SOFTGENETICS®
Software PowerTools for Forensic Analysis

GeneMarker®HTS
Forensic NGS Analysis Software

Concordance Study: High-Throughput STR/YSTR results using GeneMarker®HTS software are concordant with capillary electrophoresis results and provide additional sequence variation information.

K. Hendricks, C.S. Liu, L. Luo, J. McGuigan, T. Snyder-Leiby, N. Wan, J. Wu • SoftGenetics LLC, State College, PA

ABSTRACT

GeneMarker®HTS software was released in 2016 as a rapid, user-friendly, software for forensic mtDNA analysis. Several Forensic Science International: Genetics articles evaluating and using the software were published in 2017 (Holland *et al*., Calloway *et al*., Riman). Development of autosomal and Y-STR capabilities started in 2017; providing genotype and single nucleotide polymorphism (SNP) detection within amplicons.

While mtDNA analysis is essential in many forensic applications, STR DNA analysis impacts countless investigations and court cases. Strengths of this data include both its resolving power for excluding an individual and the ability to determine potential relationships between evidence and suspects due to Mendelian inheritance of nuclear DNA. High throughput sequencing methods provide additional resolving capabilities in situations where two individuals have the same allele sizes with differences in nucleotide repeat sequence.

To ensure continuity with the well-established capillary electrophoresis (CE) nomenclature, it is essential to confirm concordance of MPS analysis results with CE results. This presentation details a concordance study between GeneMarkerHTS STR/YSTR results and the corresponding samples analyzed by capillary electrophoresis.

National Institute of Standards and Technology (NIST), in conjunction with Promega corporation, generously supplied the fastq sequence files and the corresponding CE allele calls for 672 samples amplified with the PowerSeq® Auto/Y System and analyzed on an Illumina® MiSeq. Results of these data analyzed in m GeneMarkerHTS software were highly concordant with the CE allele calls. Summary of the allele calls concordance and examples of instances where alleles exhibited sequence variation will be presented.

Chemistries for mtDNA and STR amplification for HTS platforms enable the laboratory to have the benefits of both mtDNA and STR analysis at the same time. In addition to identification of sequence polymorphisms, advantages of HTS STR chemistries over traditional CE STR chemistries include the ability to have smaller amplicons and to analyze more loci in one reaction. GeneMarkerHTS is rigorous, user-friendly software for the analysis of HTS data for STR and mtDNA applications.

HTS STR Analysis

High-throughput sequencing data for forensic applications, database or casework, can be analyzed by selecting a built-in panel or by loading a panel for custom chemistries. Primer sequences are used to sort and trim the input reads in order to find the number of reads with each unique sequence in each amplicon in the panel. The reads are merged when using paired-end data, with overlapping sequences used to correct errors when possible. The expected repeat pattern (STR) or specific sequence (Amelogenin and SNP amplicons) is identified using regular expressions (regex strings) which are included in the panel. This matching process is used to name the allele sequences that were found, and filters exclude low-frequency sequences caused by sequencing errors.

Autosomal and Y-STR analysis includes conventional forensic nomenclature. GeneMarker HTS software capabilities include simultaneous analysis and reporting of mtDNA and STR chemistries.

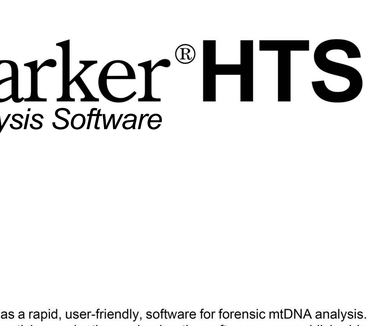

Figure 1. The analysis panel can be selected from the Panel drop down. Then import the sample files and start processing.

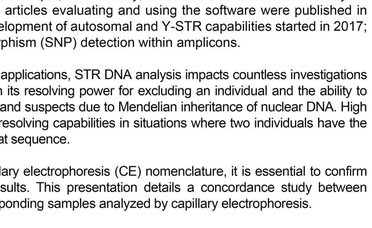

Figure 2. After the data is processed, the results can be viewed. The initial view is an overview of all the results, Auto/Y-STR, with a separate viewer if the chemistry includes mtDNA. The interactive sunburst or Category drop down can be used to narrow the scope of results being reviewed. The **overview** provides the categories of data, read count, percent of total reads and the forward/reverse counts.

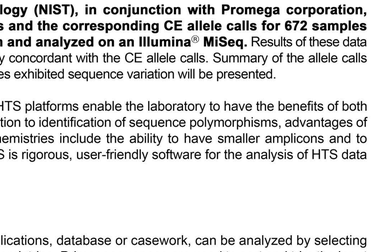

Figure 3. The **locus** view displays detailed results for each allele in the locus. Sequence information for each allele can be viewed from the **sequence** tab.

