Trace Unusable removes the trace from the assembly and flags it so that it is ignored in any subsequent assembly attempts. The trace and the trace edits are left in the Trace History with a reddish background. This process can be reversed with the Mark Trace Useable option which allows the assembler to include the trace in the consensus with all of the edit.
- Remove This Trace Edit, enabled only for the most recent Trace Edit in the Action list, undoes the selected action. In order to undo a Trace Edit that was performed before the last edit, you must individually undo all the subsequent edits starting at the last with a Remove This Trace Edit command. This command is equivalent to using Ctrl-Z in the Trace Window.
- Remove All Trace Edits undoes every trace edit performed in this project.
- Trace Properties. See Section 2.1.4.4 below

☐ Since a mistake with the Remove commands can lose a lot of work in the project, selecting the command opens a dialog box that warns of the implications and requires confirmation.

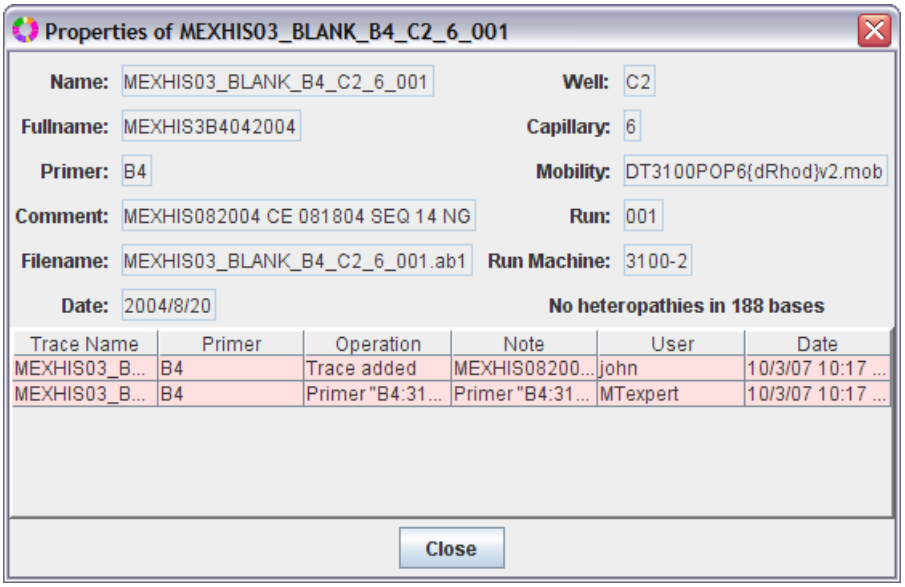



**Figure 17 Trace Properties information box.**

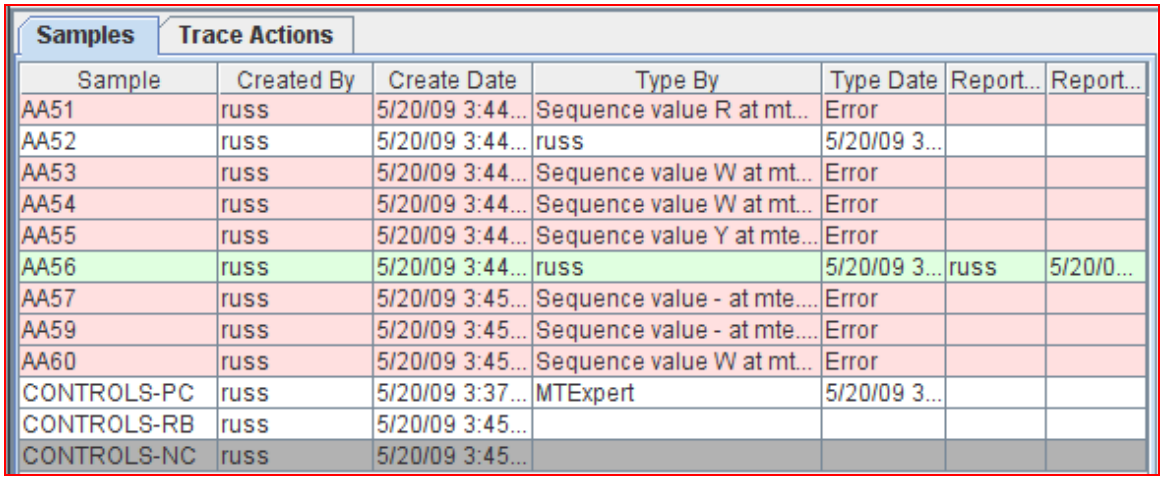
#### *2.1.4.4 Information about a Trace*

Selecting Trace Properties from the menu that appears from a right click of any trace line in the Trace Actions Window opens a handy reference window, Figure 17, showing the information about the data file that MTexpert gleaned from the `.ab1` data file name and information within the data file.

### 2.1.5 Samples Window

When a project is created or opened in a directory, MTexpert associates each `.ab1` file in the directory with a sample or control using the file name (see Section 1.3.5, How MTexpert Reads a Trace File Name). The data files for each sample are assembled into a consensus sequence for that sample and the sample name is added to the list in the Samples Window shown in Figure 18. Figure 18 shows that data for 10 samples and controls were found in the project directory. Selecting a sample in this table opens the data and results for that sample in the rest of the MTexpert Display Window and the Trace Actions Window, showing the traces and the assembled consensus sequence, and the control issues and signature for that sample. Each set of controls is treated as a unique sample. Of course there should not be any signal in the RB and NC traces, especially after trimming, so it is unlikely that there will be anything assembled in teh Assembly window. The CONTROLS sample is selected in Figure 18, allowing the analyst to see the positive control assembly in the Assembly Windows and Signature and Consensus Issue Windows and the trace actions associated with all the control files in the Trace Actions Window.

The table in the Samples Window also contains information about the status of Type approval process including when a Compare Set is saved and when a report is generated and by whom. A red background for the sample indicates that automatic processing did not successfully complete (The Type By will contain information on the errors that were encountered). A white background indicates that processing was successful - either manually or automatically - without any fatal consensus issues remaining. A green background indicates that an analyst generated a report for that sample.



Samples | Trace Actions

| Sample | Created By | Create Date | Type By | Type Date | Report... | Report... |
|---|---|---|---|---|---|---|
| AA51 | russ | 5/20/09 3:44... | Sequence value R at mt... | Error | | |
| AA52 | russ | 5/20/09 3:44... | russ | 5/20/09 3... | | |
| AA53 | russ | 5/20/09 3:44... | Sequence value W at mt... | Error | | |
| AA54 | russ | 5/20/09 3:44... | Sequence value W at mt... | Error | | |
| AA55 | russ | 5/20/09 3:44... | Sequence value Y at mte... | Error | | |
| AA56 | russ | 5/20/09 3:44... | russ | 5/20/09 3... | russ | 5/20/0... |
| AA57 | russ | 5/20/09 3:45... | Sequence value - at mte.... | Error | | |
| AA59 | russ | 5/20/09 3:45... | Sequence value - at mte.... | Error | | |
| AA60 | russ | 5/20/09 3:45... | Sequence value W at mt... | Error | | |
| CONTROLS-PC | russ | 5/20/09 3:37... | MTExpert | 5/20/09 3... | | |
| CONTROLS-RB | russ | 5/20/09 3:45... | | | | |
| CONTROLS-NC | russ | 5/20/09 3:45... | | | | |

**Figure 18. The Samples Window showing a list of 10 samples and control data found in this project.**

## 2.1.6 Multiple Contigs Display

When a sample's or positive control's traces do not assemble into a single contig, the MTexpert software attempts to display all of the consensus contigs in rough alignment with the reference sequence. Figure 19 illustrates the Assembly Map Window, the Assembly Window, and the Signature Windows displaying two un-joined contigs assembled from that samples' traces. In Figure 19 the visible region in the Assembly Map shows the gap between the contigs and the two assembled contigs are shown at the bottom of Assembly Window. In the Signature Window, the contigs are aligned as well as possible with the reference sequence and the differences are shown. Keep in mind that while the type generation rules are doing the best they can typing within a consensus fragment, there cannot be any valid type at the ends of the fragment or in the gaps between fragments.

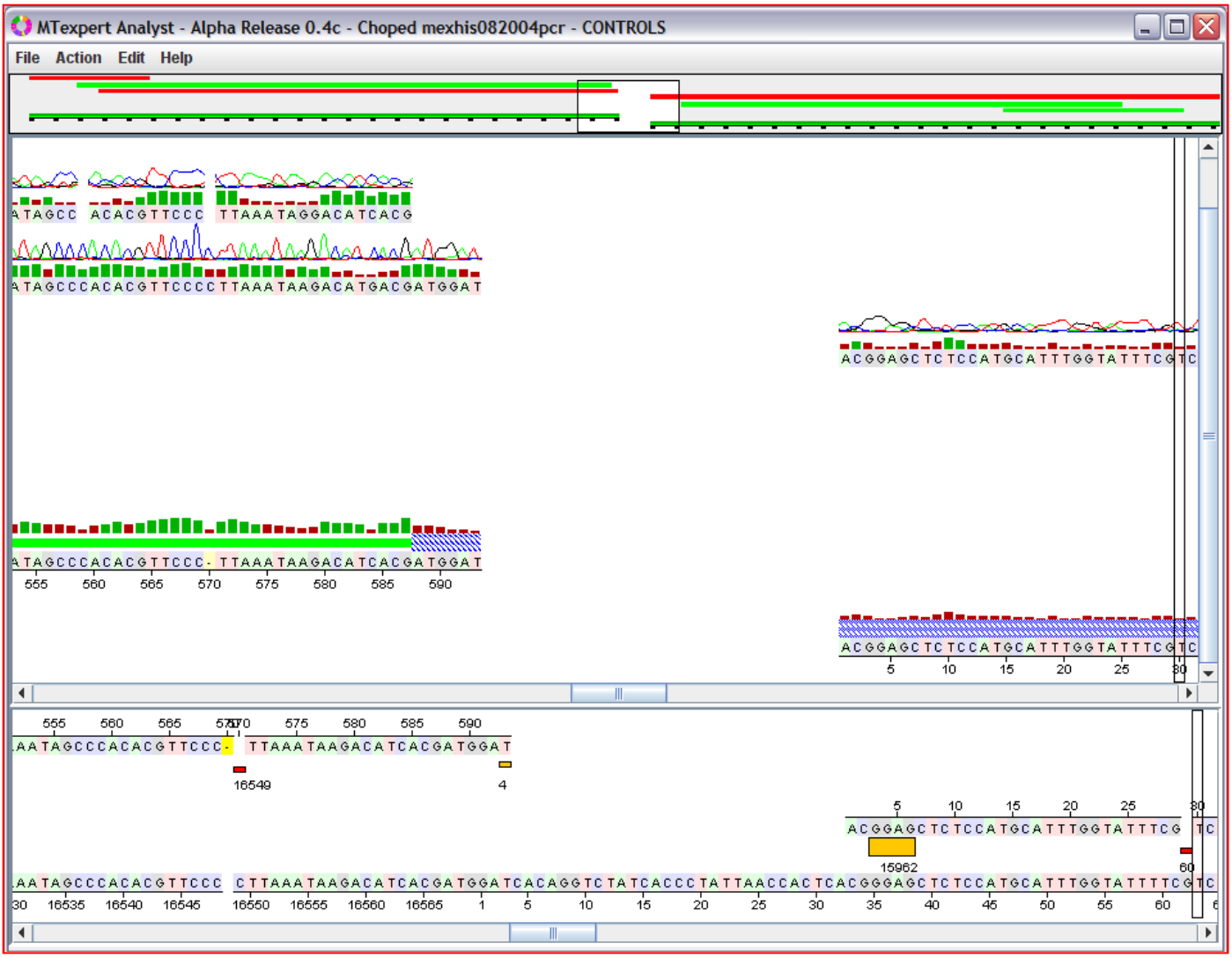



**Figure 19 MTexpert display showing two consensus sequence fragments.**

## 2.2 Consensus Flags Window

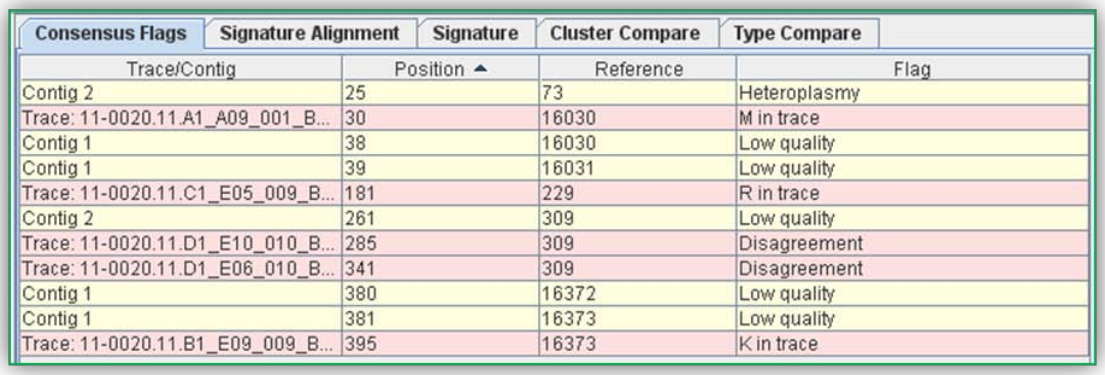
The window in the bottom right corner of the MTexpert Display has three.

