Human Identity Analysis using NextGENe® Software

John McGuigan, Kevin LeVan, Megan McCluskey, Jonathan Liu, Shouyong Ni

Introduction

There is tremendous potential for the use of next-generation sequencing in forensic analysis. It promises to make analysis faster and cheaper by combining multiple samples in a single run. It also increases the amount of information available- comparing sequence data rather than electrophoresis results allows for the comparison of Single Nucleotide Polymorphisms (SNPs) between samples. NextGENe has two new applications specifically designed to make this type of data analysis as simple as possible- one for targeted STR analysis and one for mitochondrial amplicon analysis. Samples with multiple datatypes sequenced together can be run one analysis at a time, using the unmatched reads to create other analysis projects.

For both targeted STR analysis and mitochondrial amplicon analysis the theory is the same. The most likely source of each read can be determined through alignment and sequence comparison (finding all mismatches) and the read can be counted as an observation of that allele or get marked as unreliable. False negatives are unlikely as long as there are a sufficient number of observations for each locus or amplicon.

NextGENe reports results in a comprehensive "locus" or "amplicon" report with detailed results in allele reports for each locus or amplicon. There are also graphical displays of results available, as seen in Figure 1.

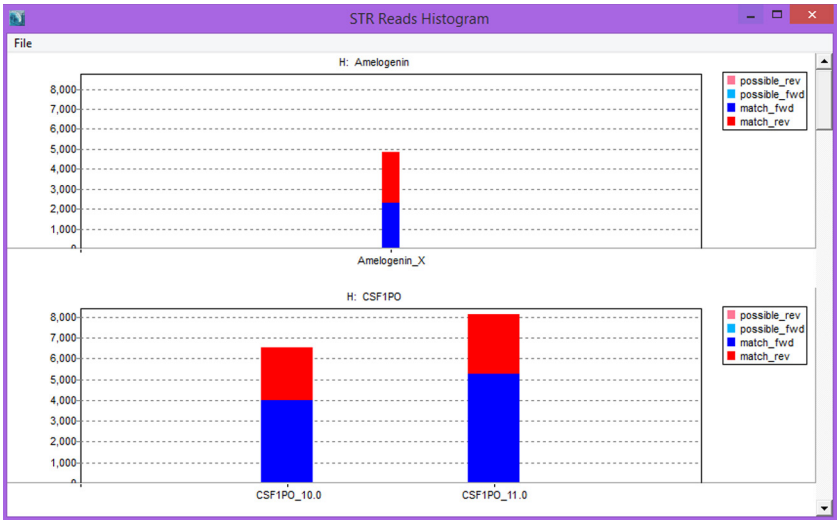


Figure 1: Histogram of results for two alleles (Amelogenin and CSF1PO).

Procedure

1. The appropriate application ("STR analysis" or "Mitochondrial Amplicon") is selected on the first page of the project wizard. A single sample is loaded for analysis on the next page (Figure 2). STR analysis requires multiple FASTA file references- one for each locus. The name of the file is used as the locus name, and the names of sequences in the file are used as known allele names. Mitochondrial analysis requires a genbank sequence of the mitochondrial genome and a single BED file specifying amplicon locations in the "Set ROI from BED file" option. If the BED file is not loaded, the software will try to estimate amplicon locations and will (if the frequency is high enough) report shorter reads as separate alleles.