CE results concordant with HTS results

National Institute of Standards and Technology (NIST), in conjunction with Promega corporation, generously supplied the fastq sequence files and the corresponding CE allele calls for 672 samples amplified with the PowerSeq® Auto/Y System and analyzed on an Illumina® MiSeq.

GeneMarker HTS software results were 99.74% concordant with the CE allele calls of 20,000 sampled loci. The concordance study evaluated Autosomal and Y-STRS. GeneMarker HTS reports the percentage of sequences for a given allele.

A call is considered:

- Homozygous when the 1st allele is >= 70%
- Homozygous-split when the 1st allele is >= 70% but the top 2 sequences are each >= 35% in that allele
- Heterozygous when the 1st two alleles are each >= 20%

Concordant	19539
Concordant (Homozygous-Split)	408

Table 1. The HTS STR/Y-STR analysis resulted in 19,539 concordant allele calls and 408 Homozygous-Split concordant allele calls.

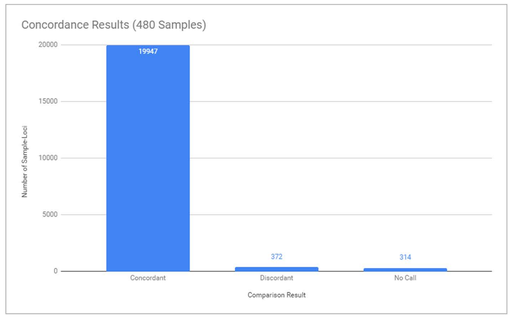

Figure 4. The HTS STR/Y-STR results were highly concordant with the corresponding CE allele calls. Only 372 allele calls were discordant and 314 no calls reported. A total of 20,000 loci were sampled.

Reason for Discordance	
Unable to Name Sequence	188
Dropout from Heterozygous	84
Discordant Allele Name(s)	71
Discordant Number of Alleles	29

Table 2. Of the discordant calls, there were 188 instances where the sequence was found and called but unable to be named and 71 instances of discordant allele names. This discordance may be due to variation in the allele sequences compared to the known information or sequencing errors.

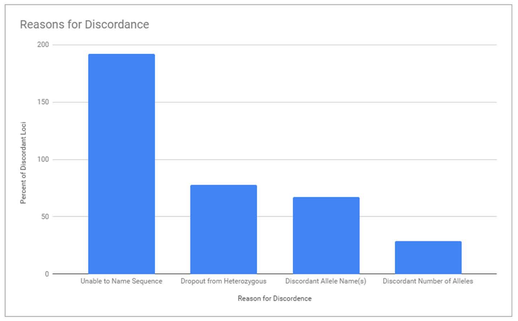

Figure 5. A graphical display of the discordance calls found in the study.

Additional Sequence Information

High-throughput sequencing can reveal additional information that is not available from the traditional CE data. Isoalleles are loci that appear homozygous in length-based measurements (such as CE), but are heterozygous by sequence. High-throughput sequencing reports the percentage of sequences for a given allele and sequence variants. This depth of information has applications in identification of individuals and relatives in single source samples and the potential for improved assignment to contributors during analysis of mixtures.

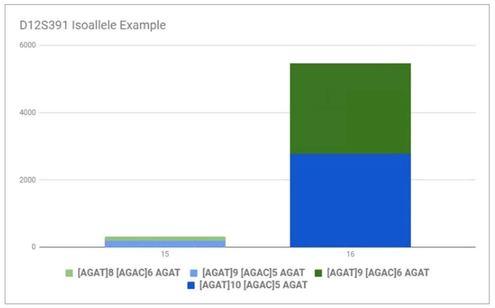

Figure 6. An example of an isoallele. High-throughput sequencing provides additional information that cannot be determined from the traditional CE data.

Conclusion

GeneMarker HTS software results proved to be 99.74% concordant with CE allele calls of 20,000 sampled loci. High-throughput sequencing can reveal additional information that is not available from the traditional CE data. The additional sequence information can be beneficial in forensic and casework applications. Strengths of this data include both its resolving power for excluding an individual and the ability to determine potential relationships between evidence and suspects due to Mendelian inheritance of nuclear DNA.

Chemistries for mtDNA and STR amplification for HTS platforms enable the laboratory to have the benefits of both mtDNA and STR analysis at the same time. GeneMarker®HTS software provides a streamlined workflow for forensic mitochondrial and STR DNA data analysis from all major High Throughput Sequencing (HTS) systems and chemistries.

If you are interested in learning more or would like a free, 35-day trial version of GeneMarkerHTS, please visit booth #512.

Acknowledgements

We would like to thank Dr. Peter Vallone at National Institute of Standards and Technology (NIST) for generously supplying data to complete the concordance study between the CE results and HTS STR/Y-STR results. We would also like to thank Promega Corporation, Madison, WI, USA for providing Autosomal and Y-STR data.

SOFTGENETICS®
Software PowerTools for Genetic Analysis
www.softgenetics.com