- The Consensus Flags Tab displays a table that lists any base or region in the consensus assembly that may need review by the analyst.
- The Signature Tab displays a table that lists the MitoTyper type generated for the sample.
- The Signature Alignment Tab provides information and options on how the consensus is aligned with the reference in the process of generating the signature.

### 2.2.1 Consensus Flags

The first tab in the bottom right MTexpert display window opens the Consensus Flags Table (Figure 20). Consensus Flags are issues in the assembly about which the analyst would like to be warned. Currently a Consensus Issue appears in the list for:

- Any heteroplasmy in the consensus sequence. Any base in the consensus sequence other than A, T, G, or C is reported as a consensus issue.
- Any low-quality base in the consensus sequence (that is eligible to be used in the signature - generated by coverage of at least two traces, one in each direction). Typically the quality threshold is set to 20 on the scale of consensus quality that runs to 100.
- Any "unused" heteroplasmy call in a trace. An unused trace het call is a het call in a trace that is not assembled into the final consensus sequence.
- Any disagreement between high quality base calls between traces (this might also be reported as a low quality base call in the consensus since it could assemble as such). Both bases must be over the trace quality threshold.
- Other indications of problems can be trapped as needed




| Trace/Contig | Position ▲ | Reference | Flag |
|---|---|---|---|
| Contig 2 | 25 | 73 | Heteroplasmy |
| Trace: 11-0020.11.A1_A09_001_B... | 30 | 16030 | M in trace |
| Contig 1 | 38 | 16030 | Low quality |
| Contig 1 | 39 | 16031 | Low quality |
| Trace: 11-0020.11.C1_E05_009_B... | 181 | 229 | R in trace |
| Contig 2 | 261 | 309 | Low quality |
| Trace: 11-0020.11.D1_E10_010_B... | 285 | 309 | Disagreement |
| Trace: 11-0020.11.D1_E06_010_B... | 341 | 309 | Disagreement |
| Contig 1 | 380 | 16372 | Low quality |
| Contig 1 | 381 | 16373 | Low quality |
| Trace: 11-0020.11.B1_E09_009_B... | 395 | 16373 | K in trace |

**Figure 20**: The Consensus Flags Table

The Position is the consensus sequence base location at which the Consensus Issue occurred (the consensus sequence bases are numbered from left to right - these are not the reference base locations).

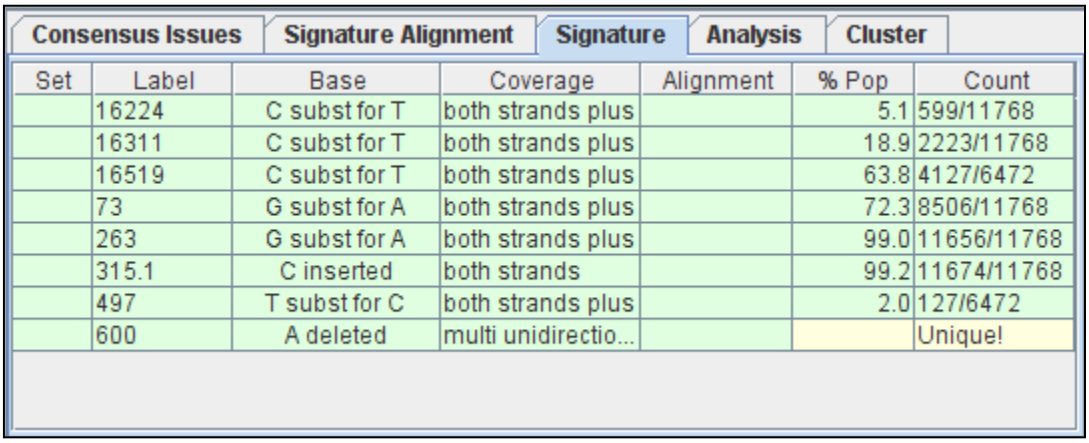
Selecting a line in the Consensus Flags list moves the Assembly Window and the Signature Window to the base location in the consensus sequence, puts the cursor at the appropriate position, and automatically opens each Trace Window and automatically highlights the base that triggered the issue, if possible.

The widths of the columns in this table can be changed by dragging the divider in the table header.

Consensus issue reporting is controlled by user preferences (Section 6.9, Preferences --> Flags). Reporting each consensus issue can be enabled or disabled and the quality thresholds can be adjusted.

### 2.2.2 Signature

The Signature (mtDNA type) is displayed by selecting the second tab in the bottom right MTexpert display window. The Signature Table (Figure 21), describes each difference between the consensus sequence and the reference sequence. The Label is the location of the difference compared to the rCRS base location.




| Set | Label | Base | Coverage | Alignment | % Pop | Count |
|---|---|---|---|---|---|---|
| | 16224 | C subst for T | both strands plus | | 5.1 | 599/11768 |
| | 16311 | C subst for T | both strands plus | | 18.9 | 2223/11768 |
| | 16519 | C subst for T | both strands plus | | 63.8 | 4127/6472 |
| | 73 | G subst for A | both strands plus | | 72.3 | 8506/11768 |
| | 263 | G subst for A | both strands plus | | 99.0 | 11656/11768 |
| | 315.1 | C inserted | both strands | | 99.2 | 11674/11768 |
| | 497 | T subst for C | both strands plus | | 2.0 | 127/6472 |
| | 600 | A deleted | multi unidirectio... | | | Unique! |

**Figure 21: The Signature Table**

Each difference between the consensus sequence and the reference sequence is placed and described according to the MitoTyper alignment and nomenclature rules. A red line in the Signature table indicates type positions that are generated from inadequate consensus coverage. Blue lines indicates type entries that are outside of the Signature Range specified in the user preferences (see Section 6.5). A yellow line indicates that the polymorphism occurs at the end of the consensus that was arbitrarily aligned with the reference sequence (see Section 1.11.3).

The columns % Pop and Count in the Signature window show the analyst the frequency in the database of each polymorphism in the sample's Mitotype. For the % Pop column we calculate the percentage of database sequences that have coverage at that position and contain that polymorphism. This allows you to immediately determine the rarity of each polymorphism in the Mitotype. The Count column the number of samples for which the position has been sequenced (which is the denominator of the % Pop calculations).

A red background for the % Pop and Count cells is a flag that the position has been sequenced before but the polymorphism has not been seen before and so should be reviewed carefully! If there are no samples in the database with coverage at that position (which would be a zero in the Count column), the two columns contain a yellow "Unique!" flag.

Selecting any line in the Signature table automatically moves the Assembly Window and the Signature Window to the base location in the consensus sequence, puts the cursor at the appropriate position, and automatically opens each Trace Window.

If two possible types are being reported for the HV2 region between 300 and 315 (see Section 1.11.1), the lines in the Signature table that vary between the different type sets

are indicated with a 1 or 2, depending on which set they belong to. Right clicking one of these lines let the user switch the display between the two types.

### 2.2.3 Signature Alignment

The Signature Alignment tab provides access to controls that affect the way MTexpert handles polymorphisms at the end of sample sequences. See Section 1.11.3, Aligning Polymorphisms at the End of Contigs. Figure 22 contains an example display. Initially the ends of the sample sequence will be arbitrarily aligned as indicated by the selected option. The analyst can specify the reference base location to which the first and last base in the contig should be aligned in the Complete Start and the Complete End boxes, specifying the Complete alignment option and then using the Recall Type button to regenerate the alignment and signature.

The table has a line for each sample sequence fragment or contig in the project, as shown in Figure 22.

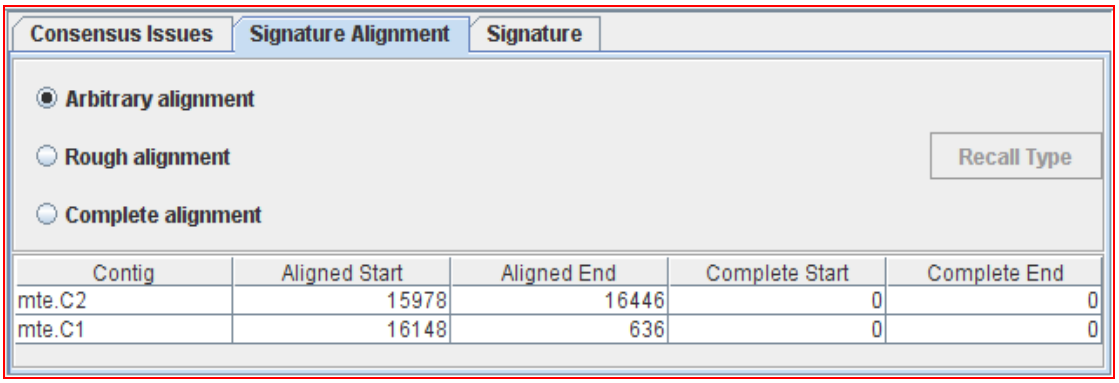



**Figure 22.** Signature Alignment Information and Options

☐ The Complete Alignment choice is not retained if the project is reloaded! Reloading the project will reset the alignment to the Arbitrary alignment.