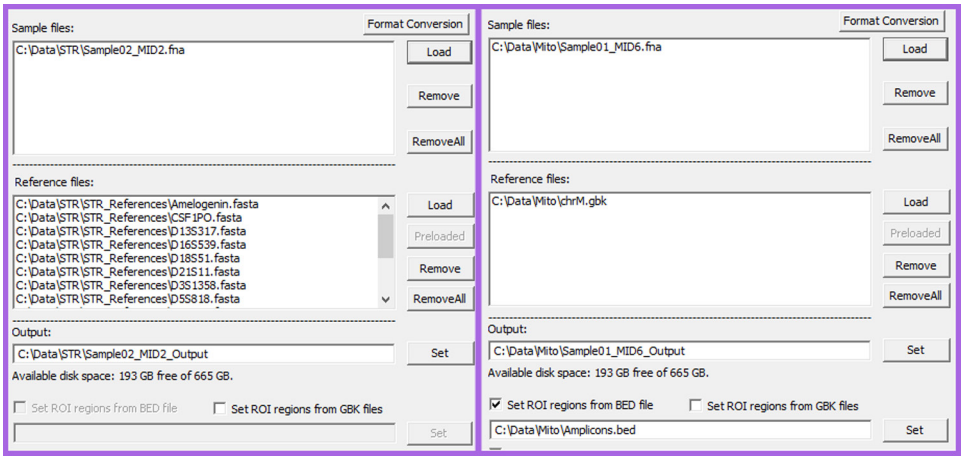


Figure 2: Loading the sample and references for STR analysis (left) and mitochondrial amplicon analysis (right)

2. Alignment and variant calling settings are adjusted in the same way as a typical alignment project. There is one additional setting for STR Analysis- "Read Length Over Reference Length". When a read aligns it must cover at least this much of the reference allele in order to be aligned. This helps prevent short reads (which terminate in a repeat region) from aligning to the shortest reference allele.

   Any potential variants that fail the mutation filters are ignored when generating consensus allele sequences. It is important to set low filters when attempting to detect heteroplasmy in mitochondrial amplicon sequencing. If the filters are too stringent the variants will be treated as sequencing errors and no heteroplasmy will be detected.
3. After alignment is finished, the projects are opened in the viewer and the appropriate reports can be shown using the report selection button (Figure 3).

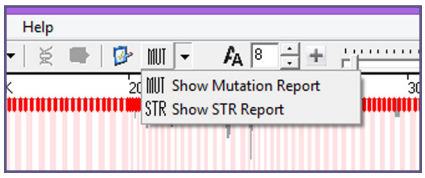
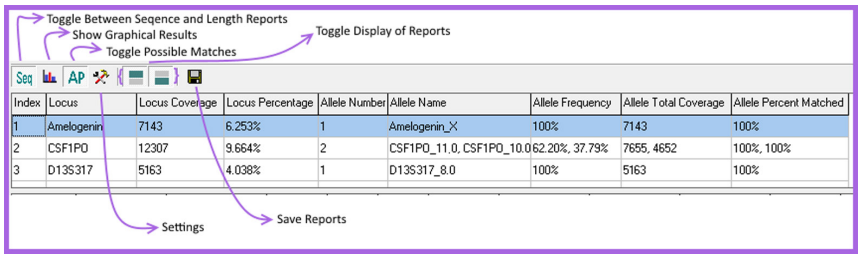


Figure 3: showing the STR report

Results

The locus report (STR) or Amplicon Report (when using the mitochondrial amplicon application) is displayed in the upper-right quadrant (Figure 4) while the allele report is displayed in the lower-right (Figure 5).



Index	Locus	Locus Coverage	Locus Percentage	Allele Number	Allele Name	Allele Frequency	Allele Total Coverage	Allele Percent Matched
1	Amelogenin	7143	6.253%	1	Amelogenin_X	100%	7143	100%
2	CSF1PO	12307	9.664%	2	CSF1PO_11.0, CSF1PO_10.0	62.20%, 37.79%	7655, 4652	100%, 100%
3	D13S317	5163	4.038%	1	D13S317_8.0	100%	5163	100%

Figure 4: Locus Report (STR Analysis)

The locus report describes:

- The coverage (number and percentage) for each locus (reads aligned in that FASTA file)
- The number, names, frequencies, and coverage of alleles that passed in each locus
- The percent matched for each allele- this may be lower than 100% if possible alleles are allowed. Possible alleles are aligned to the most similar allele, but may have sequence differences.