## 2.2.4 Type Compare

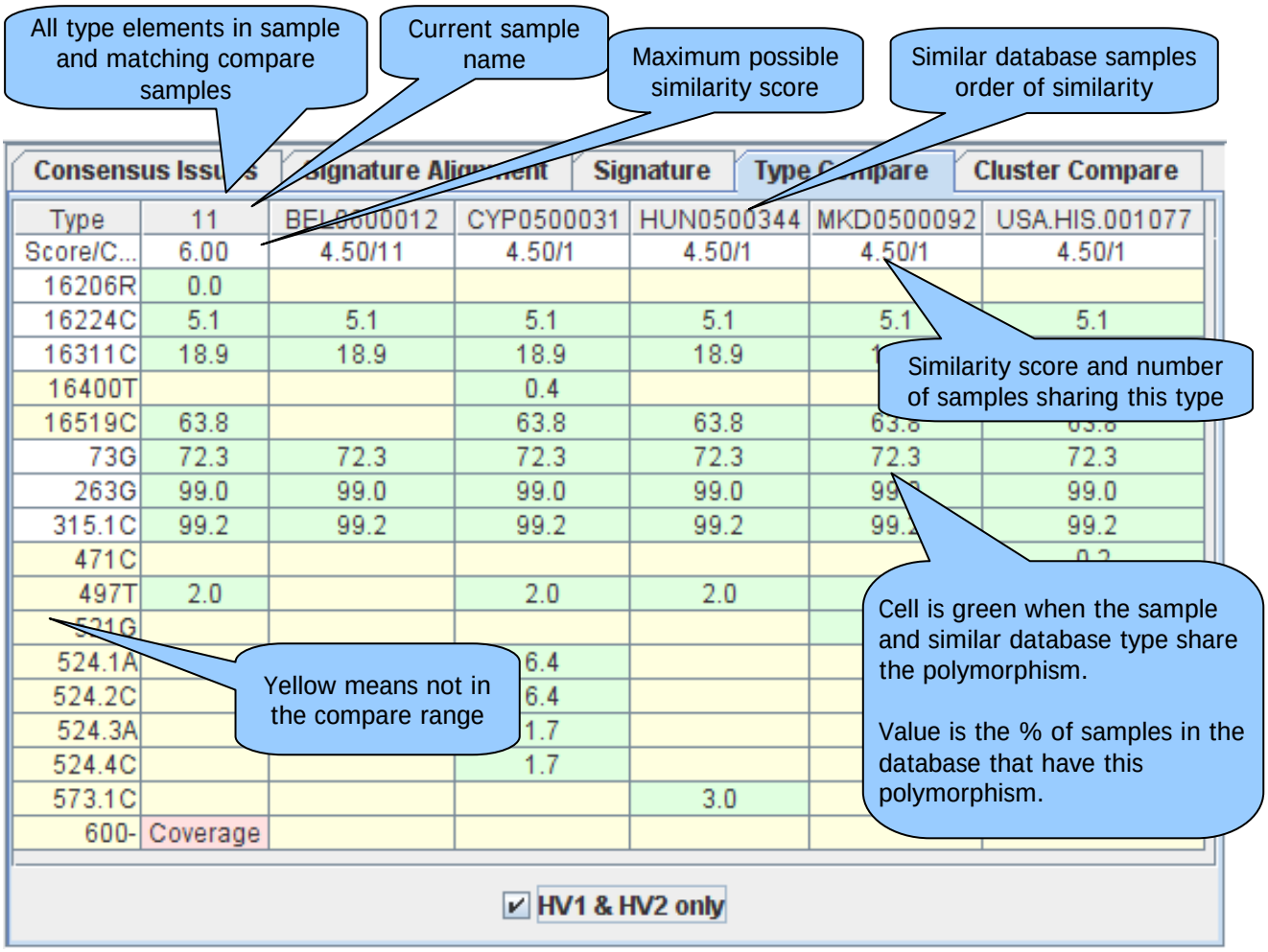



**Figure 23: Type Compare Window**

One way to identify unusual polymorphisms in a sample's Mitotype is to compare that Mitotype to all the other mitotypes in the population database. If the sample's Mitotype has been seen before, there may be increased confidence in the results. When the sample's Mitotype is similar to mitotypes that have been observed before, polymorphisms that are different from the most similar mitotypes can be reviewed to make sure the differences are valid.

The Type Compare window allows you to compare the sample's Mitotype to the most similar mitotypes in the population database and quickly see the magnitude of any differences. The Type Compare window is shown in Figure 23. You can choose to look for similar types in the entire Signature range or just in the HV1 and HV2 regions by using the checkbox at the bottom of the window.

The leftmost column, Type, contains all the type elements in the sample's Mitotype as well as any differing type elements that appear in the similar mitotypes. The sample's Mitotype is provided in column 2 under the sample's name and the most similar database mitotypes are shown from column 3 onward to the right in order of increasing difference.

The Score in the second row of the table, Score/Count, shows the maximum similarity score possible for that sample, and then shows the similarity score for each database Mitotype in the table. The score is the degree of similarity between the sample's Mitotype and the matching mitotypes in the database. A perfect score occurs when the database and sample mitotypes match completely, counting +1 for each matching type element. The similarity score drops by 0.5 for each extra type element in the sample and drops by 0.75 for each extra type element in the database Mitotype. If the sample and database Mitotype both have a polymorphism at a position but the polymorphisms themselves don't match, the match is not credited with a similarity (+0 rather than +1 for a match). This similarity metric was created to search the database and arrange this display and it does not have any particular significance. You are welcome to suggest a different similarity calculation if you want the searching or display altered.

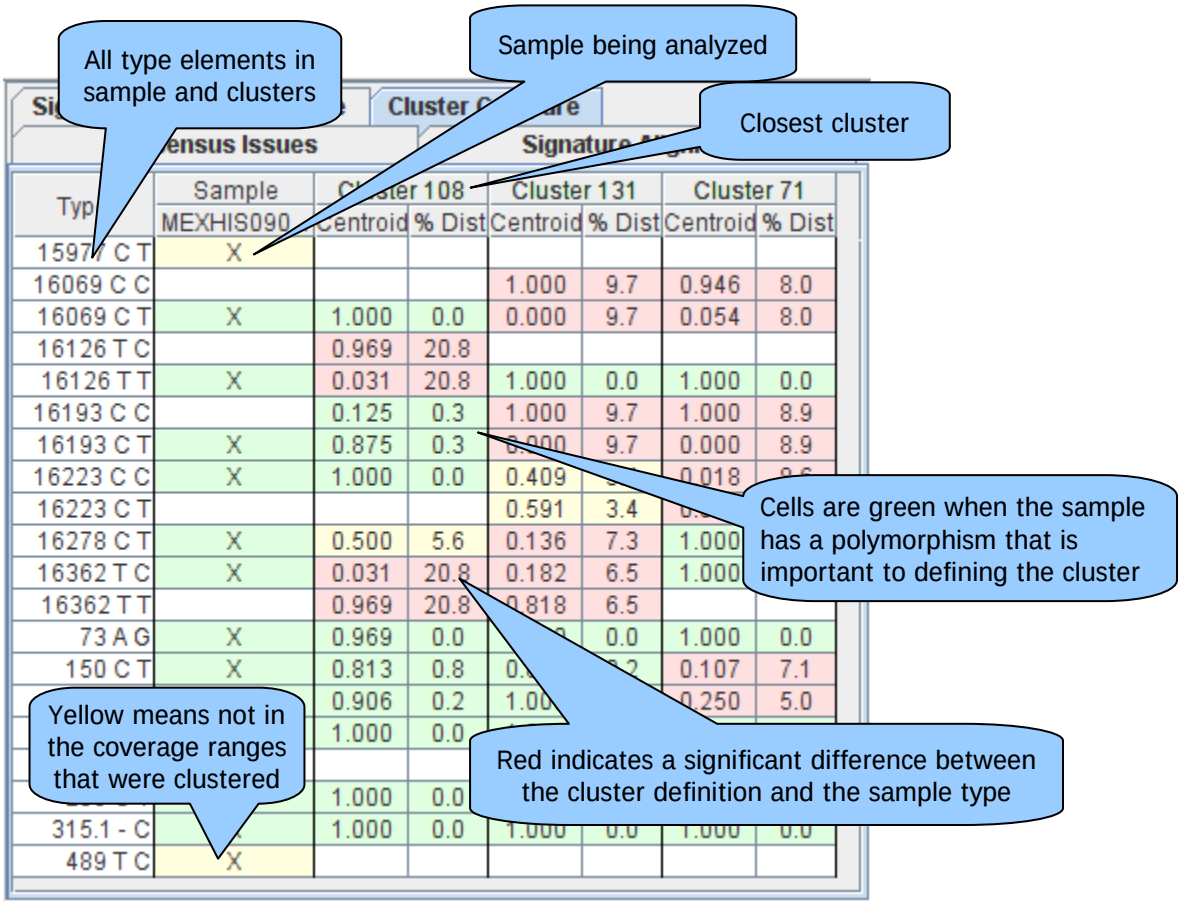
The Count displays the number of times this Mitotype has been observed in the database. (To see the names of the other samples in the database that share this Mitotype, right click anywhere in the column and select that option) You can use this to see how common the similar mitotypes are and potentially be more concerned with differences from common mitotypes than differences from uncommon mitotypes.

In addition to the Score, the cell background colors provide a way to see how similar the "most similar" database mitotypes are. The colors in each cell of the table highlight the matches and mismatches between the sample's Mitotype and the similar mitotypes in the database. For the sample Mitotype itself, the cells are highlighted green for type elements in the sample Mitotype, yellow for those absent from the sample's Mitotype, and red with the message "Coverage" for those that were outside the sequence region covered by the database. A red background also highlight when a polymorphism has never been seen before in the entire database - a significantly unusual event that should be double-checked. In the database type display, cells are colored green when the type elements match and yellow when they do not. The best way to use this window is to select each mismatching (yellow) type element's row, jumping to that location in the traces and consensus, and validate that the polymorphism is correct.

Each green cell also contains the percent of the covered samples in the entire population database that share each polymorphism (this is also the % Pop number in the Signature window). With this number, you can get an indication of how unusual each polymorphism is when compared to the entire database. Potentially, a difference between the sample and the most similar database Mitotype in an unusual

polymorphism is not as significant an indication of a potential quality problem as a difference with a polymorphism that has frequently been observed.

### 2.2.5 Cluster Compare




| Typ | Sample | Cluster 108 | | Cluster 131 | | Cluster 71 | |
|---|---|---|---|---|---|---|---|
| | MEXHIS090 | Centroid | % Dist | Centroid | % Dist | Centroid | % Dist |
| 1597 C T | X | | | | | | |
| 16069 C C | | | | 1.000 | 9.7 | 0.946 | 8.0 |
| 16069 C T | X | 1.000 | 0.0 | 0.000 | 9.7 | 0.054 | 8.0 |
| 16126 T C | | 0.969 | 20.8 | | | | |
| 16126 T T | X | 0.031 | 20.8 | 1.000 | 0.0 | 1.000 | 0.0 |
| 16193 C C | | 0.125 | 0.3 | 1.000 | 9.7 | 1.000 | 8.9 |
| 16193 C T | X | 0.875 | 0.3 | | 9.7 | 0.000 | 8.9 |
| 16223 C C | X | 1.000 | 0.0 | 0.409 | | | |
| 16223 C T | | | | 0.591 | 3.4 | | |
| 16278 C T | X | 0.500 | 5.6 | 0.136 | 7.3 | 1.000 | |
| 16362 T C | X | 0.031 | 20.8 | 0.182 | 6.5 | 1.000 | |
| 16362 T T | | 0.969 | 20.8 | 0.818 | 6.5 | | |
| 73 A G | X | 0.969 | 0.0 | | 0.0 | 1.000 | 0.0 |
| 150 C T | X | 0.813 | 0.8 | | | 0.107 | 7.1 |
| | | 0.906 | 0.2 | 1.00 | | 0.250 | 5.0 |
| | | 1.000 | 0.0 | | | | |
| | | | | | | | |
| | | 1.000 | 0.0 | | | | |
| 315.1 - C | | 1.000 | 0.0 | 1.000 | 0.0 | 1.000 | 0.0 |
| 489 T C | X | | | | | | |

**Figure 24: Cluster Compare Window**

Patterns of human mitochondrial DNA sequence variations are often conserved in clusters or haplogroups that are defined by population and ethnic groups and by geography. It is possible to identify unusual polymorphisms by comparing the sample's Mitotype to the most similar clusters of samples that have been identified in the database. The difference between a sample Mitotype and the other mitotypes of a cluster may be the result of polymorphisms that merit closer inspection, especially if the source of the sample is thought to be a member of that cluster.

The Cluster window allows you to compare the sample's Mitotype to the three closest clusters in the database. The Cluster window is shown above.

The leftmost column, Type, contains all the type elements in the sample's Mitotype as well as any other type elements that define the closest clusters but are missing in this

sample's Mitotype. An "X" in column 2 (the column is headed with the sample's name) indicates that the type element appears in this sample's Mitotype. A yellow background in column 2 means that the sample's type elements are outside the coverage used to calculate the clusters so these type elements could not be used in the sample's comparison to the clusters.

The remaining columns contain a pair of columns for each similar cluster. These columns have information that helps you analyze the relevance of the cluster comparison. The first column in a pair is Centroid, showing the polymorphism's average presenc in all the database samples that were grouped together to form that cluster. The centroid value for a polymorphism will be close to 1 if most of the samples in the cluster have the polymorphism and will be close to zero if most of the samples do not have the polymorphism.

The second column for a cluster, % Dist, helps highlight which type element disagreements cause the most dissimilarity between the sample and the cluster. A large % Dist indicates that this polymorphism difference is responsible for a significant amount of the disagreement between the sample and the cluster. The most information for quality analysis can be found from comparison of the sample to clusters in which there are only a few significant polymorphism disagreements. If each disagreement contributes less than 10% of the total difference, the sample is not a good fit with that cluster and the individual polymorphism differences are less indicative of a suspicious base call.

To focus visually where similarities and dissimilarities occur, the colors of cells highlight the matches and mismatches between the sample's Mitotype and the cluster. Cells are green when the type element of the sample and cluster agree. The cells are red when the sample's type element and the cluster value for that element disagree significantly. A yellow background highlights a moderate disagreement. Typically, two rows in the table describe each significant disagreement between the sample and type - a row for the sample polymorphism and a row for the missing element. Blank cells indicate a polymorphism absent in the type and significantly absent in the cluster – a significant agreement in the absence of the polymorphism.

The best way to use this window is to select each mismatching (red) type element, jumping to that location in the traces and consensus, and validate that the polymorphism is correct. However, even if the cluster is reported as one of the three closest, this does not mean that the comparison with the sample is relevant and can be used to find suspicious polymorphisms. The most relevant quality assessment information can be found in comparisons with clusters that have only a few significant type element differences with the sample - which are shown in the display with many green elements and a few red elements.

There is no way to assign a haplogroup to the sample being analyzed at this point. Haplogroups described in the literature are defined by polymorphisms that are outside the sequenced range and so this necessary information is not available to make a decision on haplogroup assignment. However it is possible to get an impression of which ethnic categories make up the cluster from the file names of the database samples that fell into the cluster. To see a list of the database samples associated with each cluster, right click anywhere in the cluster heading and select that option.

## 2.3 Trace Windows

An individual trace appears in its own window when the base location is double-clicked in the Assembly Window or the Signature Window (or a Consensus Issue involving that trace is selected). Trace windows can also be opened by clicking the trace action in the Trace Action Window. The windows for each trace are tiled to the left of the screen. An example is shown in Figure 25 with some components of the display labeled.

The Trace Window displays a plot of the trace electropherogram, the base calls, and the quality values as a bar graph above each base call. Moving the cursor in the display window selects the base, indicated by a black box around the base and a vertical line through the electropherogram peak that generated the base call. The cursor can also be moved point-by-point along the electropherogram scale. On the left panel of the Trace Window there is also a "Current Base" display to the right of the "Values" button. This can be handy if the scale is so large that you cannot read the individual base boxes under the peaks.

In the Trace Window, the four electropherograms are plotted from right to left for forward strands or from left to right for reverse strands. The Trace Window in Figure 25 displays a reverse strand, so the electropherogram points are plotted from right to left. The arrow above the display controls also provides a visual indication that the trace is a forward or reverse strand. In Figure 25, the arrow pointing to the left indicates a reverse strand.

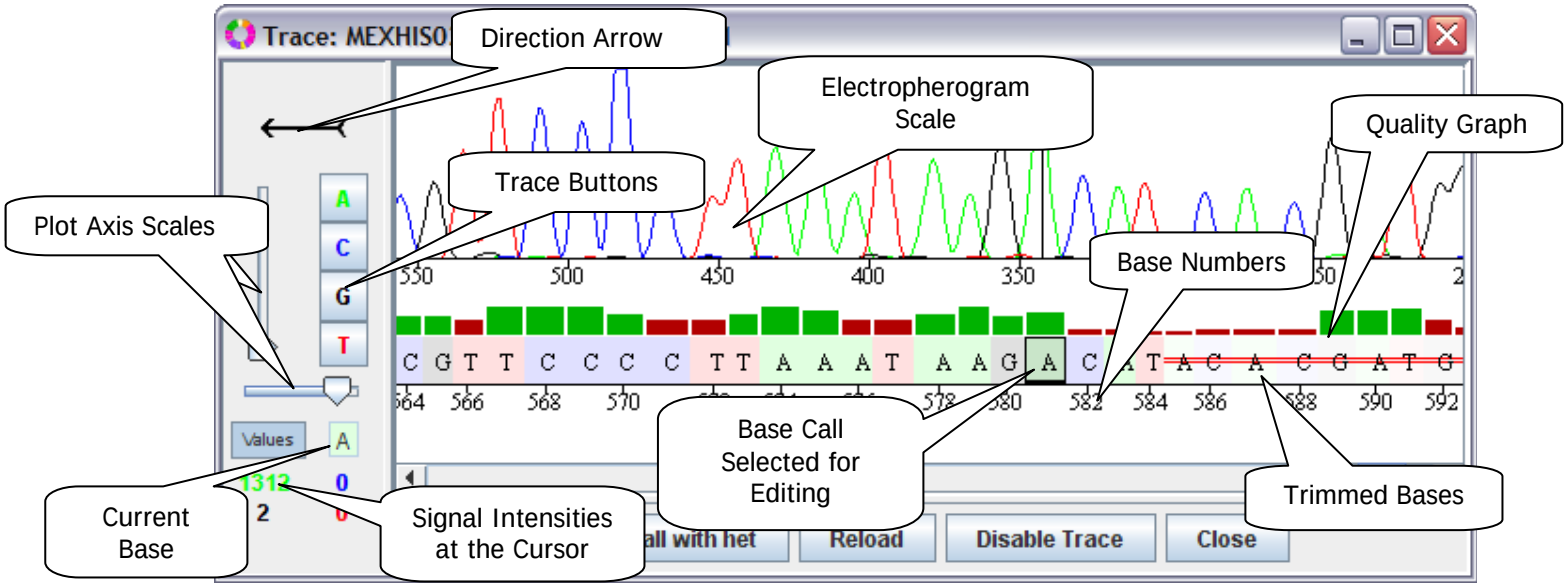



**Figure 25: Trace Window Display**

The plot of each of the four electropherograms can be switched on or off with the buttons to the left of the plot. The height and width scales of the electropherogram plot can be changed with the slider bars to the left of the plot. The lowest position of the vertical intensity slider will show the electropherogram scaled so that the highest peak just touches the top of the electropherogram window. Adjusting the intensity slider will amplify all peaks between one and five times, clipping any peaks that are higher than the display.

Selecting the Values toggle button adds a display under the check box of the absolute intensity of each electropherogram at the cursor location, as shown in Figure 25.

The base calls below the trace are always numbered from the start of the trace data file and displayed increasing to the right.

The quality of each base call is reported by the MitoTech TraceTuner base caller on a scale from 1 to 40 with 40 being the highest quality base call. The quality value for each base call in a trace is shown graphically as a bar graph above each base with a taller bar indicating a larger quality value. The bar is green if the quality is greater than 20 and red otherwise.

Right clicking any base will display the call and quality value from the TraceTuner basecalling software. A base is displayed with an underline if TraceTuner produced multiple base calls at that location - this typically occurs if there is a heteroplasmic call or another significant peak at the base location that was not the called base. The pop-up window lists the calls in decreasing quality. The highest quality call will be the one displayed in the trace display.

Any multiple calls at a base location are passed to the assembly program. The current parameters for the assembly program allow it create a mixed (heteroplasmic) base in the consensus sequence when 2 or more bases have a mixed call at that location that has a lower probability than the highest probability base call as long as both have a "reasonable" quality (at least 50% of the highest). So, if there are 2 traces aligned and each call an "A" with a lower quality "M" that is 50% of the A quality the assembly will contain the "M".

The Close button on the bottom of the window simply closes the window. The window can also be closed with the traditional  in the upper right corner. There is no command to save edits since all edits were immediately saved to the MTexpert project when they were made. The Close All Traces Action is a handy shortcut to close all the open Trace Windows at once (see Section 5).

The Reload button is used to replace the current trace in the project with the data from the original trace file in the data directory.

☐ **Important Note:** Reloading eliminates all of the edits to this trace in the project.

Use the Recall with Het or Recall w/o Het button to switch between enabling or disabling heteroplasmy base calling in TraceTuner for this trace. When this button is selected, the trace is rerun through TraceTuner and reassembled into the sample assembly – all of the edits on this trace are lost in the process. See Section 1.6 for a discussion of heteroplasmy base calls. This button is a useful alternative to manually editing each heteroplasmy call if there are a number that result from a low level of noise.

Use the Disable Trace or Enable Trace button to switch between including or excluding this trace in the project. If a trace is disabled, it is marked as unusable in the project and the background of this trace's lines in the Trace Action Window are red. This button is equivalent to the "Mark Trace as Useable" and "Mark Trace as Unusable" commands in the Trace Action Window, Section 2.1.4.

### 2.3.1 Changing the Appearance of the Trace Window

When a new Trace Window is opened, its size and scale are the same as the size and scale of the last Trace Window that was closed. To change the display, open only one Trace Window, change the size and scale, and then close it. The next Trace Window(s) opened will have the same display scale and window size.

# 3 Editing a Trace

Trace editing is performed in the Trace Window.

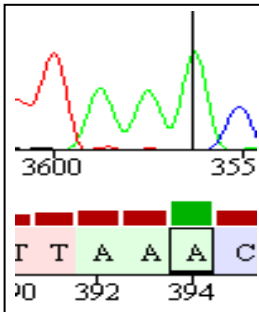



**Figure 26: A region of the Trace Window showing a base selected for editing**

Double clicking a base in the Trace Window selects it for editing, as shown by a graying of the box of the selected base. Figure 26 illustrates a base selected for editing. A range of bases can be selected with the mouse by selecting the first base and dragging the cursor over the range before releasing the mouse button.

With the base selected for editing, typing any legal IUPAC base character will replace the existing base call. The Delete key on the keyboard can be used to remove a base call or a range of base calls.

To insert a base, position the vertical line cursor at the appropriate point in the electropherogram display and right-click the mouse. A command window appears allowing you to select Insert Base at Cursor. A base is inserted between the adjacent bases and set to be the base with the highest-value electropherogram at that location. This inserted base is immediately selected for editing and so can be changed with the keyboard or removed again with the Delete key.

After every trace edit (base change, base delete, or base insertion) the consensus sequence is immediately reassembled and the Assembly Window, Signature Window, and Consensus Flags are updated to reflect the new consensus sequence. On slower computers, this may take a second or two.

Every edit to the trace is recorded as a line in the Trace Actions Window in the MTexpert display. Each edit can be undone by right clicking the corresponding line in the Trace Actions Window. Subsequent actions, if any, may have to be undone first. If it is the last edit in the project, a Ctrl-Z in the active Trace Window can also undo the last edit.

## *3.1 Control keys for trace editing.*

Selecting large stretches of a trace in the trace edit window is now has control keys that follow the Sequencher convention. To use these control keys, select the first base to be trimmed by double clicking it with the mouse. Pressing ctrl-[ (the control key and the right square bracket simultaneously) extends the selection to the left-most base and pressing ctrl-] extends the selection to the right-most base.