The locus report may be toggled between traditional length-based comparison (alleles are examined for matching lengths instead of matching sequences) and full sequence-based comparison. This will be reflected in the report and in the histogram visualization (where the x-axis will switch between discrete allele names and lengths of sequences).

The amplicon report in mitochondrial analysis describes the same information by amplicon instead of by locus with some differences. Alleles are not named and sequences aren't displayed since each allele is a different sequence. There are no "possible" matches and there isn't a way to examine lengths of alleles instead of sequences because this would be meaningless.

The allele reports are also similar for both types of analysis (Figure 5). Each allele is reported on a separate line. The mitochondrial report also includes an "Incomplete" line for counting reads that did not cross the entire amplicon if a BED file was loaded to define amplicon regions. Each allele has a reported frequency and read counts. The "Differences" column reports the number of variants compared to the reference mitochondrial genome, or, for STR analysis, the number of differences compared to the reference allele for "possible" matches. STR alleles will also have a column listing the allele name and one reporting that the consensus sequence was either "matched" or "possible".

Index	Sequence	Start	End	Frequency	Total Reads	Forward Reads	Reverse Reads	Differences
1	CACCCTATTAA	15	429	82.39%	674	466	208	3
2	CACCCTATTAA	15	429	14.79%	121	0	121	4
3	Incomplete	NA	NA	0%	0	0	0	NA

Figure 5: Allele report for one amplicon in a mitochondrial amplicon analysis

STR results can be visualized using histograms as seen in Figure 1. The mitochondrial visualization can be seen in Figure 6. Variants are shown in zoomed-out view that makes it simple to see haplotype patterns. Insertions and deletions are shown as green and red dots, while SNPs (compared to the reference sequence) are shown with the observed nucleotide.

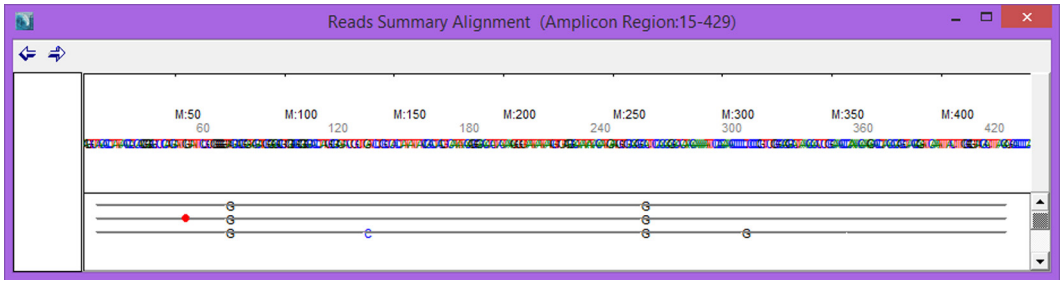


Figure 6: Reads Summary Alignment for Mitochondrial Analysis

The settings dialogs for both reports allow for filtering alleles bases on the number of reads, frequency of reads in the locus/amplicon, forward/reverse balance of the reads, or the number of differences between the consensus sequence and the reference sequence. Any reads that get filtered out are treated as an "Unknown" allele for that locus or amplicon.

Discussion

Both STR analysis and Mitochondrial Amplicon analysis have similar analysis goals. It is important in both cases to accurately identify the true sequences and to accurately quantify them. Potential false positives are removed at several different steps in the process- by the alignment, by the mutation detection filters, or by the report-specific filters.

The software organizes the results into easy-to-understand reports with a main report (locus or amplicon) which is linked to allele reports. Examining all of the potential alleles for a given loci or amplicon is as easy as double-clicking the row in the main report to show the corresponding allele report. Any allele in this secondary report can be double-clicked to navigate to supporting reads in the viewer.

Acknowledgements

We would like to thank Hanna Kim and Dr. Cassandra Calloway from Children's Hospital Oakland Research Institute for their valuable feedback during development of these tools.