- Ctrl-T (or ctrl-t) is used for manually trimming a selected region. This only works if the selected region includes either the first or the last base in the sequence. To trip a region from the end of the sequence, select the first base to be trimmed and then select to the end of the sequence with the ctrl-[ or ctrl-] and then use ctrl-T. The Trace Window display now shows trimmed bases using a red bar as well as a lighter color.
- Ctrl-R (or ctrl-r) restores original pre-trim calls, i.e., clears all the trimming in the selected region.
- Ctrl-Z (or ctrl-z) in a trace undoes the last edit (base insert, delete, trim/untrim) for the trace in the active Trace Window. All the edits can be undone in series from the most recent first. This command results is equivalent to the Remove This Trace Edit command for that trace edit recorded in the Trace Action Window.

# 4 Editing the Consensus Sequence

In order to make a change in the consensus sequence, the traces must be edited and reassembled. You cannot directly edit the consensus sequence.

# 5 Project Actions

The Action pull-down menu, Figure 27, provides direct access to some program functions. Actions that are grayed out are not available at the current state of the project.

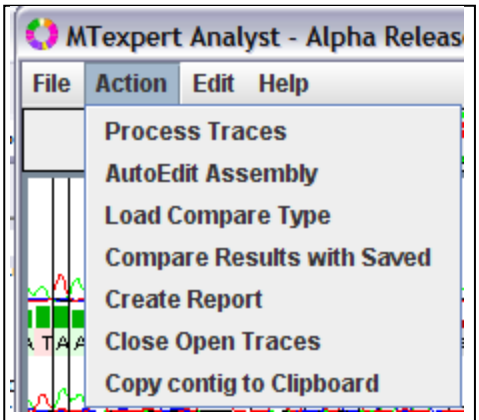



**Figure 27: Action pull-down menu**

**Process Traces** is a way to restart the automated workflow automation without saving and reopening the project. Selecting the Process Traces command starts the workflow at Open a Project in the flow chart in Section 9.1.

The Process Traces action is particularly useful to restart the automated processing when a control problem prevented the use of a primer in all the samples in a project. Rather than saving and then reopening the project, the Process Traces action can be used to force a project-wide re-execution of the automated processing starting at control validation.

**AutoEdit Assembly** runs expert rules that automatically fix specific consensus issues. See AutoEdit Rules, Section 1.9, for more information.

**Load Compare Type** and **Compare Results to Saved** are two actions that allow duplicate analyses of the data to be compared. An independent type has to be generated and saved in a .txt file format. The MitoTyper software writes a text file in a format that MTexpert can read with this command and we also try our best to read a signature report text file generated by Sequencher. Use the file browser to select the .text file containing the independent type.

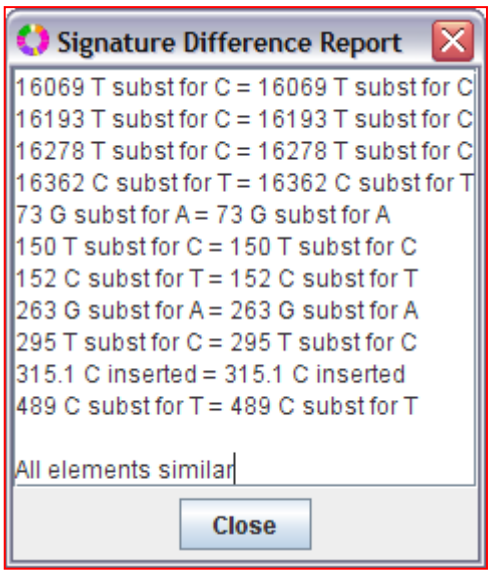



**Figure 28 Example Signature Difference Report**

**Compare Results to Saved** allows an Analyst to compare the current assembly and signature to the saved project results. This action generates a Signature Difference Report shown in Figure 28. In this window, all of the signature elements are compared with "=", "|", "<" or ">". The Report appears in an editable text window. The text can be selected and edited and copied and pasted with normal Windows commands (ctrl-c and ctrl-v).

**Create Report** is used to create an output text file or a CODIS 4.1 XML format file. Selecting this option opens a file browser dialog box that is used to name and save the output report in either the .txt or .xml format. Specifying the XML report creates a file in the CODIS 4.1 format. Two text report formats are also available from the Files of Type pull down menu. The Text Signature format saves the content of every line in the Signature table as Text. The Text Project Report option contains every line in the Trace Action Window and every line in the Signature Window.

The **Close Open Traces** action is a handy way to shut all of the Trace Windows at once when the software has opened a lot of them.

# 6 Software Preferences

The Edit→Preferences pull-down menu opens the software MTexpert Preference Settings window shown in Figure 29. The MTexpert Preference Settings window has a set of tabbed menus of options that adjust the way the software functions and the appearance of the displays. Tabs in this window select different sets of preferences organized (loosely) by software function. Some settings can only be changed by administrator-level users.

Making a change on any menu and selecting OK closes the window and implements the change in the currently-running software as well as each time this registered user (see Section 1.2.2) restarts the software. The software preferences can also be saved and retrieved from a file using the Export and Import buttons on the bottom of this window. This provides a convenient way to save and load different preferences (as .pfxml text files) for different users.

## *6.1 Display*

Options on the Display tab change the current operation and appearance of the software.

Selecting the "Open Trace Edit on File Open" box causes the program to open a Trace Window automatically when a trace file is loaded into a project.

The "Always Show Trace in Assembly" and "Always Show Quality in Assembly" boxes set the default display in the Assembly Window to include the quality bar graph and trace display. The Assembly Font Size allows adjustment of the base character font size in the Assembly Window and Signature Window. The base character font size is the fundamental limitation on how much of a sequence can be fit in the Assembly Window and Signature Window so with smaller fonts, more of the sequence can be seen.

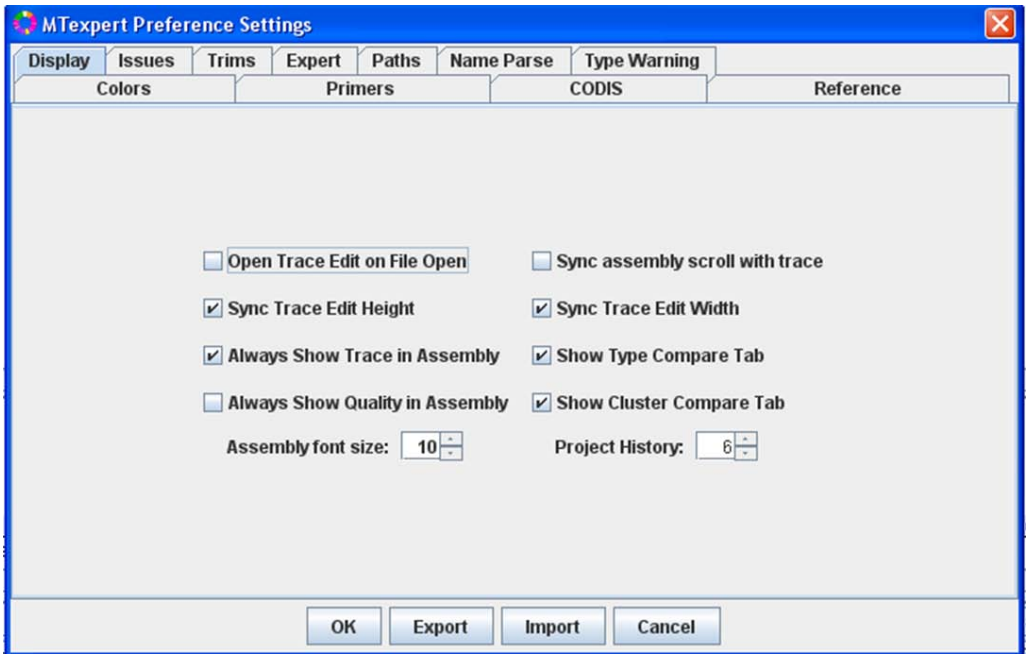



**Figure 29. The Display tab**

## 6.2 CODIS

Some text fields in the CODIS report can be saved as user preferences so that they do not have to be entered each time a CODIS report is generated. These include the Destination Lab, the Source Lab, the Specimen Lab, the "Sent by User" and the "Fragment User".

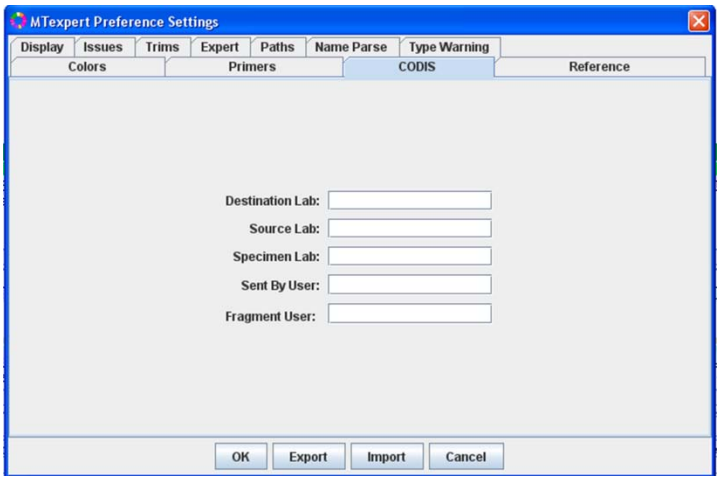



**Figure 30: The CODIS tab**

## 6.3 Paths

The Root Data Directory specifies the directory that will appear in the Create New Project and Project Open and Project Save dialogs. It is most convenient if this directory is set to be the root directory in which the sub-directories that contain each sample's sequence data are found. The TraceTuner Exec, PA Assembler Exec, and PA Parameter File Paths are set by default during the software installation. These paths point to the location of additional executables called by MTexpert. Don't change these unless you really know what you are doing. Only a user with Administrator privileges can change these paths.

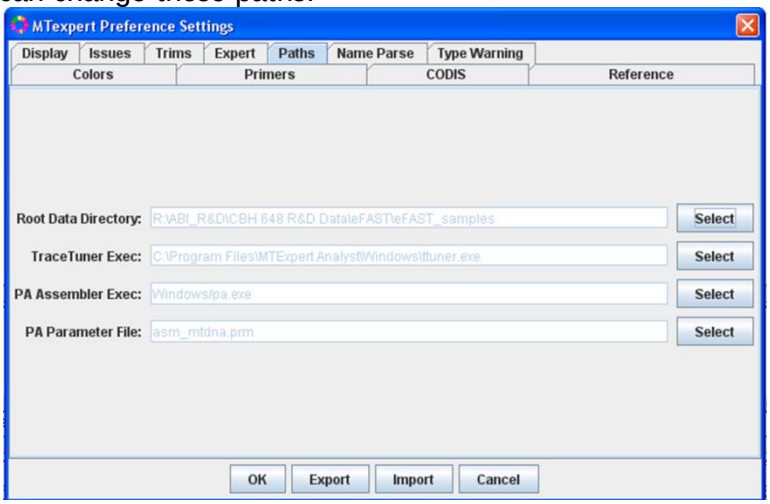



**Figure 31: The Paths tab**

## 6.4 Colors

The Color Preferences control panel allows the user to change the colors of the boxes associated with the base calls. For each base call, three different background color/shade options are provided: one for the base call box and a second when that base is selected and a third when that base is trimmed. Clicking on any box opens a color window allowing you to set the background color of that box.

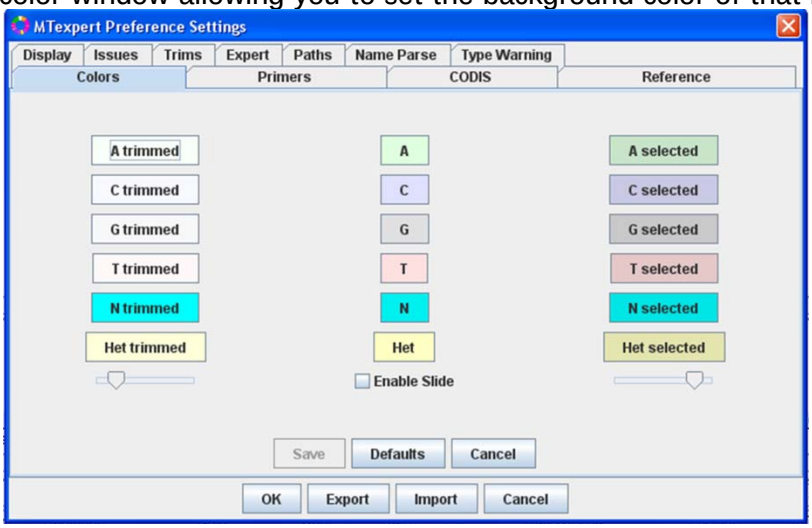



**Figure 32: User Color Preferences settings window**

By default, the color scheme uses shading of a standard color to indicate the status of that base. To adjust the density of all the shading easily, select the "Enable Slide" check box and then use the sliders to adjust the shade density.

Modified color schemes will be retained the next time that the user logs into the software.

The "Defaults" button reloads the system defaults.

## 6.5 Reference

The Reference preferences, shown in Figure 31, control how a type is generated.

The Reference Sequence File contains the base sequence of the reference mtDNA in the control region. The default file is installed with MTexpert and should NOT be changed since it is the basis for generating all the mtDNA types.

The Reference Range is the portion of the reference sequence that is sent to the MitoTyper signature code and against which the consensus sequence is aligned and typed. This parameter also controls the display width of the Signature Window – the entire Reference Range will be available in this window. The Signature Range is the range within the Reference Range in which a signature is generated. Automated MTexpert processing will generate an error if there is not at least forward and reverse coverage in this range. Polymorphic types outside the Signature Range are shown with a blue background in the Signature table and are not included in a report.

- When a project is newly created, MTexpert copies the information about the reference sequence, the reference range, and the primers into the new project file. **Once established, this information cannot be changed in a project.** Editing these settings on the Reference tab in the MTexpert Preferences Setting Window panel only changes the settings in the next new project.

  Three parameters in the Reference preferences window affect the operation of the MitoTyper rules: Enable HV2 C-Stretch, Report Threshold, and Signature Run Time. The Signature Run Time parameter controls how long the MitoTyper rules try to generate a signature for each polymorphic region. Setting this parameter to a few seconds is a reasonable default to prevent the software from appearing to hang while it is trying to generate a type for a difficult polymorphic region. If the timeout is too short, a message appears and the run time can be increased to try again as described in Section 1.11.2. Remember that if you change the Signature Run Time in order to fix a typer timeout, you will need to use the Process Traces action (Section 5) to rerun MitoTyper rules with the new run time.

  The Enable HV2 C-Stretch and the Report Threshold parameters affect the alternate typing of the 300 to 310 HV2 C-stretch region in MitoTyper. MitoTyper aligns the reference and sample sequences in this region slightly differently in order to preserve the 310T alignment and the poly-A and poly-C motifs (see Section 1.11.1). The Enable HV2 C-Stretch check box enables or disables this alternate processing. The Report Threshold causes the MTexpert software to report when the regular processing and special HV2 C-Stretch processing produce types that differ by the threshold or more places. So if the regular typing rules X number of differences from the rCRS in this region, then any HV2 C-Stretch result with fewer than X+Report Threshold number of differences from the rCRS is reported alone as the preferred type. If the HV2 C-Stretch result has X+Report Threshold or more differences from the rCRS, then both types are reported.

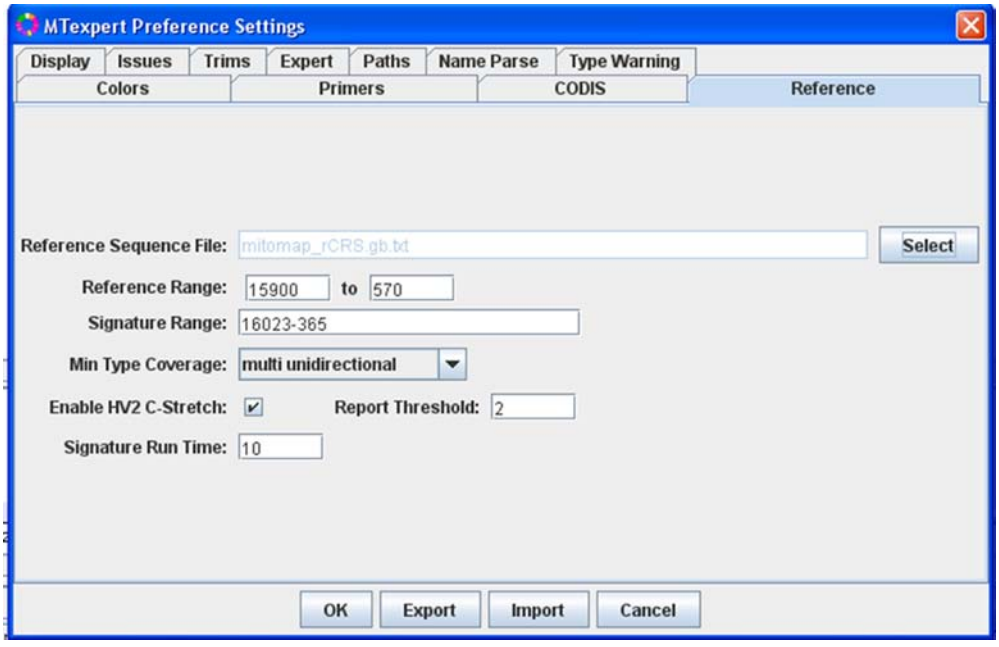



**Figure 33: Reference user preferences window**

## 6.6 Primers

The Primer File is also installed with MTexpert. The primers defined in this file are listed in the table on the Primers tab of the Preference Window. Primers can be added and removed from this file using the Add Primer and Delete Primer buttons. The Start and End and Direction references for each primer determine the location of region in the mitochondrial genome that is amplified by this primer.

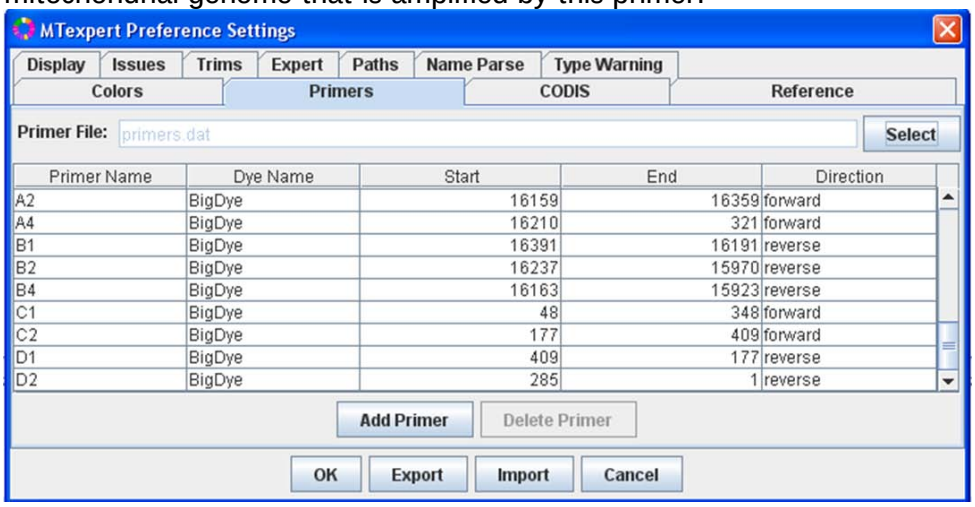



**Figure 34: The Primers tab**

## 6.7 Expert

The MTexpert software can run automated "expert" editing rules as described in Section 1.9. The rules executed by the AutoEdit Assembly command in the Action pull down menu can be included or excluded with the check boxes in this preference tab. These preferences are user specific preferences - MTexpert will remember these settings for each user (see Section 1.2.2).

## 6.8 Trim

The MTexpert software automatically trims the trace files when they are loaded (see Section 1.5.1, Automated Trimming). The heteroplasmy-based trimming can be enabled or disabled in this tab and the window size and thresholds set. Note that trace trimming occurs only when the trace is loaded - if these parameters are changed the project needs to be recreated in order to reload the traces and see the effect.

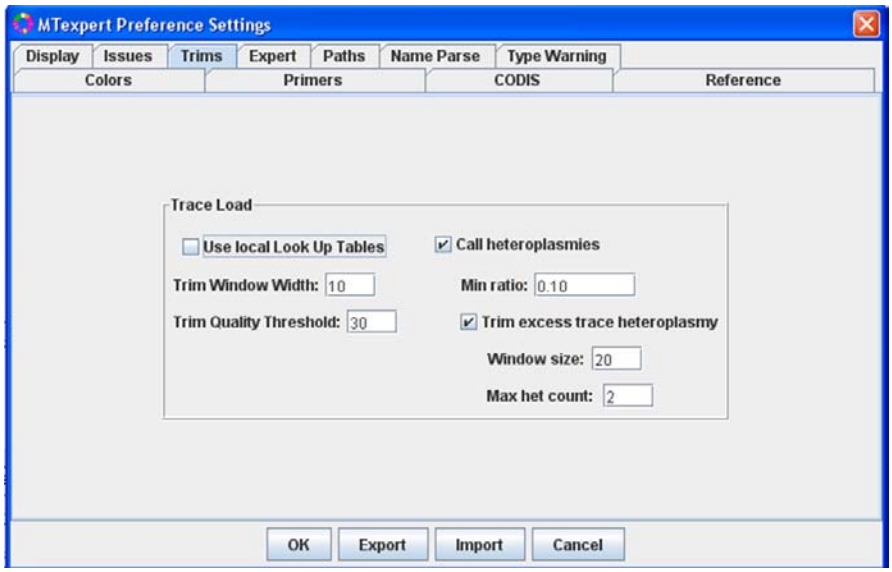



**Figure 33: The Trims tab**

## 6.9 Flags

Flags in the assembly process and the consensus assembly issues are detected and reported in the Consensus Flags list (see Section 2.2.1, Consensus Flags). If any issues are reported in the Consensus Flags list, automated completion of the sample analysis is prevented and the sample is shown as red in the Samples table.

The analysis of individual Consensus Flags can be enabled or disabled on this Preferences tab and the quality thresholds for some of the issues can be adjusted.

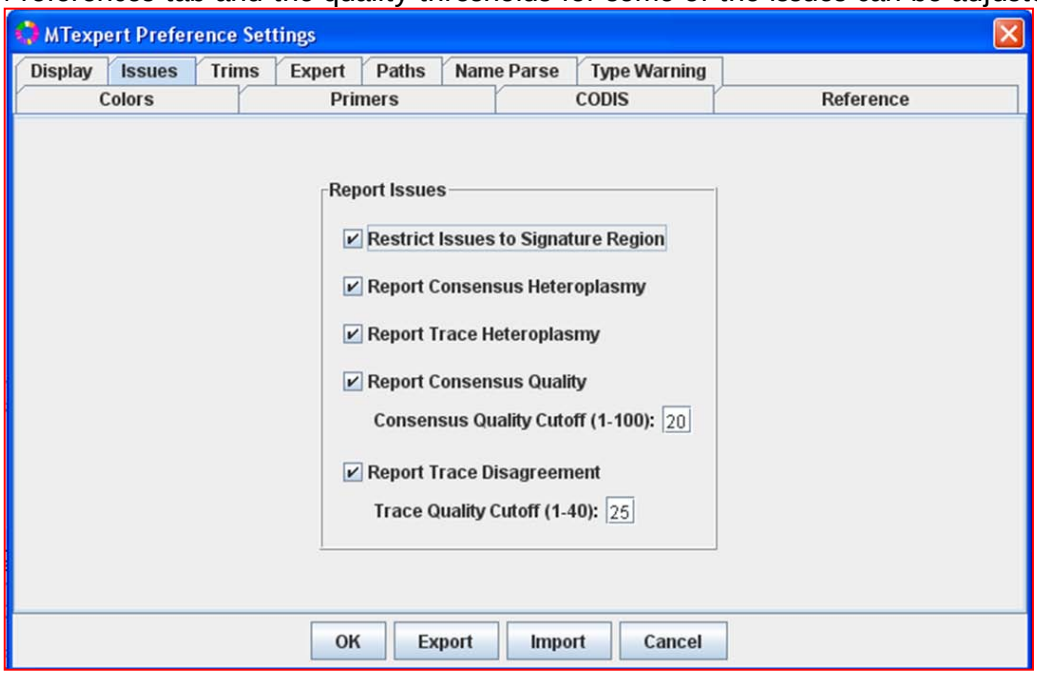



**Figure 34: The Review tab**

## 6.10 Name Parse

The Name Parse tab of the Preference menu allows user configuration of how MTexpert will read file names. It allows the user to define how the file name string will be parsed in terms of string characters and fields. This window is very useful; however, establishing naming patterns can be quite complex; it is recommended that users contact MitoTech for new file name definitions as needed should the following documentation not be sufficient.

The File Name list contains a list of example filenames for which we have developed parsers (i.e. file names that we know how to parse). When reading an ab1 file, MTexpert attempts to fit each file naming scheme that is illustrated by an example in the File Name list, in order, to the file names in the current project's directory. When the first matching file naming scheme is found, the files are read according to that scheme. In order to change the priority with which naming schemes are tried by the software, you can move the file naming scheme up and down in the order with which they will be attempted by selecting the example File Name and using the Move Up or Move Down buttons. The naming scheme that your lab expects to use should be at the top of the list to increase the efficiency of file parsing.

Each parse pattern is broken into groups. Each group may be paired to a matching description of the group. The description is one of "not used", "sample name", "control name", "amplification", "primer", "well", "capillary", "run number", "run date", "batch" or "reinjection". Figure 32 shows the Name Parse tab and indicates the function of each pane.

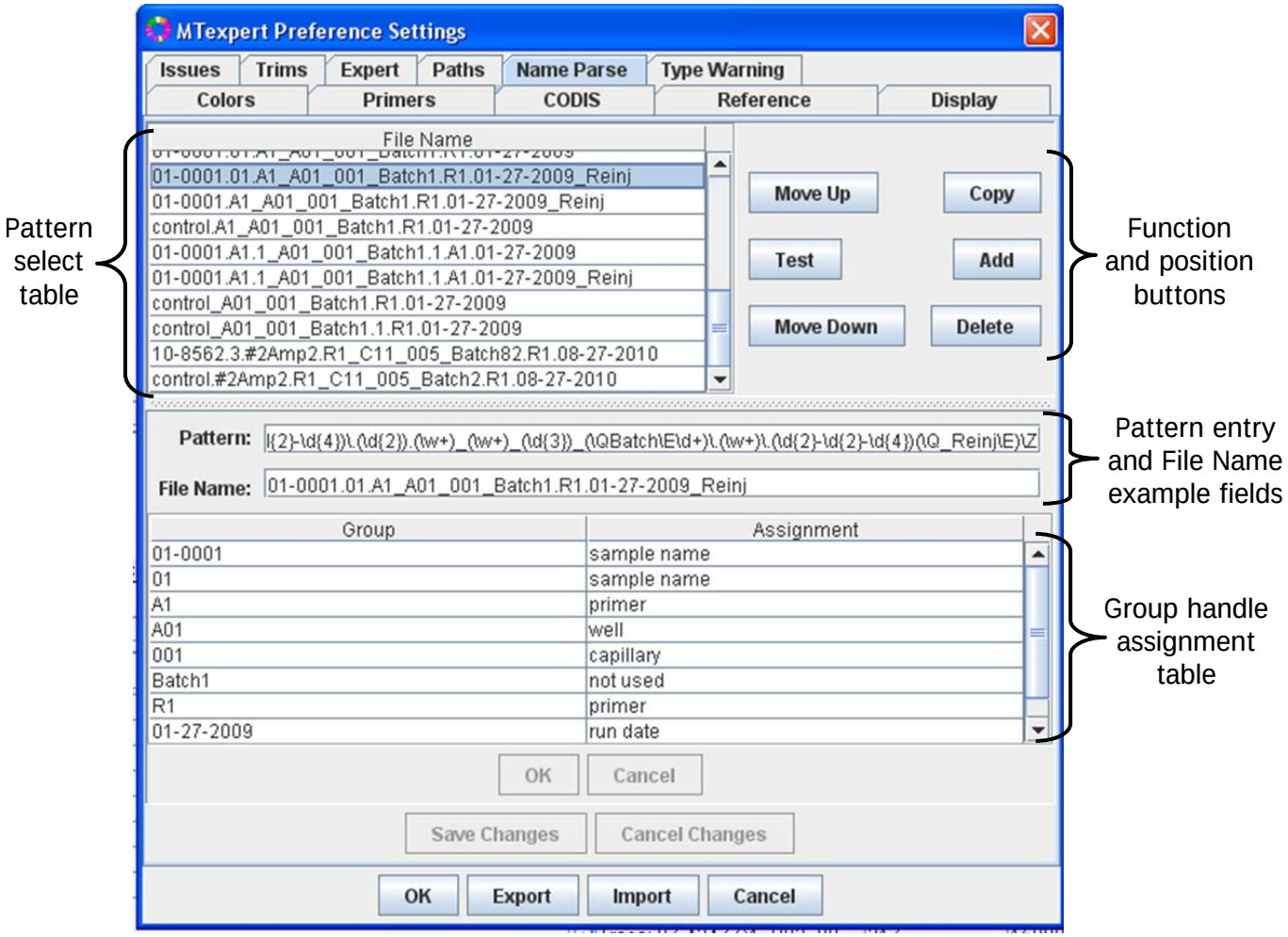



**Figure 32: Identification of the panes in the Name Parse Tab**

### 6.10.1 Definition of symbols in Name Parse patterns

All patterns are bracketed with \A and \Z to force the whole file name to be matched; this eliminates the possibility of partial matching within the file name.

Example:

\A(\d{2}-\d{4})\.(\w+?)[.](\d+)_(\w+)_(\d{3})_(\QBatch\E\d+)\.(\d+)\.(\w+)\.(\d{2}-\d{2}-\d{4})\Z

The pattern has distinct pieces - matches with parentheses around them and matches without. The parentheses define groups, and the groups map into the table below the pattern and sample. Groups are mapped to the field names using the sample file name

matched by the pattern. Internally groups have numbers - group 0 is the whole of the match, group 1 is the first parentheses, group 2 is the second parentheses, etc.

Example: This pattern has eight groups, each shown in a different color.

\A(\d{2}-\d{4})\.(\w+?)[.](\d+)_(\w+)_(\d{3})_(\QBatch\E\d+)\.(\d+)\.(\w+)\.(\d{2}-\d{2}-\d{4})\Z

Matches are reasonably simple for our use:

| | | |
|---|---|---|
| \d | → | matches a digit |
| \w | → | matches a letter or digit |
| [a-z] | → | matches a range of letters, a-z |
| [az] | → | matches either a or z |
| . | → | matches any character |
| \. | → | matches . |

[For matching more than one character]

| | | |
|---|---|---|
| * | → | matches zero or more |
| + | → | matches more than one |
| {n} | → | matches exactly n |
| (n,} | → | matches n or more |

For matching a string it may be surrounded by \Q and \E, all the string between them must be matched exactly.

Example:
\A(\d{2}-\d{4})\.(\w+?)[.](\d+)_(\w+)_(\d{3})_(\QBatch\E\d+)\.(\d+)\.(\w+)\.(\d{2}-\d{2}-\d{4})\Z

### 6.10.2 Example of File Name Pattern Parsing

The following example explains the pattern and filename shown below:

\A(\d{2}-\d{4})\.(\w+?)[.](\d+)_(\w+)_(\d{3})_(\QBatch\E\d+)\.(\d+)\.(\w+)\.(\d{2}-\d{2}-\d{4})\Z

Filename→ 01-0001.A1.1_A01_001_Batch1.1.A1.01-27-2009

| Group number | Component | Matches | Assignment |
|---|---|---|---|
| Group 1 | (\d{2}-\d{4}) | 01-0001 | sample name |
| | \. | . | terminal . |
| Group 2 | (\w+?) | A1 | primer |
| | [.] | . | terminal . (maybe should be \.?) |
| Group 3 | (\d+) | 1 | amplification |
| | _ | _ | terminal _ |
| Group 4 | (\w+) | A01 | well |

| | | | |
|---|---|---|---|
| | _ | _ | terminal _ |
| Group 5 | (\d{3}) | 1 | capillary |
| | _ | _ | |
| Group 6 | (\QBatch\E\d+) | Batch1 | batch |
| | \. | . | |
| Group 7 | (\d+) | 1 | amplification |
| | \. | . | |
| Group 8 | (\w+) | A1 | primer |
| | \. | . | |
| Group 9 | (\d{2}-\d{2}-\d{4}) | 01-27-2009 | run date |

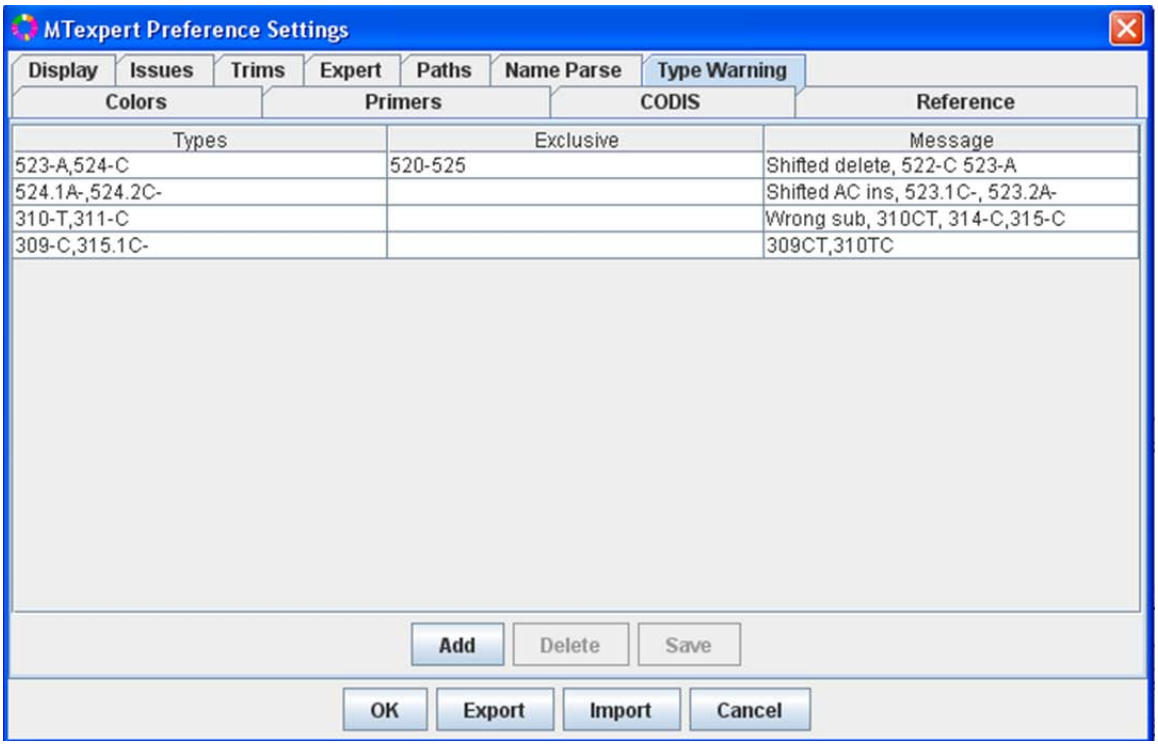
The possible patterns are applied to the filename until one matches (remember the start and end anchors), then the group data is pulled out by name. Currently the trace name really only wants sample name and primer name, but other code will want batch, amplification, etc.

It is possible that primer, amplification and others might appear more than once in the parse results, this because the plate name will often contain such fields and appear as part of the trace file name. Nothing is done with the multiple matches but it might be reasonable to compare the groups with the same assignment and fail those that match different values as an additional check on the match. Other checks might be added: comparing the matched well or capillary against the trace file's internally supplied values, etc.

If a region of a field name contains either a sample name or control type, there must be 2 different name patterns to distinguish between them, one to pick out the sample name and the other to pick out the control name. Control names are compared against a list of common synonyms:

PC = HL60, PositiveControl, Pos, POS
NC = NegativeControl, Neg, NEG
RB = ReagentBlank

## 6.11 Type Warning



# 7 Renaming Files

Often when a trace is not included in an MTexpert project it is because the file name does not comply with the required naming conventions in Section 1.3.5. One way to fix this is to close the software, navigate to the directory of files, and rename the files as needed. Alternately we have built a utility to rename files in the project directory from within the MTexpert software. Edit → Rename Traces opens the Rename Traces window shown in Figure 32.

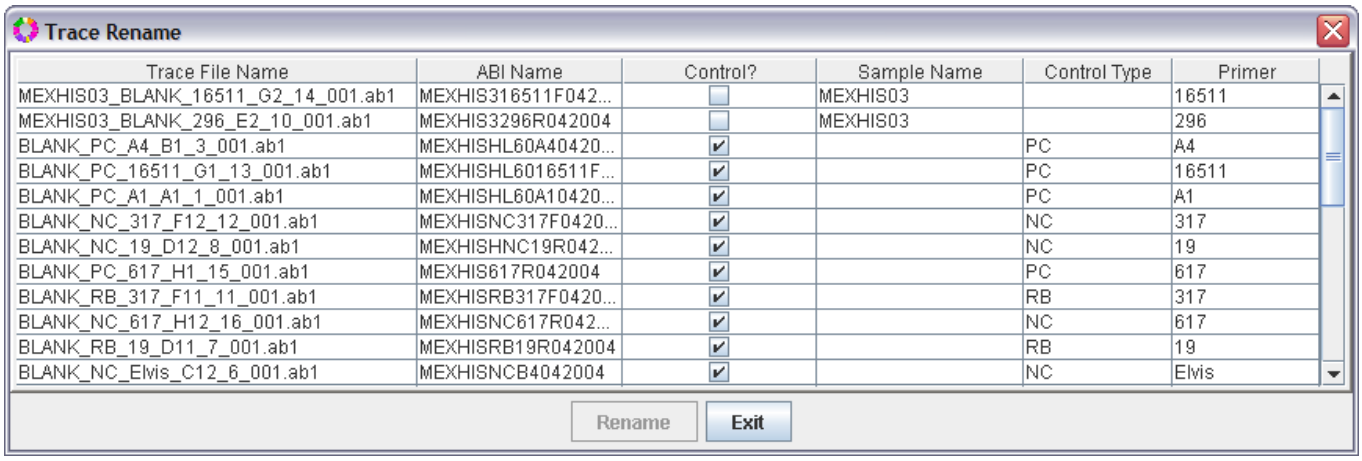



**Figure 34 The Trace Rename Window**

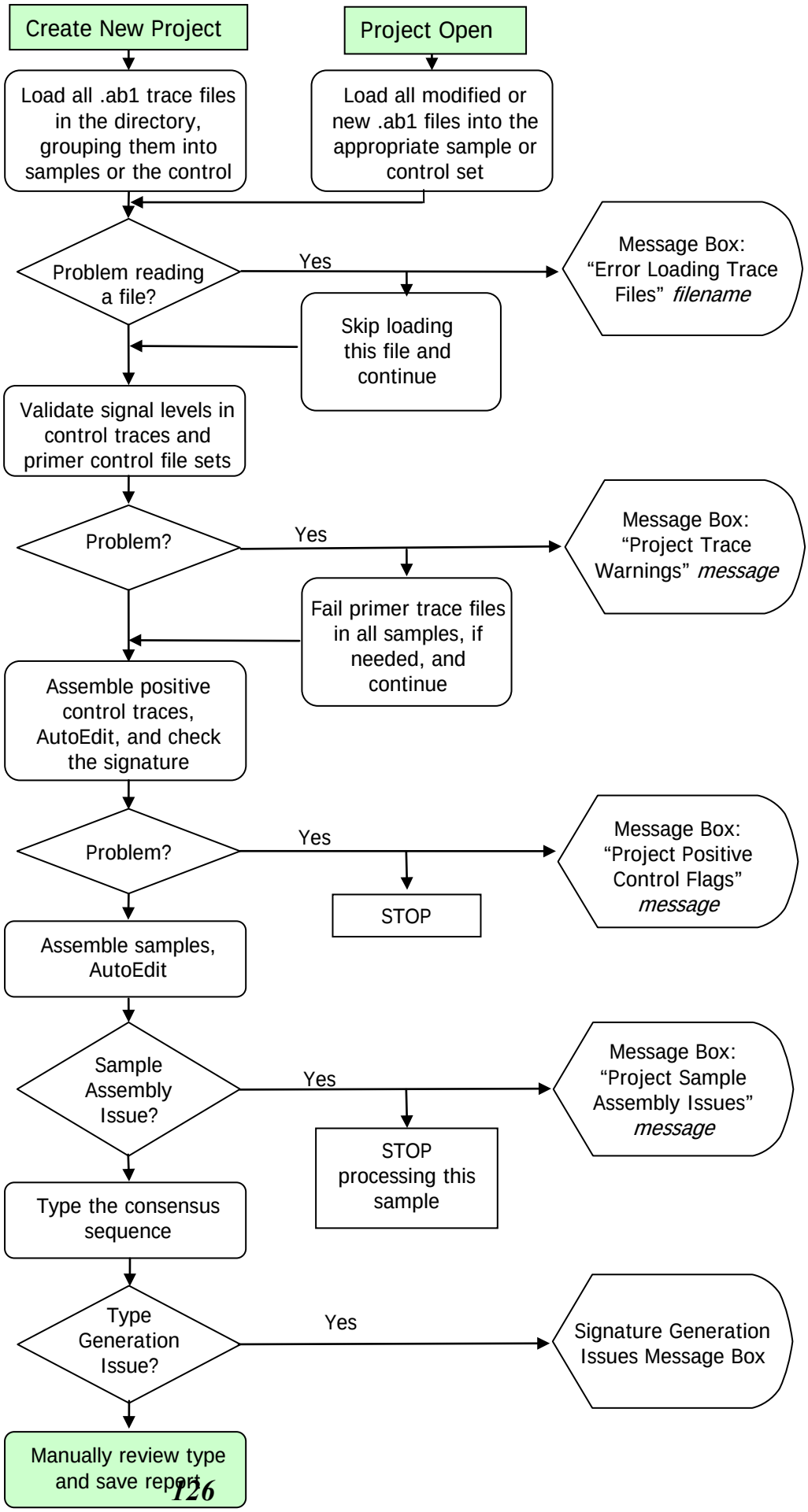
Each row in the table in the Trace Rename Window corresponds to a .ab1 file in the directory. The file name is in the left column and information read from the file's "ABI Name" in the second column from the left is whatever information was put into the Sample Name on the ABI run sheet when the data was generated. MTexpert does not read this information; it is provided for reference. The "Control?" "Sample Name" "Control Type" and "Primer" columns contain editable boxes can be used to change the sample name. When a box is edited or changed, the background changes color and the "Rename" button is enabled. Selecting Rename makes the appropriate change in the file name in the project directory and restarts the project processing workflow at Project Open (see the flow chart in Section 9.1).

# 8 Help

The Help pull-down menu accesses a splash screen with the software version number.

# 9 Workflow Diagrams

## *9.1 Top level automated data analysis workflow for a plate project*



THIS PAGE INTENTIONALLY LEFT BLANK

# *Presentations and Posters*

## *Presentations*

1. Roby R. Urban Institute Site Visit. Meeting with The Urban Institute: Fort Worth, TX; February 25, 2009.

2. Roby R., Phillips N., Thomas J., Kepler R., Elling J., and Eisenberg A. Quality Assessment and Alert Messaging Software for Raw Mitochondrial DNA Sequence Data. Conference Proceedings of the Sixteenth American Academy of Forensic Sciences 2010: Seattle, WA; February 25, 2010.

3. Phillips N. eFAST© Software: Automated Quality Assessment, Alert Messaging, File Distribution and Sample Tracking of Mitochondrial DNA Sequence Data. Proceedings of the University of North Texas Health Science Center Eighteenth Annual Research Appreciation Day 2010: Fort Worth, TX; April 23, 2010.

4. Roby R. Improving Efficiency in the (Mitochondrial) DNA Laboratory. The NIJ Conference 2010: Washington, D.C.; June 16, 2010.

5. Phillips N. and Roby, R. Expert System Rules and Software Advancements for Mitochondrial DNA Analysis. Conference Proceedings of the Seventeenth American Academy of Forensic Sciences 2011: Chicago, IL.

6. Phillips N. UNTHSC Mitochondrial Working Group: Fort Worth, TX; June 24, 2011.

7. Sprouse M. UNTHSC Mitochondrial Working Group: Fort Worth, TX; June 24, 2011.

8. Sprouse M., Kavlick M., Phillips N., and Roby R. Internal Validation of a Real-Time qPCR Assay for Human mtDNA. Conference Proceedings of the Eighteenth American Academy of Forensic Science 2012: Atlanta, GA.

## *Posters*

1. Roby R., Thomas J., Phillips N., Gonzalez S., Planz J., and Eisenberg A. High Throughput Processing and Increased Efficiency for Mitochondrial DNA Testing: Robotics, Automated Sample Tracking and Filter Metrics. Poster Presentation at the Fifteenth American Academy of Forensic Science 2009: Denver, CO; February 20, 2009.

2. Nutall K., Thomas J., Fast S., Shetty P., Vishwanatha J., and Roby R. Prostate Cancer Sample Repository of Sera and DNA. Poster Presentation at the University of North Texas Health Science Center Eighteenth Annual Research Appreciation Day 2010: Fort Worth, TX; April 23, 2010.

3. Nutall K., Thomas J., Fast S., Shetty P., Vishwanatha J., and Roby R. Prostate Cancer Sample Repository of Sera and DNA. Poster Presentation at the Fifth Annual Texas Conference on Health Disparities 2010: Fort Worth, TX; May 27, 2010.

4. Sprouse M, Phillips N, and Roby R. Internal Validation of a Real-Time Quantitative Polymerase Chain Reaction Assay for Human Mitochondrial DNA. Poster Presentation at the University of North Texas Health Science Center Nineteenth Annual Research Appreciation Day: Fort Worth, TX; April 1, 2011.